Tampereen ammattikorkeakoulu

Datan laatu koneoppimisessa

Päivi Hulkkonen

Eeva Raunnos

OPINNÄYTETYÖ
Kesäkuu 2021

Dataosaamisen ja tekoälyn koulutusohjelma (ylempi AMK)

TIIVISTELMÄ

Tampereen ammattikorkeakoulu
Dataosaamisen ja tekoälyn koulutusohjelma (ylempi AMK)

HULKKONEN, PÄIVI & RAUNNOS, EEVA:
Datan laatu koneoppimisessa

Opinnäytetyö 101 sivua
Kesäkuu 2021

Datan määrä organisaatioissa kasvaa kiihtyvällä tahdilla. Perinteisen raportoinnin ja data-analytiikan rinnalla halutaan hyödyntää tekoälyä ja koneoppimista liiketoiminnan kehittämisessä sekä uusissa liiketoimintamahdollisuuksissa. Tämän mahdollistamiseksi datan laadullisiin ominaisuuksiin tulee kiinnittää entistä enemmän huomiota.

Tässä opinnäytetyössä tutkittiin toimeksiantajaorganisaation tietokannan datan laatua ja arvioitiin sen valmiutta koneoppimisen hyödyntämiseen. Tutkimus toteutettiin Kiinteistönvälitysalan Keskusliitto Ry:n KVKL Hintaseurantapalvelulle case-tutkimuksena.

Tutkimuksen teoriaosuudessa käsiteltiin datan laadun rakentumista, laadun ulottuvuuksia ja niiden mittaamista kokonaisuutena. Lisäksi käsiteltiin tekoälyn ja koneoppimisen perusteita, erityisesti koneoppimisen ennustemallien näkökulmasta. Tutkimusosuudessa keskityttiin analysoimaan toimeksiantajan datan laatua objektiivisten mittareiden kautta. Tämän lisäksi tutkimuksessa testattiin kahta erityyppistä koneoppimismallia. Malleja koulutettiin erilaisin tavoin esikäsitellyillä data-aineistoilla. Näin osoitettiin datan laadun merkitys koneoppimisen ennustemalleille.

Tutkimuksen tulokset osoittivat datan muuttuneen merkittävästi vuosien varrella. Datan sisällön ja laadun havaittiin kehittyneen hyvään suuntaan. Koneoppimiskokeilussa ennustemallit ennustivat jopa 90 %:n tarkkuudella asunnon hinnan oikein datan laadun puhdistustoimenpiteiden jälkeen. Vaikka tulos oli varsin hyvä, ennustetarkkuutta saataisiin todennäköisesti parannettua keskittymällä datan oikeellisuuden ja oikeamuotoisuuden parantamiseen. Tulosten pohjalta toimeksiantajalle annettiin kehitysehdotuksia datan laadun kehittämiseksi. Tämän opinnäytetyön ulkopuolelle rajattiin tarkempi koneoppimismallien valintaan ja optimointiin liittyvä läpikäynti.

Asiasanat: datan laatu, koneoppiminen, asunnon hinnan ennustaminen

ABSTRACT

Tampereen ammattikorkeakoulu
Tampere University of Applied Sciences
Master's Degree Programme in Data Expertise and Artificial Intelligence

HULKKONEN, PÄIVI & RAUNNOS, EEVA:
Data Quality in Machine Learning

Master’s thesis 101 pages
June 2021

Data volumes keep increasing in organisations. Alongside traditional reporting and data analytics, the aim is to utilise artificial intelligence and machine learning in business development and new business opportunities. Machine learning requires high-quality data, and therefore it should be given more attention.

The purpose of the study was to focus on the quality of data and to evaluate its significance for machine learning. The study was conducted as a case study for Kiinteistönvälitysalan Keskusliitto Ry, KVKL Hintaseurantapalvelu.

The theoretical part of the study contained basics of data quality and strategy, quality dimensions, and measurement. Further, the theoretical part included basics of artificial intelligence and machine learning, especially from the perspective of machine learning prediction models. The research part focused on analysing data quality through objective dimensions directly related to the utilisation of machine learning. In addition, two different machine learning models were tested with differently pre-processed data sets, and thus demonstrating the importance of data quality for prediction models.

The results showed a significant change in the data over the years. The data content and quality had improved. In the machine learning experiment, the prediction models predicted the price of the apartment with up to 90 % accuracy after data pre-processing. Although the result can be considered quite good, the accuracy of the prediction could probably be improved by focusing more on machine learning models, which was not in the scope of this thesis. Based on the results, development suggestions were provided to improve the quality of the data.

Key words: data quality, machine learning, house price prediction

SISÄLLYS

1 JOHDANTO

Datan määrä organisaatioissa on kasvanut merkittävästi viimeisen vuosikymmenen aikana ja tahti yhä kiihtyy. Data on pohjana päätöksille ja datan tuoma potentiaali ymmärretään koko ajan paremmin. Siinä missä hyvälaatuinen data voi ratkaista yrityksen liiketoimintahaasteita, voi huonolaatuinen data johtaa vääriin liiketoiminnallisiin ratkaisuihin. Perinteisen raportoinnin ja data-analytiikan rinnalla halutaan hyödyntää tekoälyä ja koneoppimista liiketoiminnan kehittämisessä. Jotta tämä olisi mahdollista, tulisi datan laatuun vaikuttaviin tekijöihin kiinnittää entistä enemmän huomiota.

Tässä opinnäytetyössä tutkitaan toimeksiantajaorganisaation tietokannan datan laatua ja arvioidaan sen valmiutta koneoppimisen hyödyntämiseen. Lisäksi tutkitaan, millaisilla datan laadun parantamisen esikäsittelytoimenpiteillä voidaan vaikuttaa ennustustarkkuuteen. Tarkoituksena on antaa lukijalle helposti lähestyttävä kokonaisuus datan laatuun vaikuttavista tekijöistä. Teoriaosuudessa käsitellään datan laadullisia ulottuvuuksia ja niiden mittaamista kokonaisuutena. Tutkimusosuudessa keskitytään kuitenkin niihin ulottuvuuksiin ja objektiivisiin mittareihin, joilla on suora yhteys tekoälyn ja koneoppimisen hyödyntämisen mahdollisuuksiin. Tutkimus toteutetaan case-tutkimuksena, jonka toimeksiantaja on Kiinteistönvälitysalan Keskusliitto Ry:n KVKL Hintaseurantapalvelu.

1.1 Tutkimuksen rakenne ja rajaus

Tässä opinnäytetyössä käsitellään datan laadullisia näkökulmia. Työssä otetaan huomioon datan laatuun vaikuttavia tekijöitä laadun muodostumisen, suunnittelun ja käsittelyn eri vaiheista, aina sen hyödyntämiseen asti. Näin ollen lukija saa käsityksen, kuinka monipuolinen kokonaisuus datan laadun rakentuminen on.

Työssä on käytetty pääsääntöisesti suomenkielisiä termejä. Koska osa termeistä on haastava suomentaa aukottomasti, on termin yhteyteen lisätty englanninkielinen vastine.

Tutkimuksen teoriaosuudessa lukija johdatetaan tutkimusaiheeseen kertomalla mitä data on ja kuinka erilaista dataan voidaan luokitella. Tämä antaa pohjan lähteä tutustumaan datastrategiaan ja datan laadun hallintamalliin, sekä niiden liiketoiminnalliseen merkitykseen. Sillä, kuinka dataa varastoidaan ja käsitellään, on myös merkitystä datan laatuun. Tässä työssä sivutaan tietovaraston merkitystä data-arkkitehtuurissa, keskittyen kuitenkin suoraan datan laatuun vaikuttaviin tekijöihin. Datan laadun hallintaan on olemassa erilaisia viitekehyksiä ja arviointimalleja, joiden hyödyllisyys ja olemassaolo on hyvä tiedostaa. Tässä opinnäytetyössä nämä on kuitenkin rajattu teorian ulkopuolelle.

Yhtenä osana tutkimuksen teoriaosuutta, käsitellään datan hyödynnettävyyttä koneoppimisessa. Tässä yhteydessä kerrotaan tiiviisti tekoälyn ja koneoppimisen perusteoriaa ja käydään läpi datan esikäsittelymenetelmiä, joilla dataa voidaan valmistella koneoppimismallien kouluttamiseen. Kappaleessa esitellään pääpiirteittäin kaksi koneoppimisessa yleisesti käytettyä koneoppimismallia: satunnainen metsä sekä keinotekoiset neuroverkot. Nämä mallit antavat lukijalle kuvan datan laadun ulottuvuuksien merkityksestä koneoppimisen hyödyntämisessä ja toimintaperiaatteista koneoppimismallien takana.

Tutkimusosuudessa perehdytään toimeksiantajan, Kiinteistönvälitysalan Keskusliitto Ry:n KVKL Hintaseurantapalvelun datan laadun arviointiin ja analysointiin koneoppimisen näkökulmasta. Data-aineistoa tutkitaan datan laadun ulottuvuuksien kautta, joka tarjoaa hyvän kokonaiskuvan toimeksiantajan dataan ja sen laatuun. Tutkimuksen yhteydessä testataan kahta aiemmin esiteltyä koneoppimismenetelmää asunnon hinnan ennustamisessa. Tutkimuksen tarkoituksena ei ole toteuttaa tuotantokäyttöistä koneoppimismallia, vaan havaintojen kautta osoittaa oleellisimmat datan laadun osa-alueet koneoppimisen näkökulmasta.

Lopuksi annetaan arvio toimeksiantajan datan laadusta sekä jatkokehitysehdotuksia. Työ tiivistyy päätelmiin, jossa annetaan vastaukset esitettyihin tutkimuskysymyksiin ja arvioidaan työn onnistumista.

1.2 Tutkimuskysymykset

Tässä opinnäytetyössä pyritään vastaamaan seuraaviin tutkimuskysymyksiin:

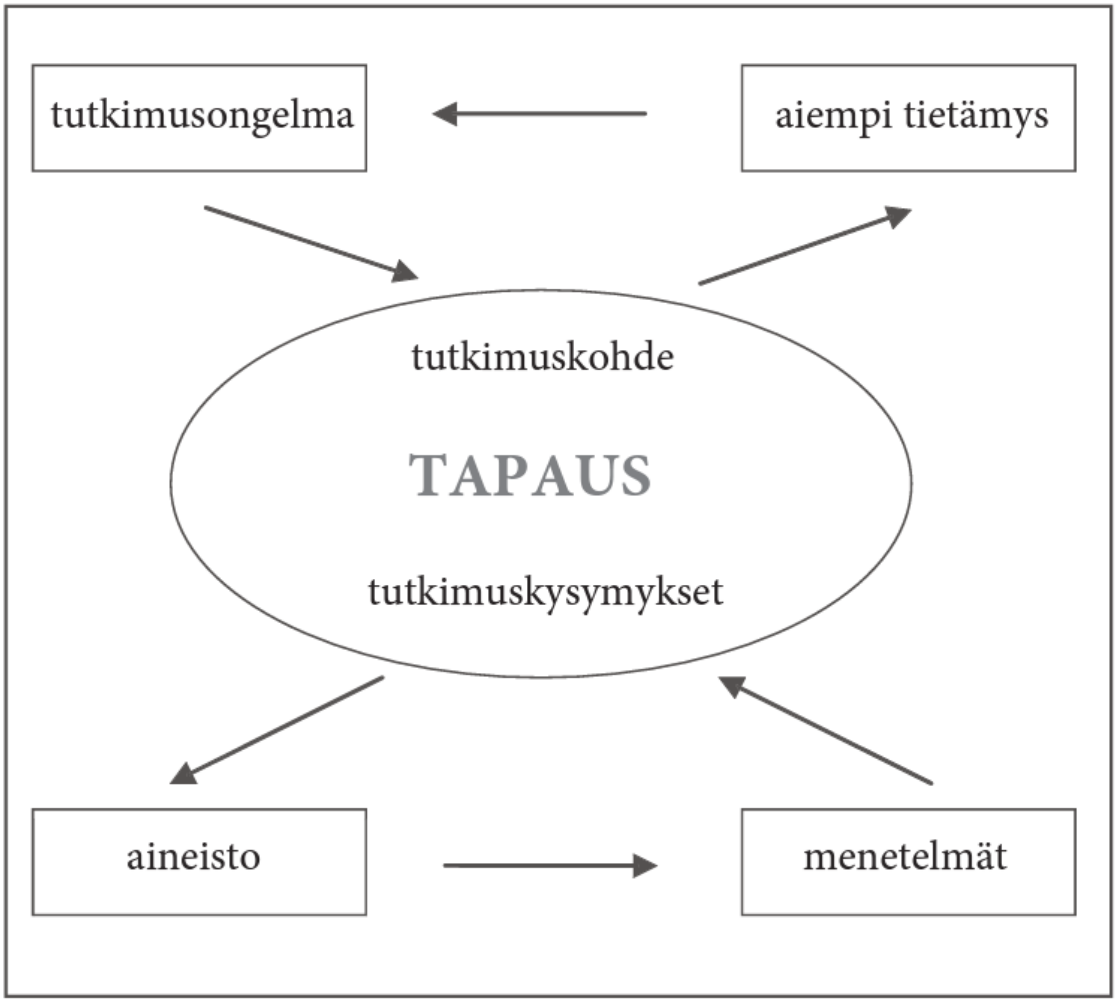
1. Kuinka koneoppiminen muuttaa datan laatuvaatimuksia?
    a. Kuinka datan laatu vaikuttaa koneoppimismallien tarkkuuteen?
    b. Kuinka datan laadullisiin ominaisuuksiin voidaan vaikuttaa?
2. Minkälainen liiketoiminnallinen merkitys datan laadulla on?

1.3 Tutkimusmenetelmä

Tässä työssä tutkimusstrategiana käytetään tapaus- eli case-tutkimusta. Työn tutkimuksellisena lähtökohtana on laadullinen, eli kvalitatiivinen tutkimus, jota täydentää kvantitatiivinen, eli määrällinen tutkimusosuus. Tämä opinnäytetyö toteutetaan parityönä, jossa molempien tutkijoiden aiempi kokemus ja vahvuusalueet rikastuttavat tutkimusasetelmaa.

Tapaustutkimuksessa tutkimusasetelma on kytköksissä aiempaan teoria- ja tutkimuspohjaan ja tutkittava tapaus muodostaa tietyn kokonaisuuden (Saaranen-Kauppinen & Puusniekka 2006). Tutkimuksessa pyritään saamaan kattava kuva tietystä ilmiöstä pyrkimättä kuitenkaan yleistettävään tietoon. Teoriaosuus luo pohjan analyysille ja johtopäätöksille. Tapaustutkimus tuottaa usein hypoteeseja ja mahdollisia jatkotutkimusajatuksia toimeksiantajalle. Tapaustutkimusta hyödynnetään paljon liiketalous-, hallinto- ja teknisissä tieteissä, joissa tutkitaan tiettyä, itsenäistä kokonaisuutta. (Aaltio-Marjasola 1999.)

Koska tapaustutkimuksessa tutkijan omat tiedot ja näkemys saattavat vaikuttaa tutkimukseen, on tärkeää, että niitä tuetaan tuloksia varmentavilla käytännöillä, kuten triangulaatiolla. Triangulaation, eli monimetodisen lähestymistavan avulla lisätään tutkimuksen validiutta. (Laine, Bamberg & Jokinen 2007, 26–29.) Kerätyn aineiston ja tutkimustulosten pohjalta muodostetaan analyysit ja ne tulkitaan johtopäätelmissä. Tapaustutkimusta toteutetaan joustavasti ja tutkimussuunnitelmaa voidaan muuttaa tarpeen mukaan tutkimuksen edetessä. (Aaltio-Marjasola 1999.) Parityönä tehtävä tutkimus lisää tutkimuksen validiutta mutta toisaalta myös tutkimusmenetelmän joustavuuden tarvetta. Kuviossa 1 on kuvattu triangulaation käyttöä tutkimuksessa (Laine ym. 2007, 27).



Kuvio 1. Tapaustutkimuksen triangulaatio, eli toisiaan täydentävien aineistojen, menetelmien ja näkökulmien käyttö (Laine ym. 2007, 27)

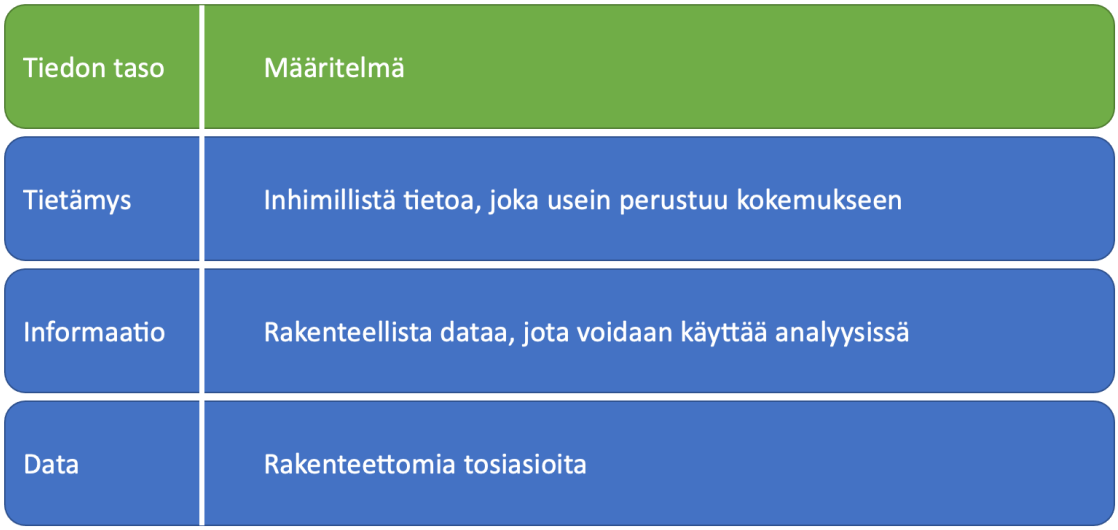
Tutkimusstrategiana case-tutkimus sopii hyvin tähän työhön, koska työssä kootaan tapauksen ympärille tutkimusaineisto yhdistämällä kvalitatiivista ja kvantitatiivista aineistoa. Kirjallisuuskatsauksen avulla haetaan ymmärrystä tutkimusongelmaan ja aihealue vaatiikin tutkijoiden omaa, varsin massiivista paneutumista alaan ja tutkimusongelmaan. Aineiston perusteella pyritään saamaan laaja kuva tutkimusongelmasta, ja tiivistää siitä lukijaystävällinen kokonaisuus. Lähteinä hyödynnetään aiheeseen liittyvää kirjallisuutta, artikkeleita, webinaareja sekä podcasteja. Aineiston hankinnassa otetaan huomioon alan nopea kehitys hyödyntämällä mahdollisimman tuoreita tietolähteitä.

2 DATA

Data on tämän opinnäytetyön keskeisin termi. Siksi työssä lähdetään liikkeelle kertomalla tiiviisti sen määrittelystä, rakenteista ja kategorioista. Datan laatu, datastrategia ja -hallinta pohjautuu näiden ominaisuuksien ymmärtämiseen.

2.1 Mitä data on?

Data on suomen kielessä yleisesti käytetty englanninkielinen termi, jota on haastava suomentaa. Mikäli käännettyä termiä käytetään, on se useimmiten tieto. Tiedolla kuitenkin viitataan kirjallisuudessa ja tutkimuksissa usein datasta työstetympään versioon. Informaatio taas puolestaan on laajempia tiedon palasia, jotka ovat saaneet merkityksen, riippuen käyttäjästä ja käyttötarkoituksesta. Data on siis kuin pieniä palasia, joiden perusteella syntyy informaatiota. (Väre 2019.) Koska tieto on käsitteenä laaja ja sillä voi olla eri tasoja, tässä työssä tullaan käyttämään raaka-aineen kuvaavampaa termiä *data*, jonka avulla pystytään rakentamaan laajempaa informaatiota ja tietämystä. Kuviossa 2 on esitelty tiedon eri tasoja.



Kuvio 2. Tiedon tasot ja niiden rakentuminen tiedon jalostuessa (mukaillen Laihonen, Hannula, Helander, Iivonen, Jussila, Kukko, Kärkkäinen, Lönnqvist, Myllärniemi, Pekkola, Virtanen, Vuori, Yliniemi 2013)

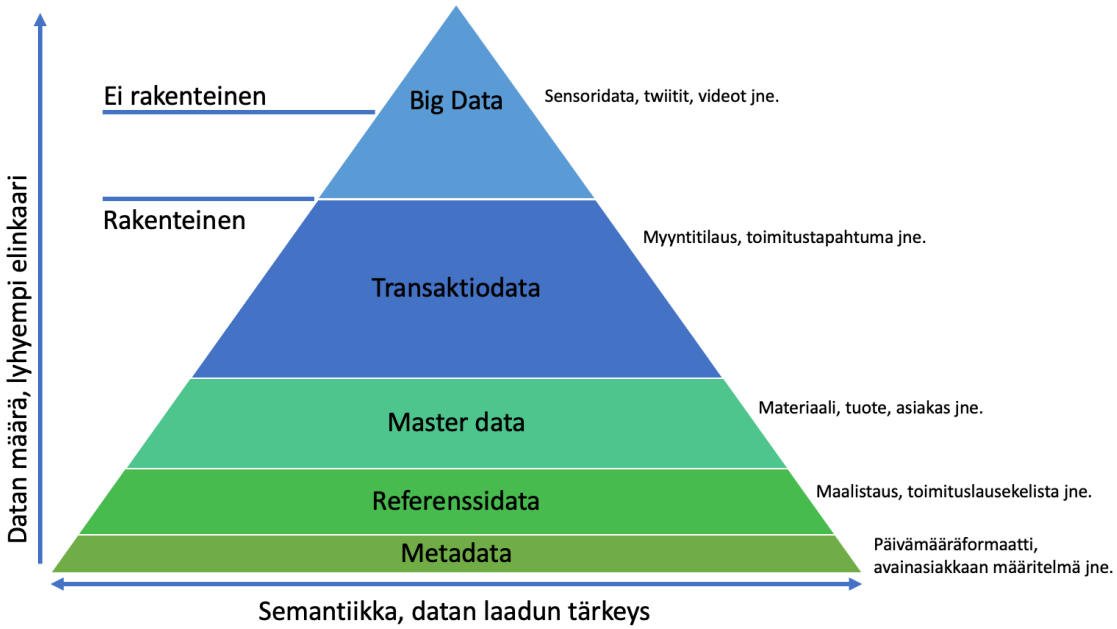
2.2 Datan rakenne

Data luokitellaan perinteisesti rakennetietomalliin perustuen rakenteiseen, puolirakenteiseen ja rakenteettomaan dataan (Hannila 2019). Toisinaan suomalaisessa kirjallisuudessa käytetään myös strukturoidun ja strukturoimattoman tiedon käsitteistä.

Yrityksissä käsitellään sekä rakenteellista, että rakenteetonta dataa (Bhansali 2013). Rakenteista dataa hallinnoidaan, tallennetaan ja käsitellään taulukoidun tietokantarakenteen kautta ja kyselyitä voidaan toteuttaa esimerkiksi SQL-kieltä käyttäen. Rakenteeton data taas ei ole tyypiltään perinteistä rivisarake-tietoa. (Hannila 2019.) Tärkeää tietoa yrityksen toiminnasta, toimialasta ja trendeistä on useimmiten juuri rakenteettomassa eli strukturoimattomassa muodossa, kuten tekstinkäsittelyasiakirjoissa, PDF-tiedostoissa, sähköpostiviesteissä, blogeissa ja verkkosivuilla. Kilpailukyvyn saavuttamiseksi suurten ja keskisuurten yritysten on voitava käyttää kaiken tyyppisiä tietoja organisaatiostaan. (Bhansali 2013.) Rakenteettoman datan määrän arvioidaan olevan yrityksissä jopa 80–90 %. Data voi olla myös luonteeltaan rakenteellisen ja rakenteettoman datan välistä. Silloin puhutaan puolirakenteisesta datasta. (Hannila 2019.)

2.3 Datan kategoriat

Dataa luokitellaan eri kategorioihin sen ominaisuuksien ja piirteiden perusteella. Datan määrittäminen oikeaan kategoriaan on olennaista datan laadun kannalta, sillä datakategorian takia, riippuvuudet ja suhteet muihin datakategorioihin voi olla merkittävässä roolissa datan laatuongelmien havainnoimisessa. Näin ollen on tärkeää ymmärtää, minkälaisesta datasta on kyse. (McGilvray 2008, 39.) Kuviossa 3 on esitelty datan kategoriat, joihin erilaiset datat jaetaan niiden ominaisuuksien ja piirteiden mukaan.



Kuvio 3. Erilaiset datan kategoriat (mukaillen Laatikainen, 2015)

Master datalla tarkoitetaan yrityksen perus- tai ydintietoa, joka on luonteeltaan melko pysyvää dataa. Se on tyypillisesti myös laajasti yrityksen eri osissa käytössä. Sen avulla saadaan vastauksia esimerkiksi siihen, mistä yrityksen toiminta muodostuu, mitä se tekee, mitkä ovat sen toimipisteet ja ketkä ovat sen asiakkaita. Master data on kriittistä dataa, eli sitä tarvitaan yrityksen päivittäisen työn sujuvuuteen ja se on myös kriittistä raportoinnin kannalta. (Väre 2019.) Koska master data on yritykselle strategisesti niin oleellista, tulee sen hallinnan olla myös strategisessa keskiössä (Niemi n.d.).

Master dataa voidaan luokitella eri tavoin. Rakenteellisesta master datasta puhutaan silloin, kun tarkastellaan yrityksen taloudellisia rakenteita tai hierarkioita. Aikasidonnainen master data puolestaan voi olla esimerkiksi hinta- tai kokoonpanotietoja. Se on käytännössä datamuotona master datan ja transaktiodatan välillä. Master data voidaan mieltää liimaksi, joka yhdistää liiketoiminta prosessit ja yrityksen IT-järjestelmät. (Väre 2019.)

Referenssidata on muun datan luokittelu- tai viitetietoa. Sen tehtävä on ryhmitellä tai luokitella muuta organisaation dataa (Väre 2019). Referenssidata tulee usein yrityksen ulkopuolelta ja on luonteeltaan standardimaista. Maa- ja valuuttakoodit ovat hyvä esimerkki standardimaisesta referenssidatasta. Monet yritykset mieltävät myös toimittajien omat tuotekoodit referenssidataksi. (Niemi n.d.)

Transaktiodata on tapahtumatietoa. Sitä syntyy eri liiketoimintaprosessien aikana eri järjestelmissä, kuten asiakkuudenhallinta- ja toiminnanohjausjärjestelmissä. Transaktiodataa ovat esimerkiksi ostotilaukset ja erilaiset kirjaukset. Transaktiodataa on siis kaikki se mitä tapahtuu yrityksen toimintaa tehtäessä. (Hannila 2019.) Toisinaan on vaikeaa erottaa master dataa ja transaktiodataa toisistaan (Väre 2019).

Big datalla tarkoitetaan nopea tempoisesti syntyvää massadataa, jota määritellään perinteisesti termeillä: massa (volume), nopeus (velocity) ja vaihtelevuus (variety). Massa kuvaa nimenomaisesti datan määrää, jota on valtavasti. Nopeudella viitataan siihen vauhtiin, jolla uutta dataa syntyy. Datan vaihtelevuudella ja monimuotoisuudella tarkoitetaan sitä, että big data voi olla monessa eri muodossa, kuten kuvina, tekstinä, twiitteina. Usein big datan määritelmään lisätään epäluotettavuus (verocity), lähes rajattomat mahdollisuudet (viability), sekä arvo (value), jolla kuvataan datan todellista tai hypoteettista arvoa tulevaisuudessa. (Törmänen 2017, 137–138.)

Oleellisia kysymyksiä big datan käsittelyssä ja sen hyödynnettävyydessä on, kuinka pystymme tunnistamaan sellaisen datan, joka tukee organisaation tavoitteita nyt tai mahdollisesti tulevaisuudessa. Myös ymmärrys siitä, kuinka dataa pystytään hyödyntämään ja yhdistelemään eri datalähteitä keskenään, tuo liiketoiminnallista lisäarvoa. Piilossa olevat riippuvuudet datan attribuuttien kesken ovat usein oleellisessa osassa big datan käsittelyä. (Törmänen 2017, 137–139.) Kuviossa 4 on esitelty big datan vektorit.

Kuvio 4. Big Datan vektorit (mukaillen Törmänen 2017, 139)

Metadatalla tarkoitetaan sellaisia tietoja, jotka kuvailevat tai luokittelevat muuta dataa. Sen avulla voidaan esimerkiksi yksittäiselle tietueelle antaa määritelmiä, kuten kentän muototietoa. Metadataa on myös opasteet, kuten minimi- tai maksimipituus, tai kentän pakollisuus. Metadata avulla pyritään varmistamaan datan laatua ja se helpottaa tiedon löydettävyyttä. (Väre 2019.)

Koska metadatalla on niin merkittävä rooli datan laadun varmistamisessa, palaamme siihen tarkemmin tulevien kappaleiden yhteydessä. Seuraavassa kappaleessa käsitellään datan laadun käsitettä ja tutustutaan datastrategiaan ja datan laadun hallintamalliin.

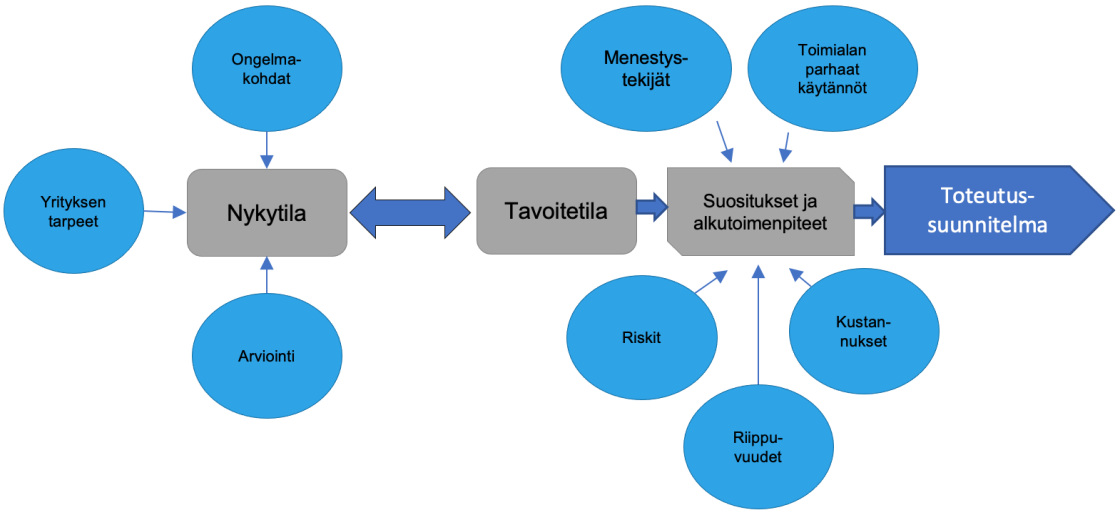
3 DATAN LAATU

Jokaisessa organisaatiossa, yrityksessä, yhdistyksessä tai laitoksessa käsitellään dataa. Mitä suurempi yritys, sitä monimutkaisemmaksi datan laadun käsite tulee (Sebastian-Coleman 2013). Datan laadussa ei ole vain yhtä määritelmää tai tiettyä tavoitetta. Asiayhteyden mukaan, datan laadulle voidaan antaa useita vaatimuksia. Tärkeintä on ymmärtää datan käyttötarkoitus ja soveltuvuus tähän tarkoitukseen. Datan voidaan todeta olevan laadukasta, mikäli se soveltuu käyttötarkoitukseen käytännössä, suunnittelussa ja päätöksenteossa. (Redman 2001.) Laadukas data ei välttämättä tarkoita sataprosenttisesti oikeita ja täydellisiä tietoja, vaan sen tulee noudattaa kyseiselle datalle asetettuja datan laadun tavoitteita (Hovi, Hervonen & Koistinen 2009). Datan laadussa on kuitenkin monta osatekijää, jotka täytyy ottaa huomioon laadukasta dataa suunniteltaessa. Käymme läpi osa-alueittain laadukkaan datan periaatteita.

3.1 Datastrategia

Datan laatuongelmat ovat yleensä piileviä ja huomataan vasta siinä vaiheessa, kun datalle olisi jo tarvetta (Hovi ym. 2009). Data voi olla monella tapaa virheellistä tai jopa käyttökelvotonta. Datan hyödyntämiseen tarvitaan datan esikäsittelytoimenpiteitä, mikä sitoo organisaatioiden resursseja. Näin ollen datan laatuun olisi syytä kiinnittää huomiota mahdollisimman aikaisessa vaiheessa. (Laihonen ym. 2013.) Datan hallinnan asiantuntija Taru Väre painottaa kirjassaan Master Data (2019), kuinka datan laadun rakentaminen alkaa jo strategiasta. Siinä missä yrityksen liiketoimintastrategia tukee yritystä toteuttamaan sen tavoitteita, tulee datastrategian pyrkiä samaan päämäärään. Siten näitä kahta ei voida erottaa erillisiksi kokonaisuuksikseen, vaan datastrategian tulee kuulua osana liiketoimintastrategiaan (Väre 2019). Toisinaan kuulee puhuttavan erikseen datan laatustrategiasta. Kyseessä on pohjimmiltaan sama asia, sillä tavoitteena on keskittyä datan ominaisuuksiin ja käytettävyyteen organisaation näkökulmasta. (Mahanti, 2019.) Laadunparantaminen tulisi siten olla osa yrityksen datastrategiaa.

Datastrategian toteutussuunnitelman avulla saadaan selville yrityksen datan tämänhetkinen tilanne. Tavoitteena on selvittää nykytila, arvioida yrityksen tarpeita ja havainnoida ongelmakohtia. Tämän jälkeen voidaan tehdä puuteanalyysi ja listata datan päivitystarpeet. (Mahanti 2019.) Kuviossa 5 on esitetty datastrategian toteutussuunnitelma.



Kuvio 5: Datastrategian toteutussuunnitelma (mukaillen Mahanti 2019)

Ellei yrityksellä ole datastrategiaa, se kannattaa luoda mahdollisimman pian. Näin ollen datasta voidaan jatkossa saada optimaalinen hyöty irti. Jos taas datastrategia on jo olemassa, on hyvä suunnitella sen ylläpitomalli datastrategian suunnitelmallisuuden varmistamiseksi. (Merilehto 2018.) Datastrategian noudattaminen on tärkeää jo pelkästään datan laadun näkökulmasta, mutta samalla se luo pohjan yrityksen järjestelmien, tietovarastojen, organisaation eri osastojen sekä henkilöstön kokonaisuudelle (Mahanti 2019).

3.2 Datan laadun hallintamalli

Datan laadun hallintamallilla tarkoitetaan toimintamallia, jonka tarkoituksena on huolehtia organisaation datan laadusta, sen ohjaamisesta, seurannasta sekä datan laadun parantamisen prosesseista. Sen tarkoituksena lisätä organisaation ymmärrystä datan laadun merkityksestä ja löytää hallitut tavat siihen, että data

sopii käyttötarkoitukseensa. Mallin tehtävänä on myös vastata datan laadun mittaamisesta, seurannasta ja parantamisesta. Nämä tulee määritellä huolella ja jalkauttaa käytäntöön koko organisaatiossa. (Väre 2019.)

Datan laadun varmistaminen on monitahoinen prosessi, johon koko yrityksen tulisi sitoutua. Datan laatustrategia ja hallintamalli tukevat tavoitteeseen pääsemistä, mutta toteutukseen liittyy usein myös haasteita. Haasteita voivat olla muun muassa (Bhansali 2013):

- Organisaatiohaasteet
    - Yhteistyö- ja kommunikaatio-ongelmat eri osastojen välillä
    - Näkemyserot sidosryhmien välillä (yhteistyökumppanit, asiakkaat, toimittajat, viranomaiset jne.)
    - Strategian ja hallintamallin puuttuminen kokonaan
    - Tehoton liiketoimintaprosessin hallinta ja suunnittelu
    - Organisaatiomuutokset
    - Globalisoituminen
    - Yhteisen käsitemallin ja nimeämiskäytäntöjen puuttuminen
- Taloudelliset haasteet
    - Laadun varmistamiseen käytettävissä oleva budjetti
    - Laadunparantamiseen käytetyn pääoman tuottoasteen laskeminen
- Henkilöstöhaasteet
    - Henkilöstövaihdokset
    - Henkilöstön tiedon ja kokemuksen puute
    - Manuaalinen datan syöttäminen ja käsittely
    - Muutosvastarinta
    - Vastuunottaminen omasta työpanoksesta
- Teknologiahaasteet
    - Monen eri järjestelmän käyttäminen
    - Datan yhdistäminen ja integraatiot

Datan hallintamallin tavoitteena on vastata näihin haasteisiin. Pelkästään datan ylläpito, puhdistaminen, varastoiminen ja turvaaminen vaatii huolellisten toimintaohjeiden ja -politiikan määrittelyä. (Laihonen ym. 2013.) Kokonaisuuden tavoitteena on selkeyttää osa-alueet ja vastuut sekä ottaa huomioon dataan liittyvät lait ja määräykset. Datan hallintamalli pyrkii varmistamaan, että organisaatiolla on käytössään dataa, jota pystytään hyödyntämään päätöksentekoon. (Bhansali 2013.)

3.3 Datan laadun merkitys liiketoiminnalle

Datan laatu on monitahoinen käsite. Se voi kuitenkin olla yksi kriittisimmistä tekijöistä esimerkiksi IT-projektin epäonnistumiseen, toimialalla erottautumiseen tai kustannustehokkuuden määrittämiseen. Datan laadun tavoitteena on tukea yritystä sen prosesseissa, parantaa tiedolla johtamista ja auttaa yritystä pääsemään tavoitteisiinsa. Yritys, osastot, työntekijät ja sidosryhmät voivat hyötyä merkittävästi hyvästä datan laadusta, mutta se vaatii koko organisaation sitoutumisen hyvän datan kulttuuriin. (Bhansali 2013.)

Datan laatu on aina kytköksissä liiketoiminnan tavoitteisiin, ja sillä on suora yhteys organisaatioiden kannattavuudelle, asiakastyytyväisyydelle ja uusiutumiskyvylle. Nykypäivänä organisaatioiden dataa hyödyntää myös eri sidosryhmät kuten asiakkaat, yhteistyökumppanit ja toimittajat. Tämä lisää painetta datan laadun alkuperän, tuotantoketjun ja tarkoituksenmukaisuuden tuntemiselle. (Seppälä 2021.)

Datan laadukkuuden näkökulma voi hyvinkin vaihdella eri osastojen sisällä yrityksessä: tietoliikenneosasto näkee asiat omalta kannaltaan, kun taas liiketoiminta ja loppukäyttäjät näkevät asian omalta kannaltaan (Bhansali 2013). Jo tämä voi aiheuttaa ristiriitoja datan laadun määrittämiseen. Kommunikaation ja yhteistyön täytyy sujua läpi organisaatiotasojen, yksiköiden ja osastojen. Datan laatuvaatimus ei voi kohdistua pelkästään yrityksen tietoliikenne- tai dataosastolle, vaan näkemystä tarvitaan läpi koko liiketoimintaprosessin aina liiketoimintajohdosta loppukäyttäjiin ja teknologiaosastolle. (Sebastian-Coleman 2013.)

Konsulttina toimiva Hannu Hannila CGI:ltä kertoo blogitekstissään ”Data-driven alkaa sanalla DATA” (2020), kuinka data tulisi nähdä yrityksissä omaisuuseränä eikä teknologisena varantona, kuten liian usein tehdään. Data, sen hallinta ja hyödyntäminen voivat olla yritykselle jopa ainoa keino erottua kilpailijoista. Data-strategiassa onnistuminen voi näin ratkaista liiketoiminnan onnistumisen tai epäonnistumisen. (Hannila 2020.)

Organisaatioiden on tärkeä ymmärtää datan laadun liiketoiminnallinen merkitys ja datasidonnaisuus. Data on organisaatiolle strateginen ja taloudellinen voimavara, josta tulee huolehtia. (Korpela 2018.)

3.3.1 Datan arvon rakentuminen

Kun pohditaan huonolaatuisen datan vaikutusta yrityksen menestykseen, on hyvä ensin ymmärtää datan arvon rakentuminen. Datan tuoma arvo voidaan laskea yksinkertaisesti vähentämällä saavutetusta liiketoiminta hyödystä siihen kohdistuneet kustannukset. Taloudellinen arvo siis syntyy, kun datan käytöstä saatu hyöty on suurempi kuin sen suunnittelusta, hankinnasta, ylläpidosta ja käytöstä aiheutuvat kustannukset. Vasta sitten, kun dataa aloitetaan käyttämään ja hyödyntämään, se alkaa tuottamaan arvoa. Sitä ennen sillä on vain potentiaalista arvoa. (English 1999.)

Huonolaatuisen datan voidaan nähdä vaikuttavan yritysten tulokseen kahta eri reittiä. Suorat kustannukset syntyvät kaikesta siitä ylimääräisestä työstä, joka seuraa huonosta datan laadusta. Tähän kuuluvat niin yrityksen prosessivirheissä syntyvät kustannukset, kuin datan korjaamiseen liittyvät kustannukset. Prosessivirheitä ovat esimerkiksi yrityksen prosessien häiriintymiset. Nämä saattavat pahimmillaan keskeyttää yrityksen toiminnan. Toinen reitti syntyy menetetyistä mahdollisuuksista. Puuttuva tieto voi johtaa tilanteeseen, jonka vuoksi ei pystytty tai ehditty reagoimaan asioihin. (English 1999.)

Mikäli organisaatio kykenee minimoimaan datan elinkaareen, kuten datan tuottamiseen ja ylläpitämiseen liittyviä kustannuksia, pystyy se samalla nostamaan datasta saatavaa arvoa. On kuitenkin huomioitava, että datan elinkaaren hallintaan

ja sen suunnitteluun liittyvistä kustannuksista ei kannata tinkiä. Datan elinkaaren tehokkuutta voidaan toteuttaa esimerkiksi laatutoimenpiteiden automatisointien avulla. Hyvin suunniteltu data-arkkitehtuuri itsessään vähentää dataan liittyviä kustannuksia ja samalla maksimoi tietojen hyödyntämisen liiketoiminnan hyödyksi. (English 1999.)

Parhaiten huonolaatuisen datan kustannukset pystytään osoittamaan konkretisoimalla jokin organisaation dataan pohjautuva prosessi ja siinä huonolaatuisen datan vuoksi syntyvät kustannukset. Usein konkretia on avain muutokseen ja datan laadun parantamiseen. Jos pystytään osoittamaan kustannus-hyöty-analyysin pohjalta, että hyöty on kehittämiseen menevien kustannusten rinnalla suurempi tietyllä aikavälillä, päästään usein askel lähemmäksi dataprosessien johdonmukaista kehitystä. (McGilvray 2008.)

3.4 Laadun kustannukset

Huono datan laatu lisää merkittävästi muun muassa IT-osaston kustannuksia. Dataa joudutaan esikäsittelemään, täydentämään ja korjaamaan jälkikäteen, mikä aiheuttaa ylimääräistä työtä. Pelkästään datan etsimiseen voi tuhlaantua merkittävästi aikaa päivittäisestä työajasta. (Bhansali 2013.)

Datan laatuvaatimuksissa täytyy ottaa huomioon asiayhteys ja datan merkitys. Joissain tapauksissa täydellisyyden tavoittelu voidaan luokitella liian kalliiksi. Tässä yhteydessä voidaan tehdä päätös, ettei pyritä täydelliseen vaan toleranssin rajoissa hyväksytään laatuvirheet tai puutteet. Jos tarpeeksi hyvä riittää, ei siihen kannata tuhlata ylimääräisiä resursseja. (Mahanti 2019.)

Datan laadun heikentymisestä puhutaan silloin, kun tietoja ei päivitetä systemaattisesti ja näin ollen ne eivät enää vastaa reaalimaailman tilaa. Yksinkertaisena esimerkkinä tästä voidaan pitää tuotteiden hintamuutoksia. Mikäli muutosta tuotteen hintaan liittyen ei päivitä tietokantaan oikea-aikaisesti, datan laatu heikkenee. Heikentynyt laatu vaikuttaa pahimmillaan läpi organisaation. Myyjät myyvät tuotteita väärillä hinnoilla tai asiakastyytyväisyys kärsii, kun odotettu hinnan alennus ei tapahtunutkaan odotetusti. (English 1999.)

Valitettavan usein jokin datan huonolaatuisuuden takia tapahtunut yrityksen kriittisiin prosesseihin kohdistunut ongelma paljastaa ongelman suuruuden, ja käynnistää kehitysprojektin datan laadun ympärillä. Kriittisiä ongelmia voivat olla esimerkiksi tuotantolinjan pysähtyminen, tuotteiden toimitusten häiriöt tai pahimmassa tapauksessa koko liiketoiminnan pysähtyminen. Tapauksen jälkeen johto haluaa varmistaa, ettei niin pääse käymään uudelleen. (McGilvray 2008.)

Oleellista datavirheiden korjaamisessa on tunnistaa todelliset syyt virheiden takana. Miksi ongelma pääsi tapahtumaan? Mikä aiheutti tämän lopputuloksen? Ongelman syntysija selviää seuraamalla datan kulkua läpi datan elinkaaren ja selvittämällä, missä ongelma on tapahtunut ensimmäisen kerran. Ongelman syy-seuraussuhde tulisi kuvata mahdollisimman tarkasti. (McGilvray 2008.)

Huonolaatuisen datan pohjalta voi syntyä virheellisiä päätöksiä ja toisaalta luottamus johdon tekemiin päätöksiin saattaa vähentyä. Huonolaatuinen data voi johtaa menetettyihin tuloihin ja se voi aiheuttaa merkittäviä mainehaittoja. Mitä pidemmälle huonolaatuinen data pääsee prosessissa kulkemaan, sitä enemmän sitä ehditään käyttämään ja virheet kertaantuvat. Varhaista korjausta voidaan siis pitää halpana korjauksena. (Korpela 2018.)

Tekoälykirjailija Antti Merilehto toteaa, ettei data ole koskaan valmista vaan se on jatkuvassa muutoksen ja kehittymisen tilassa (Merilehto 2019). Datan hyvä laatu syntyy siis systemaattisen kehityksen kautta, ja huolehtimalla koko elinkaaren aikaisista tapahtumista. Seuraavassa kappaleessa kuvataan miten datan tietovarastointi vaikuttaa datastrategiaan ja -prosesseihin.

4 TIETOVARASTOT JA DATAN LAATU

Datan varastointi, siirtäminen ja muokkaaminen vaikuttavat sen laatuun ja sitä kautta datasta saatavaan lisäarvoon. Datan keräämiseen on usein organisaatioissa tehty merkittäviä satsauksia, panostamalla tietojärjestelmiin, laitteisiin ja ihmisten kouluttamiseen. Haasteeksi muodostuu usein se, ettei kerätty data ole helposti saatavilla ja yhdistettävissä monimutkaisten ja laajojen tietojärjestelmien tietokannoista. Datan hajaantuminen puolestaan osaltaan edesauttaa datan laadullisten ongelmien syntymisessä. Hajallaan oleva data johtaa tietolähteiden hankalaan yhdistettävyyteen ja siihen, ettei datan kuvaamiseen ja muuhun elinkaaren hallintaan olla panostettu riittävästi. Aina ei siis tiedetä, mitä dataa on olemassa, mitä datamassat tarkoittavat tai miten ne rakentuvat. (Hovi ym. 2009, johdanto, 4–7.) Tässä kappaleessa pohditaan, tarvitaanko organisaatiossa tietovarastoa ja kuinka tietovaraston arkkitehtuuri ja rakenne vaikuttaa datan laatuun.

4.1 Tarvitaanko tietovarastoa?

Tietovarastoinnin (DW, Data Warehouse) ja älykkäiden raportointiratkaisujen (BI, Business Intelligence) avulla data voidaan koota yhteen ja jaella eri tarpeisiin helposti. Tietovarastoon voidaan tuoda organisaation hyödynnettäväksi johdettua informaatiota ja tunnuslukuja, ja näin päästä lähemmäs kokonaisnäkemystä; tilannetta, jossa kaikki katsovat samoja arvoja ja tunnuslukuja. Tietovarasto toimii myös yrityksen historiadatan tallennuspaikkana. Sitä voidaankin pitää niin sanotusti organisaation muistina, joka mahdollistaa erilaisten trendianalyysien toteutuksen. (Hovi ym. 2009, johdanto, 24.)

Tietovaraston rakentamiseen on useita tapoja. Helpoin tapa on rakentaa paikallinen tietovarasto (data mart), jolla saadaan nopeasti näkyviä tuloksia. On kuitenkin huomioitava, että sen laajentaminen ja yhdistäminen jälkeenpäin muihin datalähteisiin saattaa olla haastavaa. (Ari Hovi 2020.) Yrityksen keskitetyn tietovaraston (EDW, Enterprice Data Warehouse) tarkoitus on koota yhteen eri datalähteet ja toimia yrityksen yhteisenä tietolähteenä ja muistina. Sen rakentaminen voi olla työläämpää, mutta maksaa pidemmällä aikavälillä usein itsensä takaisin. (Ari Hovi 2020.)

Keskitetyn tietovaraston rakentamisen puolesta puhuu yleinen trendi, jossa data halutaan saada käyttöön reaaliajassa. Strukturoidun datan lisäksi dataa halutaan varastoida myös strukturoimattomana. Datan säilytysaika pitenee ja eri datalähteiden määrä myös lisääntyy vauhdilla. Uudenlaista liiketoiminnallista hyötyä on mahdollista saada, jos olemassa olevaan dataan yhdistetään ulkoisia tietolähteitä, kuten esimerkiksi tilastokeskuksen dataa. (Hovi ym. 2009, 14–15, 18.) Datan varastoinnin on monissa tapauksissa tuettava myös big dataa, jolloin niin sanotun tietojärven (data lake) käyttöönotto on perusteltua (Amazon 2020).

Tietovaraston tarvetta voidaan tarkastella esittämällä seuraavia kysymyksiä:

- Pystytäänkö nykyisen datan perusteella rakentamaan raportteja, joissa on yhdistettynä eri tietolähteissä olevaa tietoa?
- Onko datan perusteella helppoa tehdä trendianalyysejä, joihin peilataan historiadataa?
- Onko nykyistä/nykyisiä järjestelmiä mahdollista kuormittaa kyselyin ja analyysein, ilman että järjestelmän suorituskyky laskee merkittävästi?
- Onko data saatavissa ulos helposti ja onko datan rakenne selkeä?
- Pystyvätkö käyttäjät itse hakemaan tarvitsemansa datan?
- Onko data kuvattu niin, että käyttäjä pystyy sen täysin ymmärtämään? (Hovi ym. 2009, 9.)

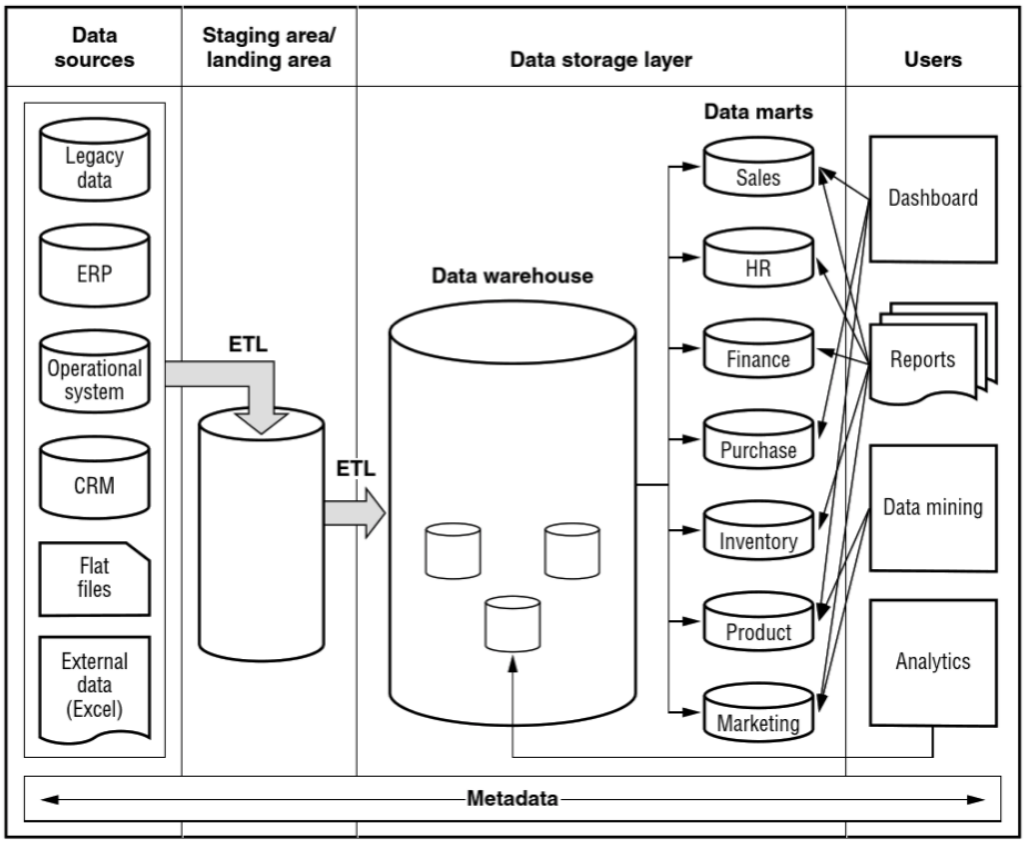
Seuraavissa kappaleessa käsitellään lyhyesti tietovaraston arkkitehtuuria, tietovaraston tietokannan mallinnusta sekä ETL- eli datan poiminta-, muokkaus- ja latausprosessia.

4.2 Data-arkkitehtuuri ja tietokannan mallinnus

Tietovarastoprojektin suunnittelussa yhdistyvät liiketoimintalähtöinen ja tekninen lähestymistapa. Toteutustavan valintaan vaikuttaa organisaation koko ja lähtötilanne, sekä tarpeet nyt ja tulevaisuudessa. Näkökulmat tietovaraston rakenteesta vaihtelevat organisaation eri tasoilla. Tietovarasto tulisikin jo alkujaan suunnitella tukemaan organisaation toimintaa monipuolisesti ja eri näkökulmista. Vaatimuksiin voivat sisäisten tarpeiden lisäksi vaikuttaa ulkoiset tarpeet. Tietovaraston ra-

kenne elää organisaation tarpeiden ja lähtökohtien mukaan. Pienen yrityksen tarpeisiin voi riittää kevyt ratkaisu, kun taas suurten organisaatioiden datastrategia ja tietovaraston rakenne voi olla moniulotteinen. (Törmänen 2017, 91–96.)

Tietovaraston kannan mallinnus on tärkeä osa teknisen projektin onnistumista. Jälkeenpäin huomatut puutteet teknisessä mallinnuksessa on hankalia ja työläitä korjata. Tietovaraston kantaratkaisun tulisi olla avoin, jotta sen käyttö eri välinein olisi mahdollista. Myös datan takaisinsyöttö operatiivisiin järjestelmiin tulisi olla mahdollista. Kantaratkaisulla on myös vaikutusta tietovaraston käytettävyyteen ja suorituskykyyn. Mikäli tietokantakyselyiden suorittaminen kestää liian kauan tai rakenne on liian monimutkainen, eivät käyttäjät jaksa käyttää sitä. (Törmänen 2017, 91-96.) Keskitetyn tietovaraston kannan mallintamiseen käytetään nykyisin paljon Data Vault 2.0. ratkaisua (Ari Hovi 2020). Paikallista tietovarastoa rakennettaessa edetään usein tähtimallin (star schema) tai lumihiutalemallin (snowflake schema) mukaisesti (Hovi ym. 2009; 37–39). Kuviossa 6 on esitetty tietovaraston rakentamisessa yleisesti käytetty rakenne.



Kuvio 6: Tietovaraston perusrakenne (Mahanti 2019, 63)

Kuten kuvio 6. osoittaa, tietovarastossa oleva data on usein peräisin eri tietolähteistä. Datan poiminta toteutetaan usein tähän tarkoitukseen suunnitelluilla ohjelmilla, joiden tehtävänä on hakea, muokata ja ladata data tietovarastoon (ETL-prosessi). Tämä tapahtuu tilannekannassa (staging area). Data varastoidaan tietovarastoalueella, joka usein sisältää keskitetyn tietovaraston. Dataa voidaan hyödyntää suoraan sen kautta, mutta sitä voidaan käyttää myös paikallisten tietovarastoratkaisuiden datalähteenä (data mart). Metadatan tallentaminen on oleellisessa osassa tietovaraston rakennetta (Mahanti 2019, 63-64.)

4.3 Metadatan rooli tietovarastossa

Tämän opinnäytetyön aiemmissa kappaleissa sivuttiin metadatan merkitystä datan laadun kannalta. Koska tietovaraston rooli yrityksen liiketoimintapäätöksissä on oleellinen, on tärkeää ymmärtää, mitä dataa tietovarastossa on ja mistä ne ovat peräisin. Juuri metadatan avulla tietovaraston datamassat pysyvät hallinnassa. (Hovi ym. 2009, 43.)

Metadata jaetaan tavallisesti kolmeen kategoriaan: liiketoimintametadataan, tekniseen metadataan ja prosessimetadataan. Metadatassa määritellään tavallisesti datan nimi, tietotyyppi, pituus, arvojoukko ja se, mistä järjestelmästä tieto on peräisin. Datan laadun kannalta on oleellista myös tietää koska data on päivitetty ja millainen laskukaava datan takana on. Datan omistaja määritellään myös metadatassa, samoin kuin käyttöoikeudet ja tietovastaava. Erillisten taulujen osalta metadata kertoo taulun nimen, kuvauksen, datan lähteen ja summataulujen osalta taso- ja summausperiaatteet. (Hovi ym. 2009, 43.)

Metadatan olemassaolo ja oikeellisuus on tietovarastoinnin, ETL-prosessin sekä datan laadun näkökulmasta oleellista. Data-asiantuntija Konsta Rönkkö IBM:ltä kertoo webinaarissa ”Kokemuksia datahankkeista & moderni analytiikka-alusta” (2020), että isoilla yrityksillä on keskimäärin käytössään yli kuusi pilveä ja näissä yli 1000 eri pilvisovellusta. Analyyseissa käytetään keskimäärin 33 datalähdettä”. Onkin selvää, että metadatalla on suuri merkitys jatkuvasti päivittyvien datamas-

sojen hallinnassa, eri tietolähteiden yhdistämisessä ja näin datan laadun hallinnassa (Hovi ym. 2009, 44). Seuraavassa kappaleessa kerrotaan tarkemmin ETL-prosessista ja sen vaikutuksista datan laatuun.

4.4 ETL-prosessi

Tietovarastoprojektissa operatiiviset tietojärjestelmät ovat tärkeässä roolissa, sillä data kertyy niihin ja niissä tiedot pidetään ajan tasalla. ETL-prosessi on tietovarastoprojektin vaihe, jossa data poimitaan (extract) eli ladataan operatiivisista tietojärjestelmistä tai muista tietolähteistä tietovarastolle sopivaan muotoon. Dataa usein myös muokataan (transform) esimerkiksi yhdenmukaistamalla ja yhdistelemällä niitä keskenään. Lopuksi valmiiksi muokattu data ladataan (load) tietovarastoon käyttöä varten. On todettu, että ETL-prosessi syö tietovarastohankkeen työajasta jopa 60–80 %. (Hovi ym. 2009, 14, 48.)

4.4.1 Datan poiminta (Extract)

Tietovarastohankkeen kannalta on olennaista päättää, mitä dataa luetaan tietovarastoon. Nykyisin on tavallista, että tietovarastoon halutaan ottaa mukaan lähes kaikki kerätty data, sillä tallennustila on aiempaa edullisempaa. Toisaalta tietovarastohankkeen koko kasvaa. Sellaista dataa, jota kukaan ei tarvitse, ei aina kannata ottaa tietovarastoon. Datan kattavuus on kuitenkin arvioitava laajemmassa ja pidempi aikaisessa perspektiivissä. (Hovi 2009, 32.)

Datan karkeisuuden, eli yksityiskohtaisuuden määrittäminen latauksen yhteydessä on datan laadun ja tietovaraston tulevan käytön kannalta oleellista. Joissain tapauksissa on järkevää lukea data summatietoina, jolloin luettavien rivien määrä pienenee. Nykyisin suositaan kuitenkin hienojakoisinta tallennustasoa, jolloin datasta saadaan myös tulevaisuudessa irti juuri halutut tiedot. (Hovi 2009, 34.)

Datan ajantasaisuus ja historiointi ovat datan laadun kannalta merkittävässä roolissa. Yleensä datan ajantasaisuus on operatiivisten tietojärjestelmien asia, mutta niissä harvemmin säilytetään historiadataa. Sen tallennus jää tietovaraston tehtäväksi. Datan poiminnassa tulisi olla selvillä, ladataanko kaikki data päivitysten yhteydessä, vai vain ne osat, jotka ovat muuttuneet sitten viime poiminnan. Datan lukemiseksi käytetään veto- tai työntömenetelmää. Joka tapauksessa on suositeltavaa, että datan poiminnan suorittavat kohdejärjestelmän ylläpitäjät. (Hovi 2009, 34, 55, 57.)

4.4.2 **Datan muokkaus (Transform)**

ETL-prosessilla on myös muita suoraan datan laatuun vaikuttavia työvaiheita. Muokkausvaiheessa muun muassa tarkastetaan ja muunnetaan data ennalta sovittuun muotoon, ennen sen lataamista käyttäjien saataville. Datalle tehdään myös oikeellisuustarkastuksia, joilla pyritään saamaan virheellistä dataa kiinni. Tarkastusten tehtävänä on muun muassa havaita tuplatietueita ja pakollisten kenttien puuttuvia ja tyhjiä arvoja, sekä raportoida niistä. Myös lukumäärätarkastukset ja erilaiset muototarkistukset kuuluvat muokkausvaiheeseen. Muototarkistuksia voivat olla esimerkiksi postinumeron muodon tai sallittujen päivämäärien tarkastaminen. Raja-arvo tarkistuksia on syytä tehdä esimerkiksi ikäsarakkeelle. (Hovi ym. 2009, 15–17, 56–57.)

Muokattaessa dataa, kiinnitetään huomiota eri järjestelmien välisien koodien yhdenmukaistamiseen ja muokataan ne helpommin ymmärrettävään muotoon. Esimerkiksi henkilön sukupuoli on voitu koodata eri tavoin eri järjestelmissä, jolloin ETL-prosessin tehtäväksi jää näiden yhdenmukaistaminen. Data myös muokataan tulevaa käyttöä varten paremmin hyödynnettävään muotoon. Esimerkiksi henkilötunnuksesta voidaan ETL-vaiheessa jalostaa syntymäaika, ikä tai ikäryhmä, joka helpottaa datan tulevaa käyttöä. (Hovi ym. 2009, 15–18.)

Postinumeron perusteella voidaan hakea postinumeroalue omaan sarakkeeseensa. Kun ikäryhmät ovat valmiiksi luotu tietovarastoon, ohjaa sen se organisaation eri osastoja käyttämään raporteissaan yhtenäisiä ikäryhmiä. Kaikki tämä

helpottaa ja nopeuttaa raportointia ja analyysien tekemistä, jolla on suora vaikutus työn tehokkuuteen. ETL-prosessi myös mahdollistaa tietojärjestelmistä erikseen johdettavien, esimerkiksi tuotekategorioiden luomisen. Tämä voidaan helpoimmillaan toteuttaa taulukkolaskelmaa hyödyntämällä ja lukemalla sen yhtenä datalähteenä tietovarastoon. (Hovi ym. 2009, 15–18.)

4.4.3 **Datan lataaminen (Load)**

Latausvaiheessa poimittu ja muokattu data ladataan tietovarastoon. Lataus on suunniteltava huolella ja esimerkiksi käsittelysäännöistä on luotava tarkat kuvaukset. ETL-prosessi voidaan toteuttaa ohjelmoimalla tai siihen tarkoitetuilla ETL-välineillä. ETL-välineiden etuna on se, että ne pakottavat yhdenmukaiseen toteutustapaan, metadatan dokumentointiin ja ne sisältävät valmiit rajapinnat datan lukemiseen ja lataamiseen eri tietolähteistä. Latausajot voivat toimia reaaliaikaisesti tai ajoittua esimerkiksi ilta- tai yöaikaan. (Hovi ym. 2009, 58.)

Kokonaisuudessaan tietovarastoinnilla ja siihen liittyvällä ETL-prosessilla on keskeinen rooli datan laatuprosesseissa ja koko data-arkkitehtuurissa. Seuraavaksi syvennytään siihen, miten datan laatua voidaan kuvata eri ulottuvuuksien kautta ja miten datan laatua tulisi mitata.

5 DATAN LAADUN OMINAISUUDET JA MITTAAMINEN

Datan laadun ominaisuuksien ja ulottuvuuksien ymmärtäminen antaa pohjan datan laadun mittaamiselle sekä laatuprosessien kehittämiselle. Tässä kappaleessa kootaan yhteen ne tärkeimmät datan laadun ulottuvuudet, joiden avulla pystytään muodostamaan kokonaiskäsitys datan laadusta.

5.1 Datan ulottuvuudet

Datan ominaisuudet ovat olennaisessa osassa laadukkaan datan tavoittelussa. Ominaisuuksien ymmärtäminen on välttämätöntä myös datan yhteen toimivuuden, siirrettävyyden ja uudelleenkäytettävyyden kannalta. Näiden ominaisuuksien avulla voidaan saavuttaa kustannustehokkaat ja yhteen toimivat järjestelmät myös datan osalta. (Bhansali 2013.)

Datan ominaisuuksiksi voidaan kutsua datalle tyypillisiä ulottuvuuksia. Ulottuvuuksia datalle voidaan löytää useita, jopa kymmeniä, mutta tärkeintä on ymmärtää oman yrityksen tarpeet ja suhteuttaa ulottuvuuksien määrittely käytettävän datan perusteella. (DAMA International 2010.)

Kuvio 7. Datan laadun ulottuvuudet (mukaillen Mahanti 2019, s. 77)

Kuvio 7 on toteutettu tietokirjailija Rupa Mahantin datan laatuun pohjautuvan kirjan Data Quality (2019) perusteella. Kuten siitä huomataan datan laadun ulottuvuuksia, voidaan jaotella hyvin hienojakoiselle tasolle asti. Useat ulottuvuudet limittyvät toistensa kanssa eivätkä kaikki ole relevantteja kaikelle datalle (Mahanti 2019). Tässä työssä datan laadun ulottuvuuksia tullaan tarkastelemaan datan hallintaan ja kehittämiseen keskittyneen yhdistyksen DAMA UK:n (2013) määrittelemien ulottuvuuksien kautta, jotka sopivat lähes kaikkeen dataan. Näitä ulottuvuuksia on kuusi ja ne ovat:

1. Datan kattavuus (Completeness)
2. Oikeellisuus (Accuracy)
3. Ainutlaatuisuus (Uniqueness)
4. Oikeamuotoisuus (Validity)
5. Johdonmukaisuus (Consistency)
6. Ajankohtaisuus (Timeliness)

Datan kattavuus

Täydellisyyttä eli datan kattavuutta voidaan mitata esimerkiksi sillä, onko tietyn tietueen arvo täytetty määritellyn mukaisesti, tai ovatko kaikki rivin pakolliset kentät täytetty. Täydellisyyden tavoittelussa tulee huomioida myös täytettävien dataarvojen asianmukaisuus ja siten käyttäjäystävällisyys. (Dama International 2010.)

Oikeellisuus

Tarkkuutta pidetään tärkeimpänä datan laadun ulottuvuutena. Toisinaan tämä ulottuvuus voidaan suomentaa myös tietojen oikeellisuutena tai täsmällisyytenä. Tällä viitataan siihen, missä määrin data edustaa tosielämän tilannetta. Mikäli tiedot on todettu kattaviksi, tulee datan oikeellisuus arvioida seuraavaksi. Käytännössä tietokannasta löytyvä data voi olla puutteellista tai jopa täysin epäolennaista. (Väre 2019.)

Ainutlaatuisuus

Ainutlaatuisuudella tarkoitetaan sitä, ettei sama tieto esiinny tietojoukossa kuin yhden kerran. Ainutlaatuisuudella on merkittävä rooli master datan tietueille sekä referenssidatan arvolistoille. Käytännössä tällä voidaan tarkoittaa sitä, että yhtä

asiakasta kohden saa olla vain yksi tietue järjestelmässä, tai että tuotehierarkiassa kukin tuotekategoria esiintyy vain kerran. Ainutlaatuisuuden tarkastaminen helpottuu, mikäli datan kattavuus on hyvä. (Väre 2019.)

Oikeamuotoisuus

Data täyttää sille asetetut muodolliset ja sisällölliset vaatimukset. Kun vaatimukset täyttyvät, helpottuu datan tekninen tulkittavuus ja siirrettävyys. Mikäli data ei täytä sille asetettuja vaatimuksia, voi sen käyttäminen erilaisissa yrityksen automatisoiduissa prosesseissa häiriintyä tai jopa estyä. Esimerkiksi sähköpostin muotovaatimukset ovat tarkat, jotta automaattiset viestit voidaan lähettää. Dataa voidaan pitää vaatimustenmukaisena, mikäli se täyttää kaikki sille asetetut muodolliset ja sisällölliset vaatimukset. (Väre 2019.)

Johdonmukaisuus

Johdonmukaisuudella tarkoitetaan yhden tietojoukon arvojen yhdenmukaisuutta toisen tietojoukon arvojen kanssa. Käytännössä johdonmukaisuudella varmistetaan kahden eri järjestelmän datan ristiriidattomuutta ja näiden yhdistämiskelpoisuutta. Johdonmukaisuus saatetaan helposti sekoittaa täydellisyyden tai oikeellisuuden kanssa. (Dama International 2010.)

Ajankohtaisuus

Ajankohtaisuuden ulottuvuus käsittää oikea-aikaisuuden sekä tiedon tuoreuden. Oleellisena asiana pidetään sitä, milloin data on saatavilla eli syötetty tietokantaan sekä milloin se on käytettävissä. (Mahanti 2019.)

Jokaiselle datalle tulisi miettiä mitkä ovat sen halutut ulottuvuudet ja ominaisuudet. (Redman 2001). Kuitenkin mikäli datan kattavuudessa on selkeitä haasteita, on järkevää keskittyä sen parantamiseen. Vasta sen jälkeen esimerkiksi oikeellisuuden ulottuvuus on järkevää ottaa tarkasteluun mukaan. (Väre 2019.)

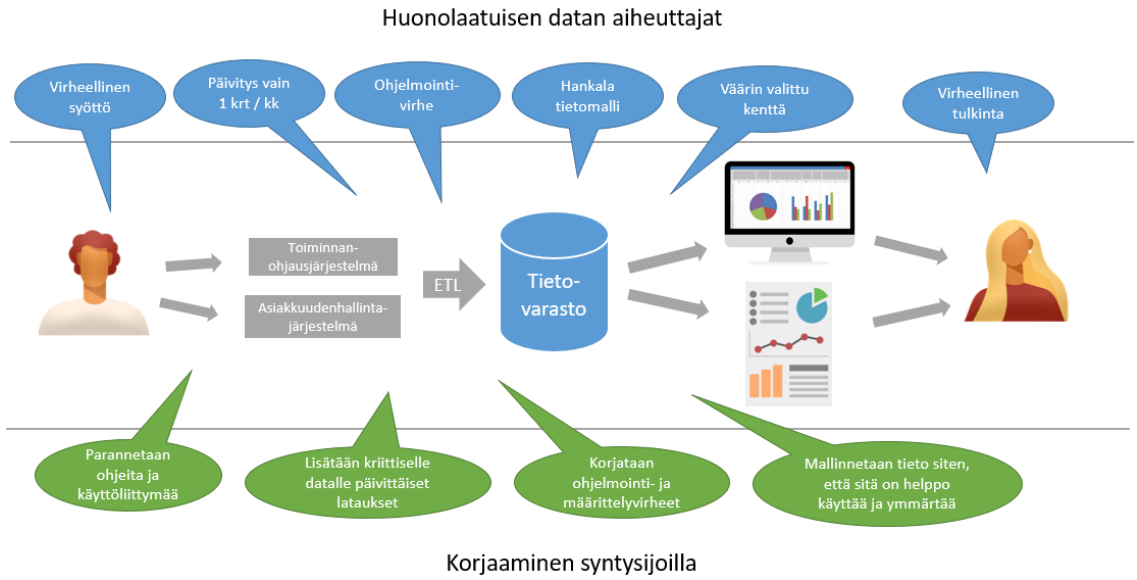
5.2 Datan laatuvirheiden syntyminen

Tekniseltä kannalta datan laatu mielletään perinteisesti IT:n ongelmana, jolloin sen tehtäväksi jää teknisten ongelmien ratkominen, kuten odottamattomien tyhjien arvojen, tuplariviesiintymien tai ylipitkien merkkijonojen ratkominen. Tämä johtaa helposti siihen, että organisaatiossa keskitytään vain tietynlaisiin datan laadun ongelmiin ja saatetaan näin kohdistaa resurssit epäoptimaalisesti liiketoiminnan kannalta. (Korpela 2018.)

Kuten aiemmissa kappaleissa jo todettiin, jotta datan laatua voidaan mitata, tulee ymmärtää ongelmien syntyperä (McGilvray 2008). Ongelmat voidaan jakaa tietojärjestelmälähtöisiin ja ihmisten aiheuttamiin laatuongelmiin. Tietojärjestelmiin liittyviä virheitä voivat olla esimerkiksi selkeät virheet ohjelmoinnissa, jotka synnyttävät virheellistä tai huonolaatuista dataa. Niitä voivat myös olla lukuvirheet, esimerkiksi tekstin, puheen tai kuvien tulkinnan ongelmien takia. Väärinymmärryksiä aiheutuu, kun dataa käsitellään väärästä lähteestä, tai jonkun välivaiheen kautta. Teknisen infrastruktuurin kuluminen voi aiheuttaa myös datavirheitä ja datan katoamisia. Siirtyminen pilvipalveluihin usein auttaa tässä ongelmassa. Organisaatioissa perinteisesti keskitytään pääasiassa järjestelmäongelmiin. (Korpela 2018.)

Ihmisten aiheuttamat laatuongelmat syntyvät usein manuaalisesti syötetystä datasta. Myös manuaalista työtä vaativat prosessivaiheet aiheuttavat ongelmia, kuten myöhästymisiä, virhesuorituksia ja henkilösidonnaisuuksia. Myös datan väärinymmärrykset ja tulkintavirheet ovat yleisiä. Käyttäjä ei välttämättä edes tiedä, mitä data kuvaa. Joissakin tapauksissa myös tahalliset väärinkäytökset voivat aiheuttaa datan laatuongelmia. (Korpela 2018.)

Datan laatuvirheet syntyvät eri vaiheissa dataprosessia. Kun tunnistetaan ongelman syntysija ja pyritään tekemään korjaus siellä, ei virhe moninkertaistu sitä mukaan, mitä pidemmälle prosessi ehtii edetä. Jotta nämä ongelmat pystytään tunnistamaan, on pystyttävä mittaamaan datan laatua. (Korpela 2018.) Kuviossa 8 on esitetty kuvitteellinen tietojärjestelmäkokonaisuus ja nostettu esiin mahdollisia virheen syntymisen paikkoja sekä niiden korjaamista mahdollisimman lähellä.



Kuvio 8. Tietojärjestelmäkokonaisuuden eri vaiheissa syntyvät dataongelmat ja niiden korjaaminen (mukaillen Korpela 2018)

5.3 Datan laadun mittaaminen

Aiemmissa kappaleissa todettiin datan laadun riippuvan organisaation vaatimuksista. Mittaamisen avulla pystytään varmistamaan datan vaatimustenmukaisuus. Kuitenkaan datan laatua ei voida mitata, mikäli sille ei ole määritelty yhteisiä ulottuvuuksia, raja-arvoja ja sekä näiden perusteella suunniteltuja mittareita. Datan laadusta ja sen mittaamisesta puhuttaessa on ymmärrettävä, että se voi tarkoittaa eri asioita eri kontekstissa ja eri organisaatiossa. Onkin oleellista määritellä yhteiset mittarit ja niiden raja-arvot. (Korpela 2018.) Yhtä tärkeää on myös tunnistaa organisaatiolle kriittinen data. Resursseja ei kannata tuhlata sellaiseen laatutyöhön, jolla ei ole selkeästi osoitettavaa liiketoimintatavoitetta. Datan laatua tulisikin kehittää siellä, missä sillä on merkitystä. (Seppälä 2021.)

Konkreettisten raja-arvomittausten määrittäminen on oleellinen osa datan laadun arviointia (Korpela 2018). Kuudelle DAMA UK:n (2013) painottamalle ulottuvuudelle on määritelty mittaristo, joka tuottaa vertailukelpoisia tuloksia datan laadun arviointiin (Korpela 2018). Seuraavassa näistä tiivistelmä.

- **Datan kattavuutta** voidaan mitata vertaamalla olemassa olevaa dataa täysin kattavaan dataan. Tästä esimerkkinä voi olla yrityksen asiakasmäärä. Mikäli todellisuudessa asiakkaita on 100, mutta järjestelmästä löytyy vain 85 asiakasta, on datan kattavuus sillä osa-alueella vain 85 %. Tämän mittarin kohdalla on tärkeää määritellä mitä täysin kattava tarkoittaa kunkin dataelementin kohdalla. Valmiin vertailuarvon löytyminen voi olla haaste. (Korpela 2018.)

- **Datan ainutlaatuisuutta** mitataan arvioimalla, löytyykö jokin todellisen maailman asia datasta useammin kuin kerran. Mittariksi voidaan näin johtaa tosiasioiden lukumäärä jaettuna datassa esiintyvien havaintojen lukumäärällä, ja laskea arvo prosentteina 0–100 % välillä. Mittaamisesta hankalaa tekee se, että mittari tarvitsee rinnalleen vertailtavaa dataa. (Korpela 2018.)

- **Ajankohtaisuudella** tarkoitetaan viivettä datan syntyajankohdan ja todellisen tapahtumahetken välillä. Mittari voidaan esittää esimerkiksi minuutteina tai päivinä. (Korpela 2018.)

- **Oikeellisuutta** mitataan vertaamalla sitä osuutta datasta, joka on muodoltaan oikein, kaikkeen mittauksen kohteena olevaan dataan. Oikeamuotoisuudella tarkoitetaan määriteltyä datatyyppiä ja sille asetettua raja-arvoa. Oikeellisuuden mittari voi siis antaa arvon siitä, kuinka monta prosenttia datasta on oikein. (Korpela 2018.)

- **Tarkkuutta** voidaan mitata tarkastelemalla, kuinka hyvin data kuvaa todellisen maailman tilannetta. Mittaria luotaessa on sovittava vaihteluväli, jolla data on organisaation mielestä riittävän tarkkaa. Mittaus voidaan toteuttaa vain tietylle osalle datajoukkoa ja tehdä sen perusteella johtopäätöksiä datajoukon tarkkuudesta. Tarkkuuden mittari voi olla raja-arvojen sisään mahtuvan datan osuus prosentteina. Käytännössä mittaus voidaan suorittaa mittaamalla manuaalisesti todellisen maailman arvo ja vertaamalla alkuperäistä mittaustulosta siihen. (Korpela 2018.)

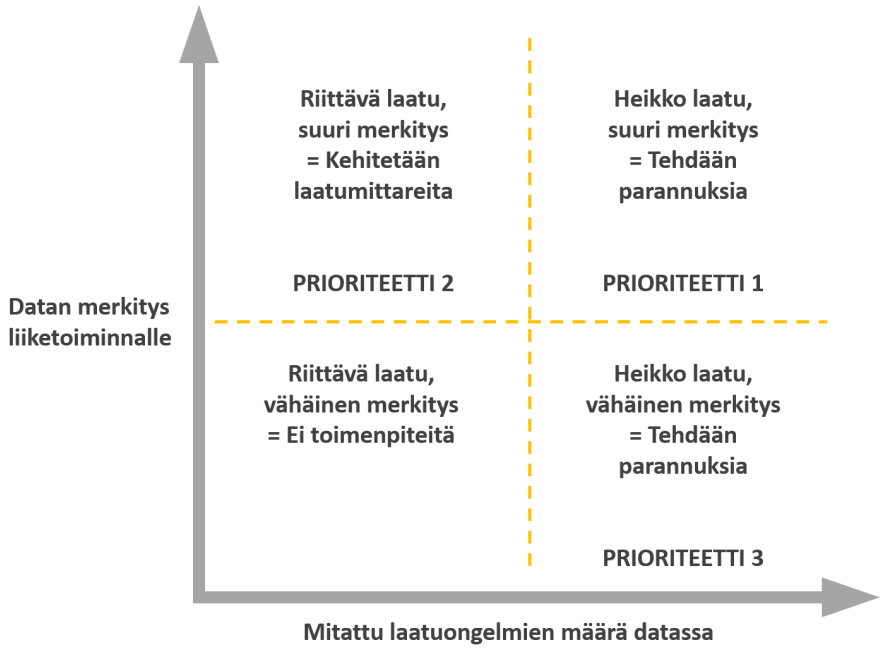
- **Johdonmukaisuutta** voidaan mitata arvioimalla, onko data saman sisältöistä eri lähteissä tai datajoukoissa. Tällä tarkoitetaan esimerkiksi sitä, onko erilaisten järjestelmien välillä olevat tiedot samat, kuten osoite. Johdonmukaisuus voidaan myös tulkita datan käyttäytymiseen liittyvänä, esimerkiksi aineiston jakaumana. Mikäli tiedetään, että tietyn mittauksen arvon tulisi kutakuinkin pysyä samansuuntaisena, mutta datan jakaumasta huomaamme, että meillä on poikkeama-arvoja, eli arvoja, jotka eivät yllä oletetun jakauman sisään, voidaan tulkita, että data ei käyttäydy johdonmukaisesti. Tästä voidaan johtaa mittari, kuinka monta prosenttia datasta mahtuu normaalijakauman sisään. (Korpela 2018.)

Datan laadun mittareita voi olla muitakin, kuin edellä mainitut. Tärkeintä on pystyä luotettavasti mittaroimaan datan laatua niiden osa-alueiden osalta, joilla kyseiselle organisaatiolle liiketoiminnallisesti merkitystä. Kun organisaatio on saanut määriteltyä riittävän laaduntason, voidaan sille asettaa mittari ja lopulta laatua mittaava prosessi voidaan automatisoida. Datan laadun parantaminen on jatkuva ja kehittyvä prosessi. (Korpela 2018.)

Hyvälaatuinen data ei siis tarkoita pelkkää datan virheettömyyttä. Hyvälaatuinen data kattaa kaikki oleelliset havainnot ja niiden tiedot. Data tulee saada käyttöön riittävän nopeasti ja mittausten tulee olla riittävän tarkkoja, eikä data saa sisältää kriittisiä virheitä. Se on sisällöltään ja muodoltaan helposti ymmärrettävää, eikä se saa sisältää ristiriitoja. (Korpela 2018.)

5.4 Toimenpiteiden priorisointi ja datan omistajuus

On todennäköistä, että kaikkea organisaation dataa ei pystytä saamaan laadullisesti hyväksyttäviin rajoihin. Tämän takia laadunparannustoimenpiteiden priorisointi on oleellista. Kuviossa 9 on havainnollistettu, kuinka datan toimenpiteiden tärkeysjärjestystä tulisi arvioida. Dataa, jolla ei ole organisaatiolle suurta merkitystä, ei kannata ensimmäisenä lähteä korjaamaan. Organisaatiolle merkityksellisen datan parantaminen tulisi olla keskiössä. On kuitenkin huomioitava, että tällä hetkellä merkityksetön data voi myöhemmin nousta merkitykselliseksi. (Korpela 2018.)



Kuvio 9. Datan laadun korjaamisen priorisointi (mukaillen Korpela 2018)

Datalla tulisi aina olla omistaja. Omistajia voi olla organisaation sisällä useita, riippuen datan luonteesta. Omistajuuden myötä mielenkiinto ja panostus laatuun kasvavat. Datan omistajan vastuu on määritellä laadun vaatimustaso, varmistaa laadun seuranta sekä kehittää olemassa olevaa dataa. Omistaja on myös vastuussa datan puutteista. Datan omistajan ei tarvitse ymmärtää teknisiä ratkaisuja, mutta hänen on oltava niistä tietoinen. (Korpela 2018.)

Tiivistettynä datan laadunhallinnan prosessi (Korpela 2018):

1. Mallinnetaan olemassa olevat datat
2. Tunnistetaan mikä data on organisaatiolle tärkeää
3. Määritellään datan laadun mittarit ja niiden raja-arvot
4. Mitataan datan laatu
5. Korjataan ongelmia priorisoinnin mukaan tärkeimmästä alkaen
6. Organisoidaan jatkuva laadun tarkkailu

6 DATAN LAATU JA TEKOÄLY

Tässä kappaleessa käsitellään datan laadun vaikutusta tekoälyn hyödynnettävyyteen. Aluksi kuvataan lyhyesti tekoälyä käsitteenä ja tutustutaan koneoppimisen eri muotoihin. Koska tämän opinnäytetyön keskeisenä teemana on koneoppimisalgoritmien hyödyntäminen asuntojen hintojen ennustamisessa, keskitytään myös tässä kappaleessa regression avulla jatkuvien muuttujien ennustaviin koneoppimismalleihin.

6.1 Tekoälyn käsite

Käsitteenä tekoäly on laaja ja moniulotteinen, eikä sille ole yhtä yleisesti hyväksyttyä määritelmää (Elements of AI 2018). Tekoälyä voidaan kuitenkin kuvailla älykkääseen toimintaan kykenevinä koneina ja ohjelmina (Itewiki n.d.). Sille ominaispiirteitä ovat autonomisuus, eli kyky tehdä tehtäviä ilman käyttäjän jatkuvaa avustusta, sekä adaptiivisuus, eli kyky parantaa suorituskykyä oppimisen kautta (Elements of AI 2018).

Tekoäly jaetaan tyypillisesti heikkoon ja vahvaan tekoälyyn. Heikko tekoäly pystyy ratkaisemaan yhtä ennalta määrättyä tehtävää kerrallaan, kun taas vahva tekoäly jäljittelee kokonaisvaltaisesti ihmisenkaltaista älykkyyttä ja jäljittelee tietoisuutta. Tällä hetkellä kaikki tekoälysovellukset kuuluvat heikkoon tekoälyjoukkoon. (Ailisto ym. 2018.) Näin ollen tekoälylle kuvaavampia termejä voisi olla esimerkiksi tukiäly tai apuäly (Merilehto 2019).

Tekoälyä esiteltiin ihmiskunnalle jo 1950-luvulla, jonka jälkeen monia tekoälyyn pohjautuvia ratkaisuja on toteutettu usealla eri tieteensaralla. 2010-luvulla koneiden laskentatehon, datan määrän ja algoritmien kehittyessä on tekoälyn kehitys lähtenyt kasvamaan eksponentiaalisesti. Koneoppimismallit tarvitsevat dataa ja massiivisia data-aineistoja. Kehittyneet tietovarastot, datan hallinta sekä niistä koituvat varastointikustannukset ovat laskeneet merkittävästi, joka on puoltanut koneoppimisen kehittymistä. (Merilehto 2018.)

6.2 Koneoppiminen

Tekoälyn yhteydessä puhutaan koneoppimisesta. Koneoppiminen pohjautuu tilastotieteeseen ja sen voidaan mieltää olevan tiedon eristämistä datasta. (Elements of AI 2018.) Sen lähtökohtana voidaan pitää sitä, että kone suorittaa oppimisprosessin ihmisen puolesta (Joutsijoki 2017). Koneoppimisessa datan tyypit jaetaan tavallisesti luokiteltuun ja luokittelemattomaan dataan. Luokittelemattomalla datalla tarkoitetaan useimmiten rakenteetonta dataa, josta on jo aiemmin tässä opinnäytetyössä puhuttu. Kun data saa merkityksen puhutaan luokitellusta datasta. Luokittelussa datalle annetaan luokka tai kategoria mihin se kuuluu. (Gollapudi 2016.) Koneoppiminen jaetaan perinteisesti kolmeen osa-alueeseen; ohjattuun ja ohjaamattomaan oppimiseen sekä vahvistusoppimiseen sen mukaan minkä luonteisesta ongelmasta on kyse (Elements of AI 2018).

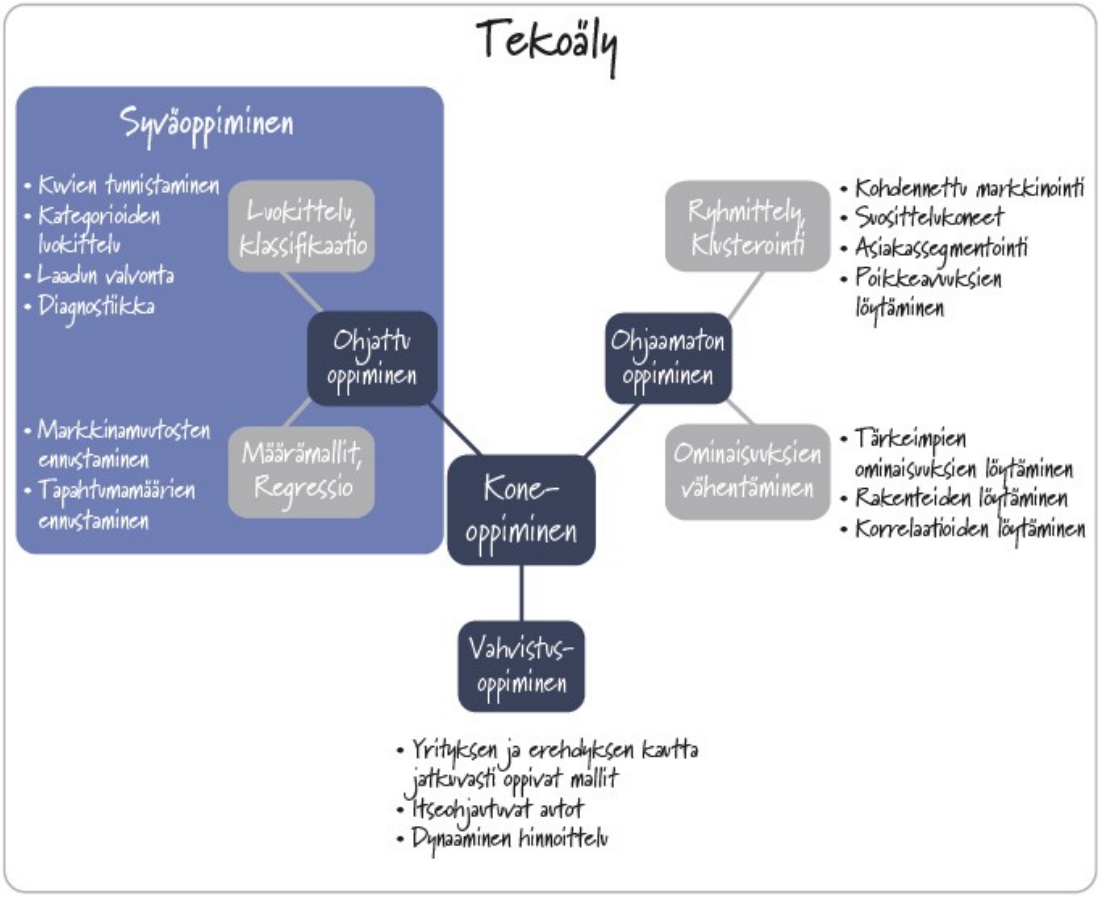
Ohjatussa koneoppimisessa hyödynnetään valmiiksi luokiteltua dataa (Gollapudi 2016). Havainnoille tehtävä luokkien määrittäminen on tärkeää tehdä huolellisesti, koska se vaikuttaa suoraan koneoppimismallin ennustuskykyyn. Luokkien määrittelyt olisi hyvä tarkastuttaa asiantuntijalla, mikäli se on mahdollista. (Joutsijoki 2017.)

Ohjatun koneoppimisen menetelmien avulla voidaan esimerkiksi tunnistaa, mihin luokkaan tietty valokuva kuuluu. Koneoppimisalgoritmille annetaan opetusaineistona valmiiksi nimettyjä kuvia, esimerkiksi valokuvia koirista vai kissoista, ja niiden perusteella se oppii tietyn luokan ominaispiirteet. Oppimansa perusteella koneoppimismalli pystyy jatkossa luokittelemaan sille entuudestaan tuntemattomia kuvia. Kun vastaukseksi halutaan vastaus kyllä tai ei, on kyse binäärisestä luokitteluongelmasta. (Elements of AI 2018.)

Asuntojen hintojen ennustamisessa ohjatulle koneoppimiselle annetaan opetusaineisoksi joukko tapahtumarivejä, joiden toteutunut myyntihinta tiedetään. Oppimansa perustella algoritmi pystyy tekemään ennusteen uuden kohteen myyntihinnasta. Tässä tapauksessa puhutaan ohjatun oppimisen regressio-ongelmasta. (Elements of AI 2018.) Esimerkkejä ohjatun oppimisen menetelmistä ovat päätöspuumenetelmät, Naiivi Bayes -luokittelija ja neuroverkkomenetelmät (Joutsijoki 2017).

Koneoppimisen ohjaamattomassa oppimisessa algoritmi pyrkii puolestaan etsimään datasta rakenteita tai ryhmiä ja näin klusteroimaan dataa (Elements of AI 2018). Esimerkkeinä ohjaamattoman oppimisen menetelmistä ovat K-means-algoritmit ja hierarkkinen klusterointi (Joutsijoki 2017).

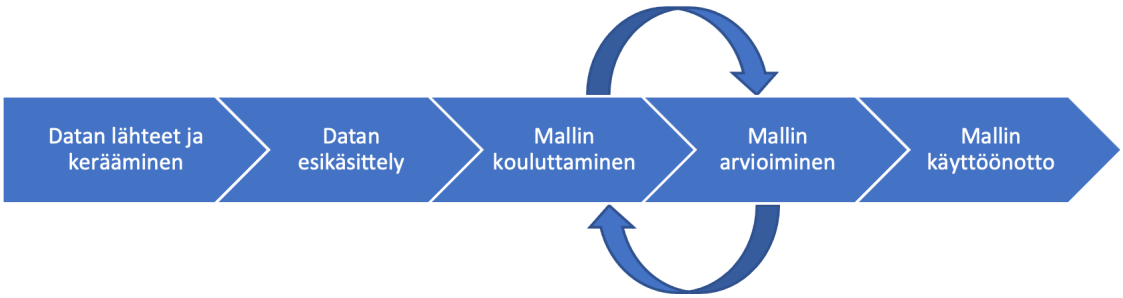
Vahvistusoppimisesta puhutaan silloin, kun ohjelmalle annetaan palautetta sen tekemien valintojen pohjalta, esimerkiksi itseajavan auton tapauksessa. Kone siis oppii saamansa palautteen pohjalta. Toisinaan saatetaan myös käyttää termiä puoliohjattu koneoppiminen, jolla tarkoitetaan koneoppimisjärjestelmää, joka hyödyntää sekä ohjatun- että ohjaamattoman oppimisen menetelmiä. (Elements of AI 2018.) Kuviossa 10 on tiivistettynä koneoppimisen muodot ja esitetty niiden tunnettuja käyttökohteita.



Kuvio 10. Tekoälyn osa-alueet (Kananen & Puolitaival 2019)

6.3 Datasta koneoppimismalliksi

Datan jalostaminen koneoppimismalliksi alkaa datan lähteiden tunnistamisesta ja keräämisestä. Seuraavaksi data valmistellaan ja esikäsitellään. Tämä prosessi on aikaa vievä, sillä esitysformaatti tai laatu voi vaihdella merkittävästi. Data voi olla rakenteellisessa tai ei-rakenteellisessa muodossa, jolloin yhteisen tietokannan muodostaminen voi muodostua jo ensimmäiseksi haasteeksi. Esikäsittelyn ja valmistelun jälkeen koneoppimismallia koulutetaan, testataan ja arvioidaan. Tätä toistetaan niin kauan, kunnes mallin todetaan olevan tarpeeksi hyvä käyttötarkoitukseen. Tämän jälkeen malli voidaan ottaa tuotantokäyttöön. (Kananen & Puolitaival 2019.) On kuitenkin huomioitava, että mallista tulee huolehtia vielä tuotantojulkaisun jälkeenkin, sillä mallin tarkkuutta voidaan joutua muuttamaan (Merilehto & Hagman 2021). Kuviossa 11 havainnollistetaan prosessia datan hankkimisesta mallin käyttöönottoon.



Kuvio 11. Koneoppimismallin prosessivaiheet (mukaillen Kananen & Puolitaival 2019)

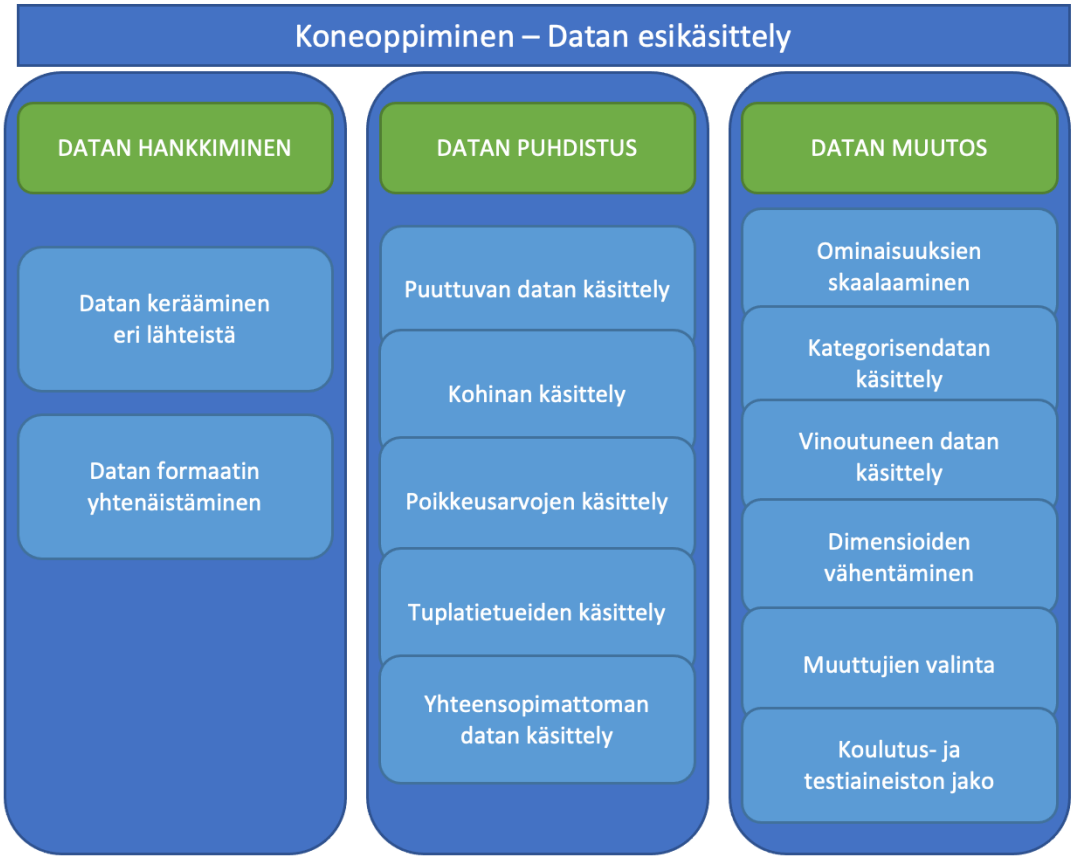
Datan laadun merkitystä ei voi korostaa liikaa, kun puhutaan koneoppimisen menetelmistä. Vaikka koneoppimista hyödyntävä ohjelmisto olisi kehitetty ja suunniteltu erinomaisesti, se voi olla vain niin hyvä, kuin sille syötetty data on ollut laadultaan. Koneoppimisprojektia voidaan verrata talonrakentamiseen. Huolellinen perustus muodostaa tukevan kivijalan, jonka päälle voidaan rakentaa seiniä ja yksityiskohtia. Mikäli maasto, johon taloa rakennetaan, on haastava, vaativat myös pohjatyöt enemmän aikaa. (Mueller ja Massaron 2016.) Seuraavissa kappaleissa tutustutaankin tarkemmin datan valmisteluun ja esikäsittelyyn sekä mallin arvioimiseen.

6.3.1 Datan esikäsittely

Datan määrä ja laatu vaikuttavat suoraan tekoälyyn ja koneoppimisen malleihin. Mitä enemmän dataa ja mitä parempi laatuista se on, sitä parempiin tuloksiin on mahdollisuus päästä. Valtaosa koneoppimisprojekteihin käytetystä ajasta kuluu esikäsittelytyöhön, kuten datan putsaamiseen ja siivoamiseen. Mikäli data olisi lähtökohtaisesti hyvälaatuista, päästäisiin toteutukseen ja tuloksiin huomattavasti nopeammin. (Merilehto 2018.)

Eräs sanonta tekoälymaailmassa kuuluu: ”roskaa sisään, roskaa ulos” (Kilkenny & Robinson 2018). Sanonnan tarkoitus on osoittaa, ettei tekoäly pysty toteuttamaan huonosta datasta hyvää koneoppimismallia. Jos koneoppimisprojektiin käytetään huonolaatuista dataa, joka on täynnä virheitä ja epäjohdonmukaisuuksia, luo se vain huonosti koulutetun koneoppimismallin, joka tuottaa merkityksettömiä tuloksia tai ohjaa prosessia kokonaan väärään suuntaan. (MLK 2019). Datan esikäsittelyn tavoitteena on parantaa huonokuntoisen datan laatua (Tamminen 2019).

Kuten kuviosta 12 huomataan, mahdollisia datan esikäsittelytoimenpiteitä on monenlaisia. Toimenpiteitä toteutetaan tarpeen mukaan (MLK 2019). Työstäminen alkaa datan hankkimisesta. Data täytyy olla saatavilla ja käsiteltävissä, jonka jälkeen siitä voidaan lähteä kehittämään liiketoimintaa eteenpäin vieviä ratkaisuja (Merilehto 2018). Usein datan keräämiseen ja yhdistämiseen tarvitaan jo toimenpiteitä, ennen kuin sitä päästään varsinaisesti esikäsittelemään (MLK 2019).



Kuvio 12. Datan esikäsittelyn toimenpiteitä (mukaillen MLK 2019)

Puuttuvien arvojen käsittely

Data voi olla puutteellista monella tapaa. Datasta voi muun muassa puuttua mittausarvoja, datan arvot on tallennettu väärin tai kokonaan kiinnostava muuttuja puuttuu kokonaan (Tamminen 2019). Useimmat koneoppimisen algoritmit eivät toimi, mikäli datassa on puuttuvia arvoja. Näin ollen on tärkeää päättää mitä puuttuvalle datalle tehdään. Data on mahdollista käsitellä muun muassa poistamalla rivi tai sarake kokonaan, korvata puuttuva tieto laskennallisella keskiarvolla, moodilla tai mediaanilla tai joissakin tapauksissa lisätä kokonaan uusi kategorinen muuttuja kuten ”*ei tiedossa*”. (MLK 2019.) Puuttuvien arvojen käsittelyyn voidaan ottaa kantaa jo datastrategiaa suunniteltaessa ja näin ollen täydentää dataa jo ETL-prosessin aikana. Siten data on jo täydennetty valmiiksi, ennen kuin sitä hyödynnetään. (Azeroual ym. 2018.)

Kohinan, poikkeusarvojen ja vinoumien käsittely

Data voi sisältää arvoja, jotka ovat epäoleellisia, virheellisiä tai jakaumasta poikkeavia (Tamminen 2019). Tällaisen datan takia mittaamiseen voi tulla epävarmuutta (Kananen & Puolitaival, 2019). Poikkeavan, vinoutuneen tai kohinaisen

datan havainnointiin ja käsittelyyn on mahdollista hyödyntää erilaisia tekniikoita. Dataa voidaan muun muassa skaalata, standardoida tai normalisoida. Helpoin tapa paljastaa poikkeavia arvoja on visualisoida dataa. Tällöin voidaan myös päättää, onko kyseessä poikkeava, vinoutunut tai kohina-arvo, ja kuinka se tullaan käsittelemään. (Hämäläinen, 2013.)

Tuplatietueiden käsittely

Tuplatietueiden osalta datasta täytyy tarkastella ja ymmärtää onko kyseessä todellakin tuplatietue vai tietue, joka on todellisuudessa esiintynyt useammin kuin kerran. Mikäli kyseessä todellakin on tuplatietue, se tulisi poistaa data-aineistosta. (MLK, 2019.)

Muuttujien vähentäminen ja valinta

Muuttujien valinta on oleellinen osa koneoppimisenprosessia. Jotta koneoppimisen algoritmit pystytään muodostamaan mahdollisimman tarkaksi, on algoritmien kehittämiseen käytettävien muuttujien valinnalla merkittävä osuus. Yleensä tarvitaan toimialaosaamista ja asiantuntijuutta, jotta oikeanlainen valinta pystytään tekemään. Muuttujien merkitystä voidaan tutkia myös selvittämällä muuttujien korrelaatiota tai käyttää matemaattista ominaisuuksien valikoimisen algoritmia. (Kananen & Puolitaival, 2019.) Toisinaan on myös oleellista muodostaa uusia muuttujia muuttamalla tai yhdistämällä olemassa olevia muuttujia (Hämäläinen, 2013).

Kategorisen datan käsittely

Kategorinen data tarkoittaa merkkijono-tyyppistä tietoa. Esimerkiksi värin esittäminen voidaan toteuttaa kategorisena datana (punainen, valkoinen). Useimmiten koneoppimisalgoritmit eivät kuitenkaan pysty käsittelemään tällaista tietoa, vaan data täytyy muuttaa numeeriseen muotoon kuitenkin siten, että datan loogisuus säilyy ehyenä. (MLK, 2019.)

Datan määrä

Koneoppimismallin kouluttaminen tarvitsee dataa, joten kysymykseen tulee, kuinka paljon dataa on tarpeeksi (Kananen & Puolitaival, 2019). Esikäsittelyssä datan määrä voi vähentyä merkittävästikin, mutta on muistettava, ettei datan määrä korvaa laatua (Tamminen, 2019). On syytä siis keskittyä hyvälaatuiseen dataan, vaikka se tarkoittaisikin datamäärän pienemistä. Tarpeeksi kattava datan

määrä riippuu ratkaistavasta asiasta ja käytettävästä koneoppimismallista. (Kananen & Puolitaival, 2019.)

Koulutus- ja testiaineiston muodostaminen

Data-aineiston esikäsittelyn viimeinen toimenpide on jakaa aineisto koulutus- ja testiaineistoihin. Koulutusdataa käytetään koneoppimismallin opettamiseen ja testausdatalla pystytään varmistamaan koneoppimismallin tarkkuus. Varsinaista datan jakosääntöä ei ole, mutta useimmiten koulutusdata on noin 70–80 % aineistosta ja testausdata 20–30 % aineistosta (MLK. 2019).

6.3.2 Opetetun koneoppimismallin arviointi

Koneoppimissovelluksia käytettäessä on hyvä tiedostaa, että sen käyttöön liittyy useita virheen mahdollisuuksia. Korkea ennustetarkkuus ei automaattisesti tarkoita parasta koneoppimismenetelmää. Mallia tulee arvioida useammalta kannalta, sillä väärin arvioidut tulokset voivat aiheuttaa vakavaakin haittaa päästessään tuotantoon. (Ruokolainen 2018.) Mallin arvioinnissa täytyy myös ymmärtää, minkälaisesta koneoppimismallista on kyse ja mitkä mittarit tai tunnusluvut ovat kyseiselle mallille oleelliset (Microsoft 2021).

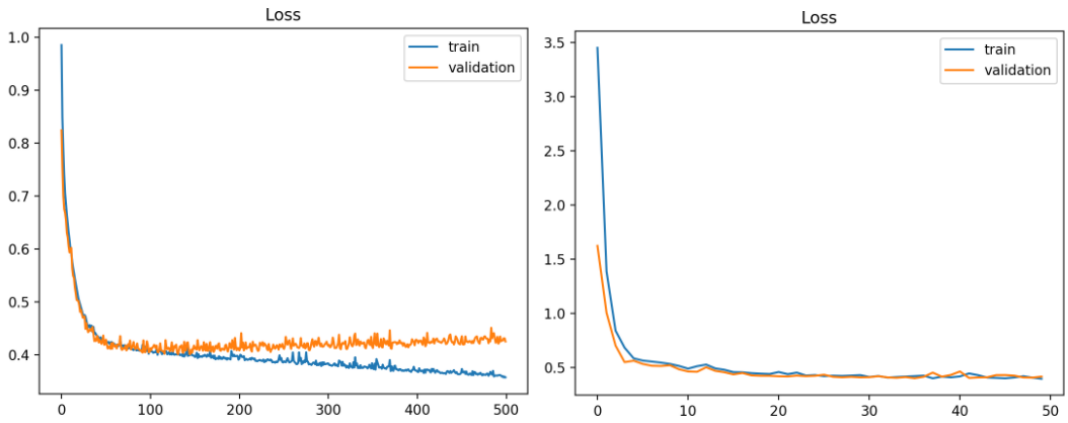
Koneoppimismallien ylisovittamisongelmaa pidetään yhtenä koneoppimisen haastavimmista ongelmista. Ylisovittamisesta puhutaan silloin, kun mallin ennustetarkkuus on opetus- ja testidataa käytettäessä toisistaan poikkeava. (Elements of AI 2018.) Yksinkertaistettuna sillä tarkoitetaan sitä, että malli oppii opetusdatan liian perusteellisesti, eikä näin pysty ennustamaan yhtä hyvin uusia havaintoja. Datajoukon tulisikin edustaa riittävän kattavasti laajempaa havaintopopulaatiota. Jotta riittävä otos havaintopopulaatiosta saadaan, datassa ei saisi olla liikaa satunnaiskohinaa, datajoukon koko ei saisi olla liian pieni, eikä eri luokkamuuttujien tasojen välillä saisi olla liikaa kombinaatioita. (Ruokolainen 2018.)

Jo aiemmin puhuttiin aineiston jakamisesta opetus- ja testidataan. Jakoa käyttämällä pystytään välttämään suurimmat koneoppimismallin opetuksessa tapahtuvat virheet. Nimensä mukaisesti opetusdatalla opetetaan malli, jotta se pystyisi

ennustamaan tuntematonta syötettä. (Elements of AI 2018.) Testidataa hyödynnetään validoimaan opetusdatalla saatuja tuloksia. Testidatan avulla voidaan siis arvioida ennustemallin tarkkuutta ja pyrkiä arvioimaan sitä, kuinka hyvin malli pystyy yleistämään oppimansa tuntemattomiin syötteisiin. (Ruokolainen 2018.)

Aina pelkkä testidatalla validointi ei riitä, vaan on käytettävä ristiinvalidointia. Ristiinvalidoinnista on monia variantteja. Yksinkertaistettuna siinä koneoppimisessa käytettävä datajoukko jaetaan useampaan opetus- ja testiaineistoon. Kun esimerkiksi kolmella aineistojaolla suoritetaan mallin opetus ja validointi, pystytään vertaamaan saatuja tuloksia keskenään ja varmentamaan saatuja opetustuloksia. (Ruokolainen 2018.)

Koneoppimismallien opetuskäyrien tarkastelu opetuksen edetessä on hyvä tapa havainnoida mallin oppimista. Oppimiskäyrästä pystytään havaitsemaan mallin yli- tai alisovittaminen, tai se, ettei malli ole ylipäätään sopiva tehtävään. Kuviossa 13 on esitetty esimerkit ylisovittavasta ja malliin sopivasta opetuskäyrästä. Kaavion vasemmanpuoleisessa reunassa on esimerkki opetuskäyrästä, joka ylisovittaa. Tämä voidaan havaita oranssin validointikäyrän suunnasta. Validointikäyrä erkaantuu sinisen opetuskäyrän todellisesta datasta. Oikeanpuoleisessa reunassa malli taas oppii hyvin, ja sininen opetusdatakäyrä sekä oranssi validointikäyrä noudattavat samaa suuntaa. Opetuskäyrä voi myös paljastaa esimerkiksi väärän suhteen opetus ja testiaineiston välillä. (Brownlee 2019.)



Kuvio 13. Koneoppimismallin opetuskäyrät (Brownlee 2019)

Lasse Ruokolainen tiivistää blogissaan ”Ylisovittaminen ja kuinka sen kanssa voi tulla toimeen” (2018), kuinka koneoppimisen sudenkuoppien kanssa voi tulla toimeen. Tärkeintä on miettiä tarkkaan etukäteen, mitä on tekemässä ja kuinka ongelmaa tulisi lähestyä.

Koneoppimismallien arviointiin voidaan käyttää useita eri mittareita, joista jokaisella on omat hyvät ja huonot puolensa. Tässä opinnäytetyössä tullaan arvioimaan koneoppimismallien suorituskykyä MAE ja R2 -arvojen perusteella. Ne ovat toimivia mittareita jatkuvaa muuttujaa ennustettaessa regressiomalleissa (Willeams 2019).

- **MAE** (Mean Absolute Error) eli keskimääräinen absoluuttinen virhe on arvo, joka osoittaa, kuinka lähellä ennusteet ovat mahdollisiin tuloksiin nähden. Mitä pienempi arvo, sitä paremmin malli ennustaa. (Willeams 2019.)

- **R2 Score**, eli niin sanottua regressiopistettä voidaan käyttää arvioitaessa regressiomallin toimintaa. Se antaa arvon 0 ja 100 % väliltä. Jos malli saa arvon 1 (100 %), tarkoittaa se muuttujien korreloivan toisiaan täydellisesti. Mikäli arvo jää alhaiseksi, muuttujat eivät korreloi hyvin mallissa, eikä näin ollen regressiomalli ei ole kelvollinen. (Willeams 2019.)

Seuraavaksi esitellään kaksi ohjatun oppimisen koneoppimismenetelmää, joita tullaan tässä opinnäytetyössä testaamaan. Nämä koneoppimismenetelmät ovat hyvin erityyppisiä ja toimivat eri periaatteilla.

6.3.3 Keinotekoiset neuroverkot

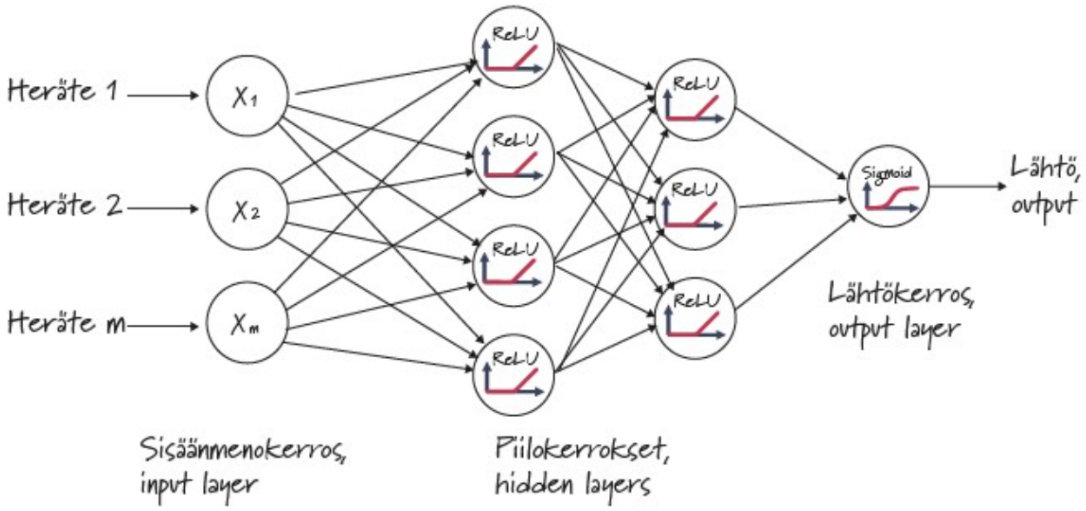
Keinotekoiset neuroverkot (Artificial Neural Networks, ANN) ovat koneoppimisen työkaluja, jotka mukailevat ihmisen aivotoimintaa. Neurotiede tutkii aivojen ja hermoston toimintaa ja pyrkii kehittämään niitä matkivia malleja ja näin parempia ratkaisuja tekoälyyn ja koneoppimiseen. Neuroverkkojen erityispiirteinä voidaan

pitää neuroverkon rakennetta, joka pystyy prosessoimaan tietoa toisistaan riippumatta ja näin käsittelemään samanaikaisesti suuria tietomääriä. (Elements of AI, 2018.)

Neuroverkkojen yhteydessä puhutaan syväoppimisesta. Yksinkertaistettuna syväoppimisessa yhdistetään yksinkertaisista prosessointiyksiköistä, eli neuroneista koostuvia kerroksia yhteen siten, että syntyy verkko, jonka läpi järjestelmän prosessoima tieto kulkee. Verkon kerroksellisuus eli neuroverkon syvyys mahdollistaa datan muodostamien rakenteiden ja säännönmukaisuuksien oppimisen. (Elements Of AI, 2018.) Neuroverkko saa eri painokertoimia, joiden perusteella eri syötteiden välinen painoarvo määrittyy. Kun neuroverkkoa koulutetaan, pyritään löytämään optimaaliset painokertoimet mahdollisimman oikean ulostulostuloarvon saavuttamiseksi. Neuroni toimii aktivointifunktion avulla. Käytetyimpiä aktivointifunktioita ovat Sigmoid ja Rectifier Linear Unit eli ReLU-funktio. (Kananen & Puolitaival 2019, 129, 132.)

Asunnon hintaa ennustettaessa sisään annettavat syötteet, eli muuttujat, ovat aineiston sisältämät tiedot myydystä kohteesta. Näitä ovat esimerkiksi postinumeroalue, asunnon tyyppi, asuinneliöt ja asunnon kunto. Datan kulkiessa neuroverkon läpi, nämä syötteet saavat erilaiset painokertoimet sen mukaan, kuinka suuri merkitys niillä on ulostuloarvolle, eli asunnon hinnalle.

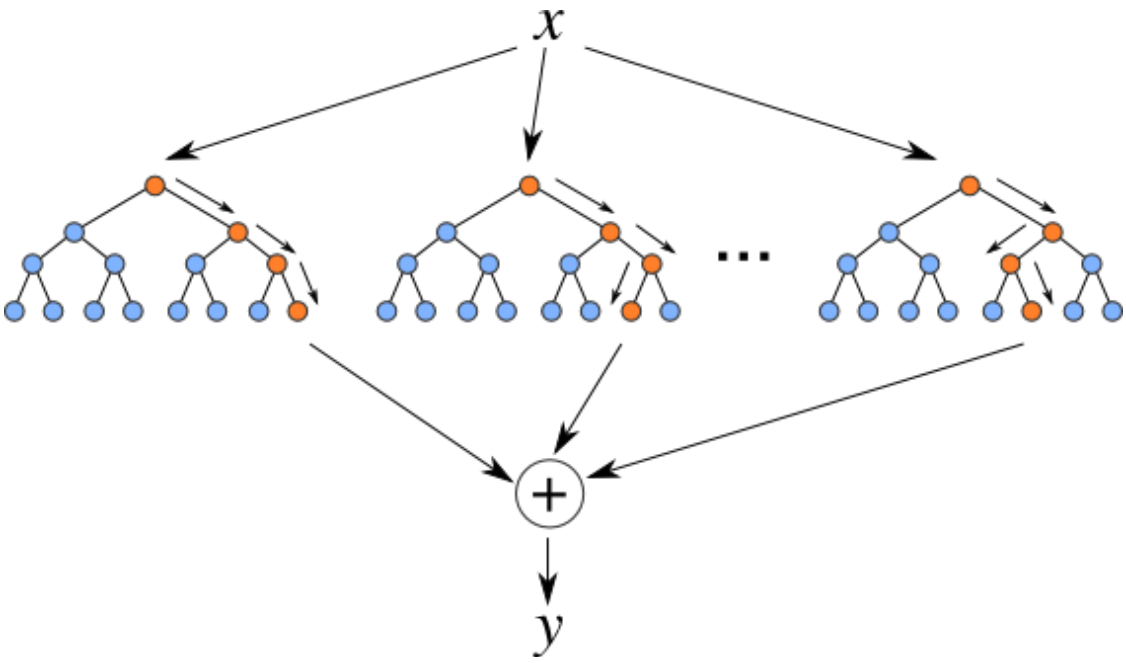
Neuroverkkoja käytetään paljon ennustemalleissa, mutta niiden avulla voidaan analysoida monentyyppistä dataa (Kananen & Puolitaival 2019, 133). Esimerkiksi CNN (Convolutional Neural Networks) soveltaa nimensä mukaisesti matemaattista konvoluutiota laskennassaan ja se sopii erityisesti kuvadatan käsittelyyn. RNN (Recurrent neural network) puolestaan soveltuu hyvin signaalidatan käsittelyyn, kuten esimerkiksi aikasarjadatan analysointiin ennakoivan huollon tehtävissä. RNN soveltuu myös kirjoitetun tekstin ja tallennetun äänen analysointiin. (Kallio 2018.) Kuviossa 14 on esitetty neuroverkon perusrakenne.



Kuvio 14. Neuroverkon perusrakenne (Kananen & Puolitaival 2019, 133)

6.3.4 Satunnainen metsä

Satunnainen metsä (Random Forest, RF) on lukuisista päätöspuista koostuva menetelmä sekä luokittelu- että regressio-ongelmiin. Päätöspuu koostuu ihmisenpäättelykykyä mukailevasta päättelyketjusta, jossa jokaisessa solmukohdassa tehdään päätös, mihin suuntaan päätöspuussa edetään. Satunnainen metsä koostuu näistä päätöspuiden yhdistelmistä. (Kananen & Puolitaival 2019, 125–126.) Kuviossa 15 on kuvattu satunnaisen metsän yksinkertaistettu rakenne.



Kuvio 15. Satunnaisen metsän yksinkertaistettu malli (Bakshi, 2020)

Satunnainen metsä on ohjattua oppimista, joka sopii useisiin erilaisiin käyttötapauksiin. Käytännössä satunnainen metsä muodostaa päätöspuukokoelman keskimääräisen ennusteen sen sijaan, että ennustettu arvo otettaisiin vain yhdestä puusta. Mallin avulla voidaan päästä tarkkoihinkin tuloksiin. Sen haasteena on kuitenkin mallin ylioppiminen tai mallin monimutkaisuus, mikäli satunnaismetsä sisältää päätöspuita runsaasti. Täten on huomioitavaa, ettei yksittäisten puiden määrän kasvattaminen loputtomasti ole tarkoituksenmukaista, vaan on syytä hakea optimaalista puiden määrää, josta satunnainen metsä siten koostuu. (Bakshi, 2020.)

Sekä päätöspuita että niiden yhdistelmistä muodostuvaa satunnaista metsää käytetään etenkin sellaisten ongelmien ratkaisemiseksi, missä tarkoituksena on ennustaa jatkuvaa arvoa kuten hintojen ennustamista. Malli on suosittu etenkin sen yksinkertaisuuden, mutta tarkan ennustemallin muodostumisen ansiosta useilla eri aloilla kuten asuntokauppa, pankkisektori, osakemarkkinat, lääketiede ja verkkokauppa. (Donges, 2019.)

6.4 Datan merkitys ja koneoppimisen hyödyntäminen

Robert Broända Bilot Groupilta kertoo IBM:n järjestämässä webinaarissa (2020), ettei 81 % yrityksistä ymmärrä vaatimuksia, joita tekoäly datalle asettaa. Hän kertoo myös, että 73 % yritysten datasta ei ole käyttökelpoista tai se on epäluotettavaa tai analysoimatta. Myös datan laadun erityisasiantuntija Thomas C. Redman painottaa artikkelissaan (2018) datan laadun merkitystä tekoälyn hyödyntämisen yhteydessä. Mikäli datan laatu on huonoa, koneoppimisen työkalut ovat hyödyttömiä. Jos data on puutteellista, vääristynyttä tai väärinymmärrettyä, voi se aiheuttaa koneoppimismallien vääristymää. Koneoppimisen algoritmit voivat vääristyä datan laadun takia ja tällöin pieneltä tuntuvalta virheellä voi olla kauaskantoiset vaikutukset ongelman moninkertaistuttua organisaation eri osiin ja prosessin eri vaiheisiin. Tekoälystä on kuitenkin moneksi sen jälkeen, kun laatuhaasteet on korjattu. Tekoälyn avulla voidaan muodostaa monikäyttöisiä työkaluja ja tehokaita apuälyjä yrityksen erilaisiin tarpeisiin. (Redman 2018.)

Microsoft toteutti vuonna 2018 yhteistyössä tilintarkastus- ja konsultointiyritys PricewaterhouseCoopers:n kanssa kyselyn 20 suureen suomalaiseen organisaatioon (Uncovering AI in Finland, 2018). Kyselyn tarkoituksena oli kartoittaa suomalaisten yritysten valmiutta tekoälyn hyödyntämiseen liike-elämässä. Kyselyn tärkeimmäksi havainnoksi tunnistettiin, että tekoälyn käyttöönottoon tulee tutustua mahdollisuuksien mukaan niin pian kuin organisaatio kokee olevansa siihen valmis. Seuraavaksi tärkeimpänä asiana tekoälyprojektissa ymmärrettiin datan laadun merkitys. Tekoäly on täysin riippuvainen datasta, joten laadun tulee olla käyttökelpoista ja datan perusteet kunnossa. Moni yritys myönsikin datan laatuongelmat. Hyvänä puolena useimmat yritykset olivat tunnistaneet datan laadun tuomat haasteet, jolloin siihen on mahdollista puuttua jo ennen tekoälyprojektin käynnistämistä. (Uncovering AI in Finland, 2018.)

Tekoälyn ja koneoppimisen hyödyntäminen liiketoiminnassa nähdään yleistyvän tulevaisuudessa (Uncovering AI in Finland, 2018). On kuitenkin syytä pohtia sitä, lisääkö tekoälyinnovaatio yrityksen taloudellista arvoa ja mikä on sen liitetoiminnallinen tarve. Tämän lisäksi täytyy ymmärtää mitä tekoälyn käyttöönottaminen vaatii. Ihmisten ja tiimien sitoutumisen lisäksi datalla on oleellinen osa. Datan hyödyntäminen voi olla haastavaa, mikäli sen ominaisuudet eivät ole selkeät ja sitä joudutaan käsittelemään, rikastamaan tai korjaamaan. Datan laadulla on näin ollen oleellinen merkitys. Tekoälysovellusten toteutus yrityksen omiin järjestelmiin onnistuu parhaiten, mikäli yrityksellä on käytössään hyvä data-arkkitehtuuri ja tietovarastointi, joiden avulla dataan päästään helpommin käsiksi sekä siiloutumisen ongelmat on jo ratkaistu. (Kananen & Puolitaival, 2019.)

7 TUTKIMUKSEN TOTEUTTAMINEN

Asuntojen ja kiinteistöjen arvon määrittäminen on haastava tehtävä. Yksittäisen kohteen arvioiminen voi poiketa huomattavasti muun alueen keskineliöhinnasta. Suuria eroja hintatasossa selittää muun muassa kohteiden sijainti, kunto, varustetaso ja näkymät. Myös kysynnän ja tarjonnan välinen suhde vaikuttaa kohteiden hintaan: mitä vähemmän tarjontaa suositulla sijainnilla, sitä korkeammat myyntihinnat ja päinvastoin. Hintojen määrityksessä on siis useita merkittäviä tekijöitä, joita on osittain hankala kvantifioida dataksi.

Asuntojen hinnan määrittämiseen ja pääsääntöisesti kiinteistönvälittäjän työtä helpottamaan tarjoaa Kiinteistönvälitysalan Keskusliitto Ry KVKL Hintaseurantapalvelua, jossa voi vertailla toteutuneita kauppahintoja ja kohteiden tietoja aiemmista asuntokaupoista. Tämän tutkimuksen tarkoituksena on tutkia kyseistä tietokantaa ja analysoida sen tuottaman datan laatua ja hyödyntämismahdollisuuksia koneoppimismallien muodostamiseen. Tutkimuksessa keskitytään erityisesti arvioimaan datan laatua objektiivisten mittareiden näkökulmasta, joilla on suora merkitys tekoälyn ja koneoppimisalgoritmien toiminnalle. Tarkoitus on selvittää, mahdollistaako datan laatu tekoälyalgoritmien hyödyntämisen asuntojen hinnan ennustamisessa nykyisessä muodossaan.

7.1 Toimeksiantajan esittely

Kiinteistönvälitysalan Keskusliitto Ry toimii kiinteistönvälitysalan edunvalvojana ja puolestapuhujana. Liitossa on yhteensä seitsemän jäsenryhmittymää, joiden yhteisenä foorumina liitto toimii. Jäsenryhmittymät ovat: Kiinteistömaailma Oy, OP Koti, Suomen Kiinteistönvälittäjät SKVL ry, Aktia Kiinteistönvälitys Oy, RE/MAX Finland, Sp-Koti Oy ja Realia Group Oy. (Kiinteistönvälitysalan Keskusliitto Ry, n.d. a; Kiinteistönvälitysalan Keskusliitto Ry, n.d. c.)

Liitto on listannut internetsivuillaan tärkeimmiksi tavoitteikseen asuntokaupan läpinäkyvyyden edistämisen, alan yhteiskunnallisena vaikuttajana ja edunvalvojana toimimisen, tutkimuksen ja koulutuksen kehittämisen sekä eettisen ja juridisen sääntelyn edistämisen. Liitto tekee töitä alan ammattilaisten paremman

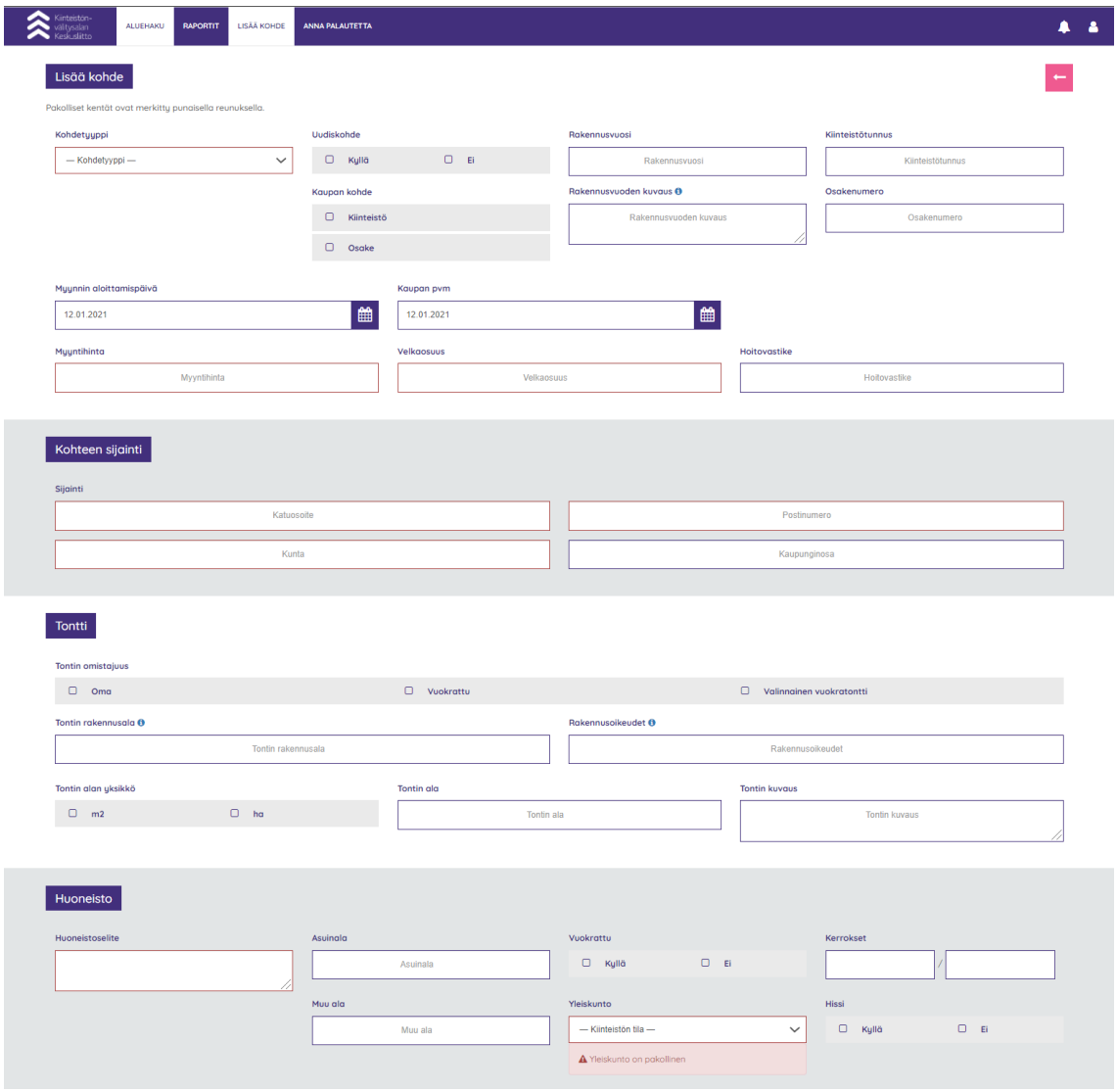
osaamisen tukemiseksi sekä tarjottujen asiantuntijapalvelujen laadun kohottamiseksi. (Kiinteistönvälitysalan Keskusliitto Ry, n.d. a; Kiinteistönvälitysalan Keskusliitto Ry, n.d. c.)

Liitto tekee esityksiä ja lausuntoja kiinteistönvälitysalan kysymyksissä viranomaisille ja tutkii ja tilastoi kiinteistönvälitysalaa. Se ylläpitää yhteyksiä kiinteistöalan järjestöihin Suomessa ja ulkomailla. Liiton tehtävä on myös ylläpitää hyvän välitystavan ohjetta, joka tärkeä osa alan itsesäätelyä. (Kiinteistönvälitysalan Keskusliitto Ry, n.d. a; Kiinteistönvälitysalan Keskusliitto Ry, n.d. c.)

7.2 Palvelukuvaus

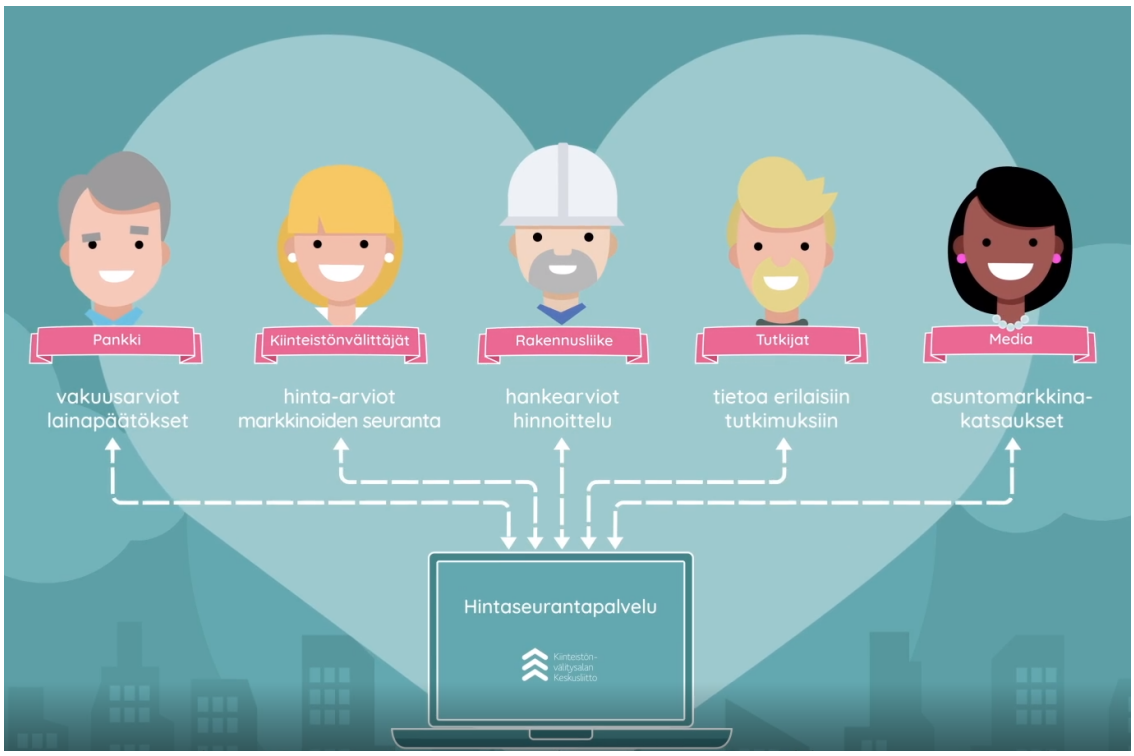
KVKL Hintaseurantapalvelu on Kiinteistönvälitysalan Keskusliitto Ry:n ylläpitämä tietokanta ja verkkopalvelu, joka on tarkoitettu ensisijaisesti kiinteistönvälitys- ja rakennusalalla toimivien yritysten avuksi, esimerkiksi myyntikohteiden markkinahinnan määrittelemisessä. Palvelussa on suomalaisten kiinteistönvälittäjien tekemien kauppojen tiedot vuodesta 1999 lähtien. Palvelussa on mukana yli 800 välitysliikettä ja merkittävä määrä rakennusliikkeitä. (Kiinteistönvälitysalan Keskusliitto Ry, n.d. b.)

Liiton kehitysjohtaja Mikko Hämäläinen kertoo (2021), että palveluun kerätään integraation kautta asuntokauppadataa kiinteistönvälittäjien ja rakennusliikkeiden käyttämistä kiinteistönvälitysalan järjestelmistä, kuten KIVI- ja PDX- toiminnanohjausjärjestelmistä. Tiedot voi myös syöttää palvelun selainpohjaisen käyttöliittymän kautta. (Hämäläinen 2021.) Kuviossa 16 on havainnollistava esimerkki palvelun käyttöliittymästä.



Kuvio 16. Ote Kiinteistönvälitysalan Keskusliitto Ry:n KVKL Hintaseurantapalvelun käyttöliittymästä (Kiinteistönvälitysalan Keskusliitto Ry KVKL Hintaseurantapalvelu, n.d.)

Palvelu mahdollistaa eri sidosryhmien tiedonsaannin. Palvelussa on tarjolla valmiita raportteja ja kuvaajia eri käyttäjien tarpeisiin. (Kiinteistönvälitysalan Keskusliitto Ry, n.d. b.) Kuviossa 17 on esimerkki eri palvelua hyödyntävistä sidosryhmistä.

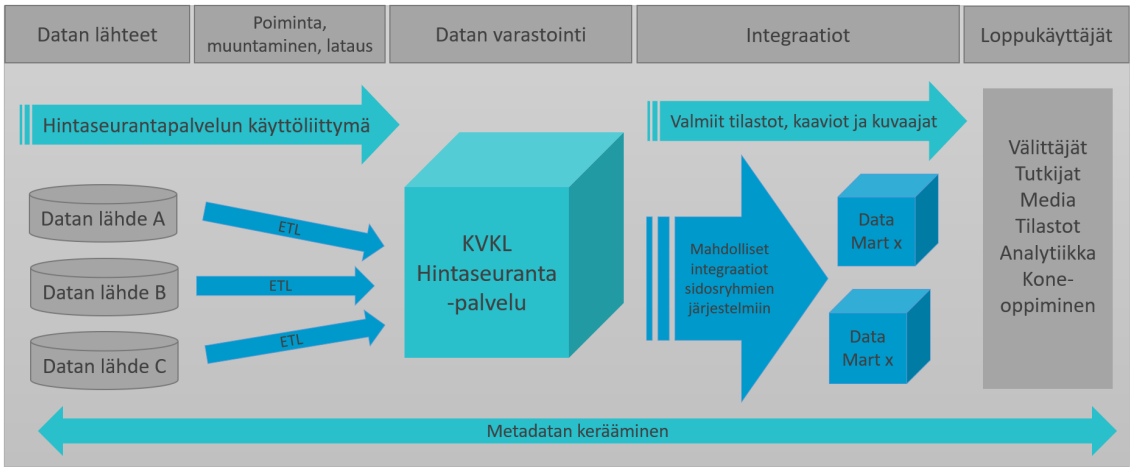


Kuvio 17. Kiinteistönvälitysalan Keskusliitto Ry:n KVKL Hintaseurantapalvelun sidosryhmiä (Kiinteistönvälitysalan Keskusliitto Ry, n.d. b)

Palvelun käyttö edellyttää, että kauppatiedot myydyistä kohteista toimitetaan jokaisen kuun 15. päivään mennessä ja kohteen tiedot pyritään julkaisemaan palvelussa pääsääntöisesti heti. Tämä takaa ajantasaisen datan. Datan kattavuuden kerrotaan olevan 75 % kaikista Suomessa myydyistä vanhoista asunnoista. Uusien asuntojen osalta kattavuuden arvioidaan olevan 30–40 %. (Kiinteistönvälitysalan Keskusliitto Ry, n.d. b.)

7.3 Palvelun tekninen arkkitehtuuri

Kiinteistönvälitysalan Keskusliitto Ry:n KVKL Hintaseurantapalvelu on tekniseltä arkkitehtuuriltaan tietokanta, joka toimii keskitettynä tietolähteenä alalla. Palvelun tiedot kerätään integraatioiden kautta eri ERP- eli toiminnanohjausjärjestelmistä. Yhtenä tiedonkeruukanavana toimii toimeksiantajan oma, selaimessa toimiva käyttöliittymä, jonka avulla käyttäjät voivat syöttää kauppatietoja palveluun. Integraation kautta kerättävät tiedot kulkevat ETL-prosessin läpi, jonka jälkeen ne julkaistaan palvelussa. Kuviossa 18 on esitetty pääpiirteinen kuvaus palvelun teknisestä arkkitehtuurista.



Kuvio 18. Pääpiirteet Kiinteistönvälitysalan Keskusliitto Ry:n KVKL Hintaseurantapalvelun teknisestä arkkitehtuurista

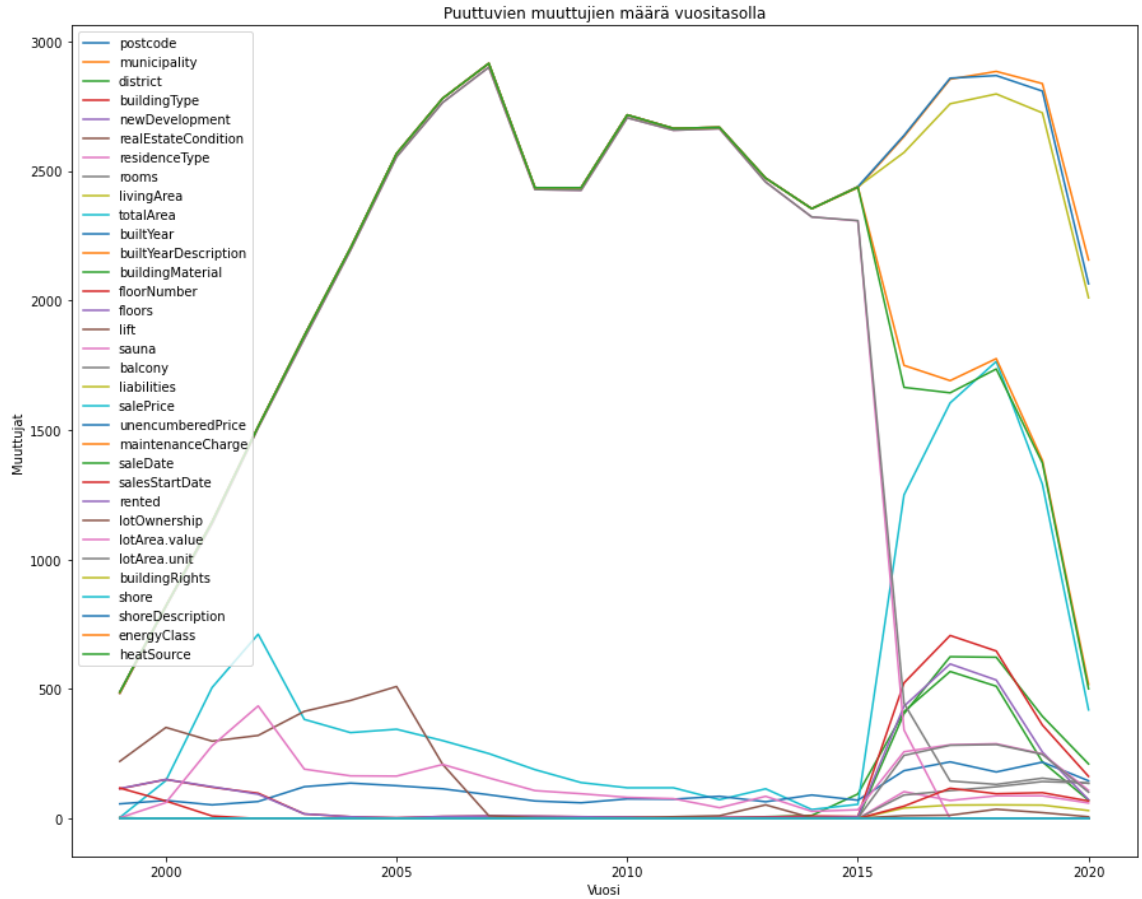
Kappaleessa 8 tutustutaan tarkemmin toimeksiantajan dataan ja tarkastellaan sen rakennetta ja laadullisia ominaisuuksia.

8 DATAAN TUTUSTUMINEN JA LAADUN ARVIOINTI

Tässä kappaleessa tutustutaan tarkemmin toimeksiantajan data-aineistoon. Data-aineiston tarkastelussa keskitytään datan laadun ulottuvuuksien analysoimiseen koneoppimisen näkökulmasta hyödyntäen kuutta DAMA UK:n (2013) käyttämää ulottuvuutta. Ulottuvuuksien mittaamiseen on useita erilaisia menetelmiä, joita käsittelimme kappaleessa 5. Tutkimuksen näkökulmasta keskitymme analysoimaan objektiivisia mittareita, kuten datan kattavuus ja datan tarkkuus. Tarkoitus saada yleiskäsitys tietokannan datasta koneoppimisen hyödyntämiseen. Subjektiivisten ulottuvuuksien analysoinnin, kuten data-aineiston järkevyys, rajaamme työmme ulkopuolelle.

8.1 Datan kattavuus ja aineistorakenne

Tietokannassa on kattavasti tietoa asuntojen myyntitoteumista aina vuodesta 1999 lähtien. Tapahtumarivejä on lähes 1,5 miljoonaa ja yksi tapahtumarivi käsittää kymmeniä eri sarakkeita, jotka tässä tapauksessa voidaan mieltää muuttujiksi. Tietokannalle on tehty muutoksia vuosien mittaan, joten aineiston kattavuus vaihtelee eri ajankohtina. Näin ollen aineiston analysointia varten selvitämme ensin aineiston kattavuutta tutkimalla puuttuvien arvojen määrää. Kuviossa 19 on esitetty puuttuvien arvojen määrä eri vuosina. Visualisoimalla puuttuvien arvojen määrää pystytään nopeasti päättelemään aineiston kattavuuden muutoksia vuosien aikana.



Kuvio 19. Puuttuvien arvojen määrä vuosittain

Kuvion 19 on tarkoitus antaa yleisnäkymä puuttuvien arvojen määrästä. Tässä kohtaa yksittäistä muuttujaa ei ole kuitenkaan tarkoitus analysoida vaan muodostaa näkemys kokonaistilanteesta. Mitä korkeampi käyrä, sitä enemmän puuttuvia arvoja aineistossa esiintyy. Kuviosta voidaan siten havaita, että tietokantaan on tehty muutoksia vuosien saatossa. Kuviosta 19 voidaan myös havaita vuonna 2015–2016 tapahtuneen muutoksia datan muuttujien määrässä. Silloin data-aineistoon on lisätty lisää muuttujia, joita aiemmin ei ollut mukana. Tämä näyttäytyy kuviossa 19 korkeina käyrinä. Näin ollen voidaan todeta, että varsinaista puuttuvien arvojen tilannetta täytyy tarkastella kulloinkin voimassa olevan datan tallentamisvuoden mukaan. Tämä on hyvä ottaa huomioon myös koneoppimismallin rakentamisessa, mikäli ennustemallin muodostamisessa käytetään sellaisia muuttujia, joiden arvoja ei ole tallennettu järjestelmään eli tietoa ei ole saatavissa.

Tutkimukseen käytetty data-aineisto on rakenteellisessa muodossa. Data-aineisto koostuu asuntomyynnissä tapahtuneista tapahtumariveistä ja sarakkeista. Jokainen sarake sisältää kenttätiedon, jossa määritetään tietotyyppi. Tietotyyppi on metadatatietoa, jota käytiin tarkemmin läpi kappaleissa 2.3 ja 4.3. Tietotyyppi

rajaa mahdollisuuden täydentää dataa vain ennalta määritellyllä tavalla. Näin ollen jokaiselle kentälle on erikseen määritelty muoto, jossa data pystytään tallentamaan. Tässä yhteydessä huomionarvoista on, että osa arvoista on korvattu oletusarvoilla. Oletusarvojen käyttämisestä puhuttiin teoriaosuuden kappaleessa 6.3.1. Näin ollen on mahdollista, ettei arvo vastaa todellisen maailman tilannetta. Asiaa tarkastellaan lisää seuraavassa kappaleessa. Taulukossa 1. tarkastellaan aineiston rakennetta ja puuttuvien arvojen määrää. Toimeksiantajan tietokannassa kerätään kuitenkin tätä kuvausta kattavammin tietoja myydyistä kohteista.

Taulukko 1. Muuttujien tietotyypit ja puuttuvat arvot

	Muuttuja	**Muuttuja suomeksi**	**Tietotyyppi**	**Puuttuvien arvojen määrä 1999 – 2020 prosentteina**
1	postcode	postinumero	Numeerinen	0,00
2	municipality	kaupunki	Merkkijono (vapaateksti)	0,00
3	district	kaupunginosa tai kylä	Merkkijono (vapaateksti)	3,84
4	buildingType	asuntotyyppi	Merkkijono	0,00
5	newDevelop-ment	uudiskohde	Boolean	0,00
6	realEstate-Condition	huoneiston kunto	Merkkijono	0,00
7	residen-ceType	huoneistokuvaus	Merkkijono (vapaateksti)	0,81
8	rooms	huoneluku, keittiötä ei lasketa	Numeerinen	1,23
9	livingArea	asuinala m2	Numeerinen	0,44
10	totalArea	muu kuin asuinala m2 (muu ala)	Numeerinen	19,98
11	builtYear	rakennusvuosi	Numeerinen	4,89
12	builtYear-Description	rakennusvuoden selite	Merkkijono (vapaateksti)	99,79

13	buildingMaterial	rakennusmateriaali	Merkkijono (vapaateksti)	4,52
14	floorNumber	asuinkerros	Numeerinen	6,00
15	floors	kerrosten lukumäärä	Numeerinen	4,93
16	lift	hissi	Boolean	0,00
17	sauna	sauna	Boolean	73,00
18	balcony	parveke	Boolean	74,30
19	liabilities	lainanosuus	Numeerinen	0,00
20	salePrice	myyntihinta	Numeerinen	0,00
21	unencumberedPrice	velaton myyntihinta	Numeerinen	0,00
22	maintenanceCharge	hoitovastike	Numeerinen	0,00
23	saleDate	myyntipäivä	Päivämäärä	0,00
24	salesStartDate	myynnin aloituspäivä	Päivämäärä	1,26
25	rented	vuokrattu	Boolean	0,00
26	lotOwnership	tontin omistajuus	Merkkijono	6,00
27	lotArea.value	tontin ala	Numeerinen	6,78
28	lotArea.unit	tontin alan määrä	Merkkijono (vapaateksti)	2,46
29	buildingRights	rakennusoikeuskaava	Merkkijono (vapaateksti)	98,82
30	shore	oma ranta	Boolean	0,00
31	shoreDescription	rannan kuvaus	Merkkijono (vapaateksti)	99,64
32	energyClass	energialuokka	Merkkijono (vapaateksti)	87,26
33	heatSource	lämmitysmuoto	Merkkijono (vapaateksti)	86,82
34	auction	huutokauppa	Boolean	100,00

Kuten aiemmin on jo todettu, aineistorakenteeseen on tehty muutoksia vuosien varrella, joka selittää useiden kenttien korkeat puutteelliset tiedot.

Data-aineiston tietotyypit:

- **Boolean** tietokentät määrittelevät onko asia tosi (true) vai epätosi (false). Näitä tietokenttiä käytetään ilmaisemaan esimerkiksi, onko kohteessa parveketta tai saunaa. Näiden tietojen täydentäminen ei ole pakollista, joten mikäli käyttäjä ei ole lisännyt kyseistä tietoa, tietokantaan täydentyy automaattisesti ennalta määritelty tieto, joka on määritelty aineistokuvauksessa.

- **Päivämääräkentät** sisältävät tietoa myynnin aloittamispäivämäärästä ja kaupantekopäivästä. Tätä tietoa pystytään hyödyntämään, esimerkiksi laskemalla kauanko keskimäärin asunnon myyntiaika on milläkin kohteella, alueella tai esimerkiksi onko poikkeamia eri vuosien tai vuodenaikojen välillä.

- **Numeeriset kentät** sisältävät tietoa erilaisista numeerisista arvoista kuten myyntihinta, postinumero ja kohteen huoneiden lukumäärä.

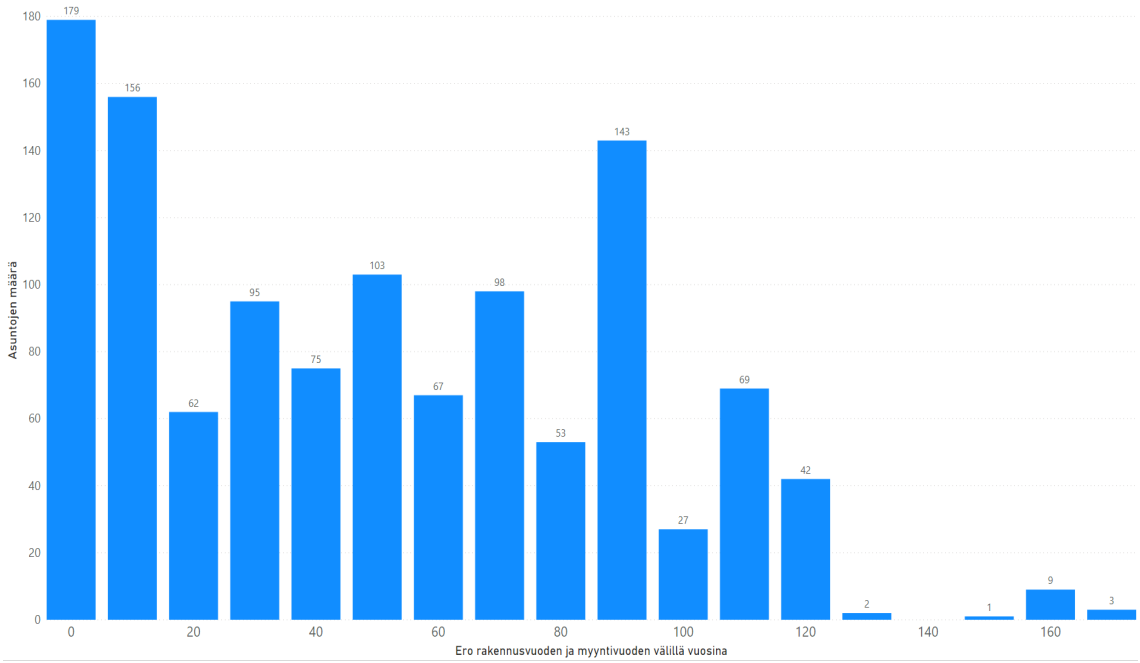
- **Merkkijono (vapaateksti)** -tyyppisiä kenttiä on aineistossa eniten. Nämä kentät mahdollistavat asuntoon liittyvien huomioiden lisäämisen. Koska merkkijono(vapaateksti)-tyyppisissä kentissä annetaan mahdollisuus kirjoittaa vapaata tekstiä, on niiden käsittely koneoppimisessa tällä hetkellä haastavaa. Käsittelyyn on mahdollisuus hyödyntää luonnollisen kielen tekniikoita, mutta me rajaamme tutkimuksessamme vapaatekstikenttien käsittelyn työn ulkopuolelle.

- **Merkkijono** -tyyppiset kentät sisältävät pelkästään ennalta määriteltyä tekstiä tai merkkejä, eli niin kutsuttuja pudotus- tai alasvetovalikkoja. Näissä kentissä määritellään valmiiksi vapaasti syötettävien arvojen sijaan joukko arvoja, joista käyttäjän on mahdollisuus valita kohteelle sopiva tieto. Tällaisia kenttiä ovat tietokannassa muun muassa kohteen tyyppi (esimerkiksi omakotitalo, rivitalo tai kerrostalo) ja kohteen kunto (esimerkiksi hyvä / tyydyttävä / välttävä). Näiden tietokenttien hyödyntäminen koneoppimisessa on mahdollista, sillä erilaisten arvojen määrä on rajattu.

8.2 Datan oikeellisuus

Vaikka data-aineisto olisi täydellistä eli kattavaa, on datan oikeellisuudella merkittävä vaikutus datan tulkittavuuteen. Mikäli aineistossa tiedot on täytetty, mutta ne ovat vääriä, voi se aiheuttaa vaikeuksia aineiston todenmukaiselle tulkinnalle. Tietojen paikkansapitävyyttä voidaan muun muassa mitata määrittelemällä rajaarvoja, joiden perustella tarkastetaan datan oikeellisuus. Yhtenä tällaisena mittarina voidaan tarkastella uudisrakennuksen tietokenttää. Tämä tietokenttä on boolean – typpinen ja oletusarvona tietokenttään täydentyy ”false” eli tieto ei päde, mikäli käyttäjä ei erikseen ole antanut kenttään arvoa. Koneoppimismallin käyttämisessä kentällä on kuitenkin merkitystä, mikäli sitä käytetään muuttujana mallin opettamiseen. Tällöin tiedon oikeellisuudella on merkitystä.

Tutkimalla tarkemmin tietokanta-aineistoa voidaan arvioida pitääkö tietokentän data paikkaansa. Kuviossa 20 on analysoitu uudiskohteiksi merkittyjä asuntoja, joiden rakennusvuoden ja myyntivuoden välinen aika on yli viisi vuotta. Näiden asuntojen tiedoista on laskettu rakennusvuoden ja myyntivuoden välinen aika vuosina ja osavälit (bins) on eritelty kymmenen vuoden välein. Ensimmäiset viisi vuotta on jätetty kuvaajasta pois, sillä niiden oletetaan osuvan toleranssiin varsinaisten uudiskohteiden myynnissä.



Kuvio 20. Ero rakennusvuoden ja myyntivuoden välillä, yli 5 vuoden ero huomioidaan, osaväli (bins) = 10 vuotta.

Kuvio 20 osoittaa, että asuntoja on merkitty runsaasti uudiskohteeksi vielä jopa kymmeniä vuosia rakennuksen valmistumisen jälkeen. Ongelma toistuu myös toisinpäin, eli asuntoa ei ole merkitty uudiskohteeksi, vaikka asunto on myyty rakennuksen valmistumisvuonna tai lähivuosina. Seuraavaksi esitettävissä taulukoissa 2 ja 3 on nostettu esiin esimerkkirivejä datasta, jossa on huomattavissa tämä ilmiö.

Taulukko 2. Otos aineistosta: kohdetta ei ole merkitty uudiskohteeksi, vaikka todennäköisesti kyse voi olla uudiskohteesta

Rakennusvuosi	Myyntivuosi	Merkitty uudiskohteeksi
2020	2020	FALSE
2019	2020	FALSE
2016	2017	FALSE
1999	1999	FALSE
2000	2000	FALSE

Taulukko 3. Otos aineistosta: kohde merkitty uudiskohteeksi, vaikka todennäköisesti se ei ole uudiskohde

Rakennusvuosi	Myyntivuosi	Merkitty uudiskohteeksi
1997	2003	TRUE
1988	2008	TRUE
2000	2013	TRUE
1963	2015	TRUE
2013	2020	TRUE

Puuttuvan datan osalta voidaankin harkita, onko oletusarvojen käyttäminen järkevää. Uudiskohteen kohdalla puuttuvan tiedon kenttään täydennetään arvo ”false” eli epätosi. Kuitenkaan tämä ei ole koneoppimisen näkökulmasta suositeltavaa, sillä kone olettaa todellisen arvon olevan todellakin ”false”. Parempi onkin jättää tietokenttä täydentämättä ja antaa arvo ”null” eli tyhjäarvo.

Mikäli kyseessä on rakennuksen laajentaminen, perusparannus tai korjaus- / kunnossapito, voidaan rakennusta tietyin ehdoin pitää uudenveroisena. Kuitenkin on pohdittava, pitäisikö tällaiselle vaihtoehdolle olla oma erillinen tietokenttänsä. Näin ollen pystytään määrittelemään tarkemmin, millaisesta kohteesta on tapauksessa kyse. Tämän ansiosta datan laatua ja tarkkuutta saadaan parannettua.

8.3 Ainutlaatuisuus

Data-aineistossa on datan ainutlaatuisuudella merkitystä eli toisin sanoen tuplatietueita ei kuulu aineistossa olla. Mikäli tuplatietueita löytyy, voivat ne kokonaisuuden kannalta vääristää datan ainutlaatuisuutta ja tulkittavuutta. Myös koneoppimismalliin sillä on merkitystä. Koneoppimismalli menettää tarkkuuttaan, mikäli opetusdatassa on virheitä kuten tuplatietueita.

Tarkastuksen perusteella tuplatietueita löytyi 0,8 % tietokannasta vuosien 1999–2020 välillä eli yhteensä 11953 kappaletta. Tämän datan osalta täytyisi selvittää onko kyseessä varsinaisesti tuplatietue vai onko mahdollisesti kyseessä kuitenkin oma ainutlaatuinen myyntitapahtuma, joka muuttujien määrän ollessa vähäinen tulkittaan koneellisesti tuplatietueeksi. On kuitenkin huomioitava myös se, että aineistossa on tapahtunut muutoksia vuosien saatossa. Näin ollen viimeisen viiden vuoden (2015–2020) ajalta ei löytynyt yhtään tuplatietuetta. Vaikuttaisi siis siltä, että ainakin viime vuosien myyntitapahtumat ovat varsinaisia myyntitapahtumia eikä tietoja ole syötetty useampaa kertaa järjestelmään. Sama asunto on voitu myydä vuosien varrella useasti, jolloin niiden kuuluu olla oma ainutlaatuinen tapahtumansa ja näin ollen data on oikein tietokannassa.

8.4 Oikeamuotoisuus

Datan oikeamuotoisuudella on oleellinen osuus datan tulkitsemiseen. Oikeamuotoisuutta voidaan valvoa teknisesti oikeanlaisen tietotyypin eli metadatan määrittelyllä. Tämän lisäksi voidaan huomioida esimerkiksi data-aineiston tulkittavuuteen liittyviä asioita kuten vapaatekstikenttien osuutta tai rakenteisuutta. On erityisen tärkeää huomioida loppukäyttäjät ja muut sidosryhmät, jotka hyödyntävät tietokannan tietoa. Datan esitystavan tulisi olla helposti ymmärrettävissä ja siten oikeamuotoista.

Kuten olemme käyneet teoriaosuudessa läpi, datalle voidaan tehdä erilaista tulkintaa ja visualisointeja. Kuitenkin etenkin vapaatekstikenttien käsittely aiheuttaa usein tulkittavuusongelmia. Osalle vapaatekstikenttiä on perusteltua käyttää ky-

seisen muotoista tietotyyppiä, jolla pystytään antamaan lisätietoa myyntikohteesta. Osalle kentistä olisi syytä kuitenkin pohtia vapaatekstikentän tarpeellisuutta. Otetaan esimerkkinä kohteen lämmönlähteen tietokenttää. Koska käyttäjällä on mahdollisuus lisätä tieto vapaatekstinä, on lämmönlähteiden ilmaisussa useita eri tapoja. Tarkalleen ottaen aineisto sisältää erilaisia arvoja 17024 kappaletta. Vielä tarkemmin tarkasteltaessa huomataan, että pelkästään kirjoitusasulla on merkitystä, kuten esimerkiksi isoilla ja pienillä kirjaimilla. Nämä tulkittiin aineiston analysoinnissa erillisiksi lämmönlähteiksi. Lämmönlähteen puuttuvia arvoja on yli 85 %, mutta kuten havaittiin data-aineistosta, tietokannan rakennetta on muutettu vuosien saatossa ja tämä kenttäkin on tullut vasta myöhemmin aineistoon mukaan. Kuviossa 21 on esimerkkejä käyttäjien syöttämistä lämmönlähteistä.

Arvo	Määrä	Esiintyvyys aineistossa(%)
Kaukolämpö	54540	3.9%
Kauko	22838	1.6%
kaukolämpö	20317	1.4%
Sähkö	17081	1.2%
Maalämpö	5156	0.4%
sähkö	4677	0.3%
Kaukolämpö, vesikeskuslämmitys	3497	0.2%
Suora sähkölämmitys	3348	0.2%
Kaukolämpö, vesikiertoinen lattialämmitys	3127	0.2%
Öljy	2999	0.2%
Other values (17014)	49011	3.5%
(Missing)	1228794	86.8%

Kuvio 21. Lämmönlähteen eri arvot aineistossa

Taulukossa 4 on tutkittu lämmönlähdettä vieläkin tarkemmin. Taulukosta havaitaan esimerkiksi kaukolämmön kohdalla, että saman tiedon esittämiseen on käytetty montaa eri tapaa.

Taulukko 4. Otos lämmönlähteen erilaisista kuvauksista

Lämmönlähde
Kaukolämpö (Vesikiertoinen lattialämmitys)
Kaukolämpö, vesikiertoiset patterit
Kaukolämpö, vesikiertoinen lattialämmitys
Kaukolämpö, koneellinen ilmanvaihto
Kaukolämpö, vesikeskuslämmitys
Kaukolämpö, asunnoissa vesikiertoinen lattialämmitys
Kaukolämpö, vesikiertoinen lattialämmitys,
kaukolämpö, vesikiertoinen patterilämmitys. Mahdollisuus maalämpöön.
Kaukolämpö, lattialämmitys
Kaukolämpö, vesikeskuslämmitys patterein.
Kaukolämpö, Ilmanvaihto: Koneellinen poisto
Kaukolämpö (vesikeskuslämmitys)
kaukolämpö, vesikiertoiset patterit/lattialämmitys
Kaukolämpö/vesikiertoinen lattialämmitys
Kaukolämpö, vesikiertoinen patterilämmitys
Kaukolämpö, vesikiertoisin patterein
Kaukolämpö, vesikiertoiset patterit.
Kaukolämpö, vesikiertoinen lattialämmitys. Ilmalämpöpumppu
Kaukolämpö, lämmönjakotapana vesipatterit, koneellinen poistoilmanvaihto
Kaukolämpö, vesikeskuslämmitys. Painovoimainen ilmanvaihto.

Näin ollen voidaan miettiä, onko lämmönlähteen vapaateksti -tietotyyppi oikea. Tietotyypin muuttaminen ennalta määrättyyn muotoon rajaisi datan monimuotoisuutta ja mahdollistaisi datan hyödyntämisen muun muassa koneoppimismalleihin. Tilastokeskus on Suomen valtion virasto, jonka tarkoitus on toimia muun muassa tietopalveluna. Tilastokeskus määrittelee listauksen lämmönlähteistä, jotka on esitelty taulukossa 5 (Tilastokeskus, n.d.). Säännönmukaisen tiedon hyödyntäminen lisää mahdollisuuksia datan käyttämiseen, mutta myös samalla rajaa mahdollisuuden tarkemman kuvauksen antamisesta. Näin ollen ennalta määritellyn tietokentän lisäksi mahdollinen vapaatekstikenttä lämmönlähteen tarkempaan kuvaamiseen voisi olla tarpeen. Näin ollen data saadaan tallennettua en-

nalta määritellyssä muodossa sekä tarvittaessa yksityiskohtaisemmat tiedot ilmoitetaan lisäkentässä. Koska data saapuu toimeksiantajan tietokantaan pääsääntöisesti muista järjestelmistä, täytyisi muutos toteuttaa yhteistyössä datan lähtöjärjestelmien toimittajien kanssa.

Taulukko 5. Listaus lämmönlähteistä (Tilastokeskus, n.d.)

Luokitus on seuraava:
• Vesikeskuslämmitys
• Ilmakeskuslämmitys
• Suora sähkölämmitys
• Uuni- tai kamiinalämmitys
• Ei kiinteää lämmityslaitetta
• Tuntematon

Taulukossa 4 on myös merkintöjä ilmanvaihtoon tai viilennykseen liittyen. Tällä hetkellä tällaisille tiedoille ei ole omia tietokenttiään, mutta niiden tarpeellisuuden arviointi voisi kuitenkin olla tarpeellista.

Vapaatekstikenttien käsittelyyn kuitenkin myös mahdollista hyödyntää luonnollisen kielen koneoppimistekniikkaa (natural language processing). Tämä mahdollistaa avainsanojen löytymisen sanojen segmentoinnin keinoin. Avainsanan pelkistämisentekniikat mahdollistavat sanan muuttamisen perusmuotoon, jolloin vapaatekstin hyödyntämisen mahdollisuudet laajenevat.

8.5 Johdonmukaisuus

Datan yhdistettävyys muihin tietokantoihin tuo datalle lisää mahdollisuuksia ja näin ollen mahdollistaa laajempia käyttökohteita. Datan yhdistämisen keskeisenä piirteenä on datan johdonmukaisuus muihin aineistoihin. Tutkimusaineiston datassa on ilmoitettu muun muassa osoitetieto. Maantieteellinen sijaintitieto voidaan ilmoittaa usealla eri tavalla ja osoitteen kirjoitusmuoto voi olla hyvin vaihteleva. Osoitetiedon perusteella data voidaan yhdistää esimerkkisi alueelliseen tut-

kimiseen ja analysointiin. Mitä enemmän erilaisia kirjoitusmuotoja on, sitä epäjohdonmukaisemmaksi datan hyödynnettävyys muuttuu. Katuosoitteen kirjaamisessa haastavuuden tuo kirjaamisen monivivahteisuus, jolloin sen hyödyntäminen on haastavampaa. Katuosoitteen lisäksi aineistossa kerätään erilliseen tietokenttään postinumerotieto. Postinumero on Suomessa säännönmukainen, joten sen oikeellisuuden tarkastaminen ja näin ollen yhdenmukaistaminen muihin aineistoihin tulisi olla suhteellisen selkeää. Kuitenkin aineistossa postinumeroille löytyi useita satoja virheellisiä postinumeroita. Osa postinumeroista sisälsi väärän määrän numeroita, selkeästi virheellisiä postinumeroita tai kenttään oli lisätty muita kuin numeerisia arvoja. Esimerkkejä virheellisistä arvoista esitetty taulukossa 6. Kappaleen 8.1 taulukosta 1 huomattiin postinumeron tietotyypin olevan kehitystyön myötä nykyisin numeerinen. Kuitenkin yhä viime vuosina aineistoon on päätynyt virheellisiä postinumeroita. Lähtöjärjestelmään olisi ehkä syytä miettiä, olisiko järkevää kontrolloida postinumeroiden kirjaamista virheellisyyksien ehkäisemiseksi.

Taulukko 6. Otos virheellisesti ilmoitetuista postinumeroista

Postinumero	**Virheen syy**
100	Puutteellinen postinumero
6880	Puutteellinen postinumero
12345	Virheellinen postinumero
99999	Virheellinen postinumero
656300	Liian pitkä postinumero
-	Postinumeroa ei ilmoitettu
21320.	Piste postinumeron perässä
8100 Kontiolahti	Lyhyt postinumero. Tämän lisäksi postitoimipaikka lisätty samaan kenttään.

8.6 Ajankohtaisuus

Datan ajankohtaisuutta voidaan analysoida vertaamalla tapahtuman todellista syntyhetkeä sekä ajankohtaa, jolloin tapahtuma on päivitetty tietokantaan käyttäjien saataville. Tässä opinnäytetyössä datan analysointiin käytettävässä aineistossa ei ole ilmoitettu datan julkaisemisen ajankohtaa, joten tarkkaa analysointia

toteutuneen myyntihetken ja tietojen julkaisemisen välillä ei pystytä toteuttamaan. Kuitenkin toimeksiantaja julkaisee tiedot pääsääntöisesti heti tietokannassaan, kun se on syötetty järjestelmään.

9 TUTKIMUSDATALLA ENNUSTAMINEN

Tässä kappaleessa tutustutaan koneoppimismallien testaamista varten rajattuun tutkimusaineistoon sekä käsitellään aineistolle tehtyjä esikäsittelytoimenpiteitä. Kappaleesta 9.3 alkaen esitellään eri periaattein esikäsitellyt testidata-aineistot ja käydään läpi niillä toteutetut koneoppimiskokeilut ja esitellään saadut tulokset.

Kappaleessa on hyödynnetty datan visualisointia tapana tutustua tutkimusdataan tarkemmin. Usein visualisointi paljastaa myös datan laatuvirheitä ja auttaa ymmärtämään dataa paremmin. Visualisoinnit on toteutettu Python-ohjelmointikieltä käyttäen sekä Microsoft Power BI -ohjelman avulla.

9.1 Muuttujien valinta ja aineiston rajaus

Varsinainen koneoppimismallien testaaminen toteutetaan rajatulla aineistolla. Rajaamme datan seuraavilla tiedoilla:

- Alue: Tampere
- Asuntotyyppi: Kerrostaloasunnot
- Myyntipäivä: 1.1.2014 - 31.12.2019

Rajattu aineisto sisältää 20997 tapahtumariviä. Datamäärä on jo melko kattava, jolla pystymme tekemään analysointia ja testaamaan koneoppimismalleja. Tarkoitus on tutkia, voidaanko mallia kouluttaa tarkemmaksi parantamalla aineiston laatua vai onko aineisto jo tarpeeksi hyvälaatuista, jolloin koneoppimisen algoritmit pystyvät muodostamaan tarpeeksi tarkan ennustemallin.

Muuttujien valinta on oleellisessa osassa koneoppimismallin tarkkuuden määritelmässä, kuten kappaleessa 6.3.1 todettiin. Koneoppimismalliin tulisi sisällyttää relevantteja muuttujia, joilla voi olla merkitystä ennusteen lopputulokseen. Muuttujien ensimmäinen rajaus toteutetaan seuraavin periaattein:

- Kentän tietotyyppi tulee olla boolean, numeerinen tai merkkijono
- Merkkijono(vapaateksti)-tyyppiset muuttujat jätetään käsittelystä pois niiden haasteellisen käsittelyn takia
- Muuttujan koetaan olevan oleellinen myyntihinnan määrittämiseksi

Muuttujien määräksi saadaan 12 muuttujaa sekä velaton myyntihinta, jota malli pyrkii ennustamaan.

- postcode / postinumero
- newDevelopment / uudisrakennus
- realEstateCondition / asunnon kunto
- rooms / huoneiden lukumäärä
- livingArea / asunnon koko
- builtYear / rakennusvuosi
- floorNumber / asuinkerros
- sauna / sauna
- balcony / parveke
- maintenanceCharge / hoitovastike
- saleDate / myyntipäivä
- lotOwnership / tontin omistus
- unencumberedPrice / velaton myyntihinta > Ennustettava muuttuja

Kuviossa 22 on esitetty valitut muuttujat. Kuvio kiteyttää datan puutteet ja muodon muuttujien tasolla. Mikäli rivin arvojen määrä on 20997, tarkoittaa se, ettei rivi sisällä puuttuvia arvoja (Non-Null Count). Kuviosta voidaan heti kuitenkin huomata osan muuttujien kohdalla datan olevan puutteellista. Lisäksi kuviosta huomataan datatyypien (Dtype) kolmenlaista jaottelua: float64 on numeerinen, bool on boolean ja object on kategorinen muuttuja. Näin ollen havaitaan, että dataa joudutaan esikäsittelemään ennen koneoppimismallin muodostamista.

```
RangeIndex: 20997 entries, 0 to 20996
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   postcode              20997 non-null  float64
 1   newDevelopment        20997 non-null  bool
 2   realEstateCondition   20997 non-null  object
 3   balcony               15081 non-null  object
 4   maintenanceCharge     20997 non-null  float64
 5   saleDate              20997 non-null  object
 6   livingArea            20995 non-null  float64
 7   rooms                 20970 non-null  float64
 8   builtYear             19905 non-null  float64
 9   floorNumber           19474 non-null  float64
 10  sauna                 15278 non-null  object
 11  unencumberedPrice     20997 non-null  float64
 12  lotOwnership          20924 non-null  object
dtypes: bool(1), float64(7), object(5)
```

Kuvio 22. Tutkimusdatan muuttujat python – ohjelmointikielellä analysoituna.

9.2 Datan esikäsittely

Tässä opinnäytetyössä halutaan tutkia toimeksiantajan tietokannan datan laatua koneoppimismallien hyödyntämiseen. Samalla tutkimme, millaisilla datan laadun parantamisen esikäsittelytoimenpiteillä voidaan vaikuttaa ennustustarkkuuteen.

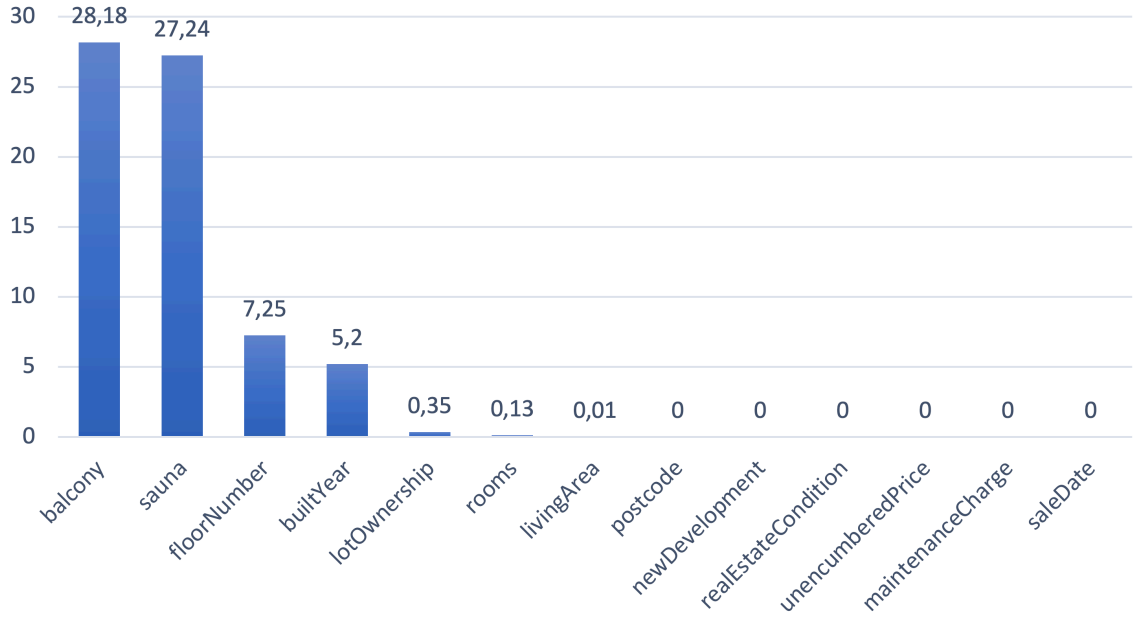
Datan laatu ja esikäsittely on koneoppimisprojektien keskiössä. Kuten aiemmin tässä opinnäytetyössä todettiin, on datan laadulla ja sen parantamiseen tarvittavalla esikäsittelyllä merkittävä rooli koneoppimismallin muodostamisessa. Esikäsittely on algoritmien oikeinmuodostumisessa oleellista ja siihen voi kulua koneoppimisprojektissa jopa pidempi aika kuin itse koneoppimisalgoritmien luomiseen. Mitä parempi laatuista data on, sitä vähemmän aikaa kuluu datan esikäsittelyyn.

Datan esikäsittely- ja rajaustoimenpiteet sekä koneoppimiskokeilut toteutetaan kokonaisuudessa Python – ohjelmointikielellä, josta löytyy laaja kattaus erilaisia

kirjastoja muun muassa matemaattisiin laskelmiin, visualisointeihin sekä koneoppimismalleihin. Python myös mahdollistaa monipuolisten esikäsittelytoimenpiteiden toteuttamisen.

Puuttuvien arvojen käsittely

Kuten jo aiemmin huomattiin, data on osittain puutteellista. Otannalle tehdyn analysoinnin perusteella havaittiin, että puuttuvia arvoja on etenkin muuttujien sauna ja parveke (balcony) arvoissa. Kuviossa 23 voidaan huomata, että myös muutamassa muussa muuttujassa on puutteellista tietoa.



Kuvio 23. Prosenttiosuus muuttujien puuttuvista arvoista

Datan puuttuvia tietoja tai virheitä korjataan ensisijaisesti korvaamalla ne oikeilla arvoilla. Mikäli oikeita arvoja ei ole tiedossa, voidaan tiedot korvata esimerkiksi muuttujan keskiarvolla tai mediaanilla. Vaihtoehtoisesti puutteelliset muuttujat voidaan jättää kokonaan mallin koulutuksesta pois tai jättää muuttujat ennalleen, mutta poistaa kaikki muuttujan puutteelliset rivit.

Kategoristen arvojen käsittely

Koneoppimisalgoritmien kannalta on tärkeää, että kaikki muuttajat on muunnettu numeeriseen muotoon. Aineistoon tutustuessa havaittiin erilaisia tietotyyppejä muuttujien välillä. On kuitenkin oleellista, että kategoriset muuttujat muutetaan

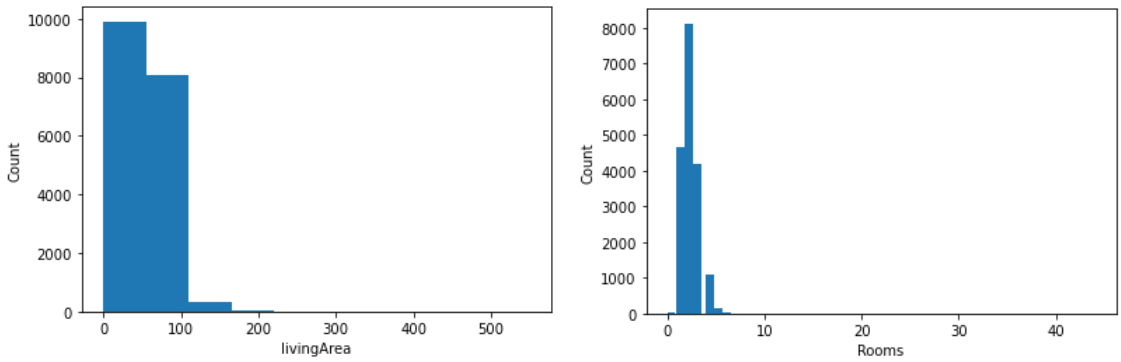
numeeriseen muotoon. Näin ollen esimerkiksi huoneiston kuntoluokitus muutettiin numeriseen muotoon:

- 'condition_unknown': 0
- 'condition_poor': 1
- 'condition_decent': 2
- 'condition_good':3
- 'condition_excellent':4
- 'condition_new':5

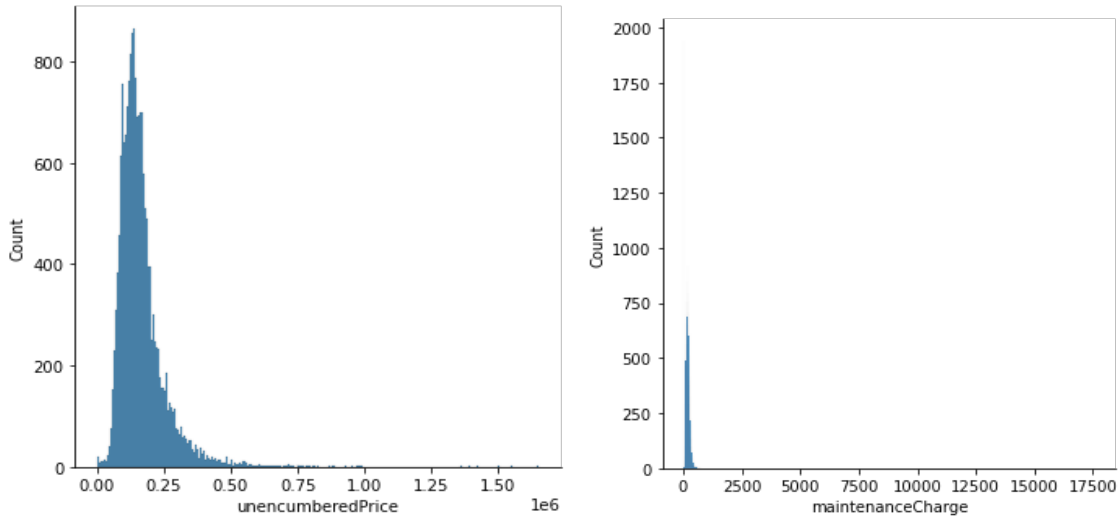
Kategoristen muuttujien osalta päivämäärien osalta myyntipäivä (saleDate) muutettiin pelkkään vuosiluku -tietoon.

Poikkeamien käsittely

Poikkeamien käsittelyyn on monta mahdollista tapaa. Aineiston poikkeavuuksien havainnoiminen on lähtökohta ja parhaiten tämä onnistuu visualisoimalla lähtötilanne. Kuviosarjassa 24 ja 25 on visualisoitu aineiston jakauman lähtötilannetta. Kuvioiden X-akselien iso skaala kertoo aineiston sisältävän normaalijakauman ulkopuolisia arvoja.



Kuvio 24. Datan jakauma asuinalan ja huoneiden lukumäärän osalta puuttuvien arvojen poistamisen jälkeen, mutta ennen poikkeamien käsittelyä.



Kuvio 25. Datan jakauma velattoman myyntihinnan ja vastikkeen osalta puuttuvien arvojen poistamisen jälkeen, mutta ennen poikkeamien käsittelyä.

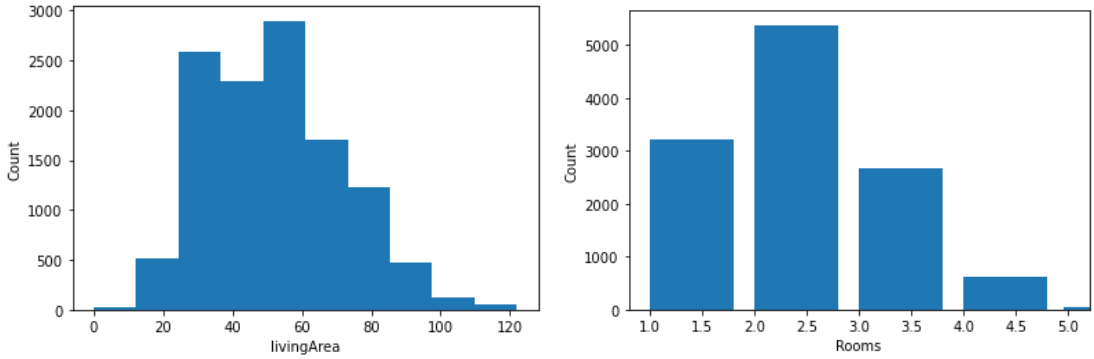
Taulukko 7 osoittaa selvemmin ongelman. Aineiston jakauma on useiden muuttujien kohdalla vinoutunut ja näin vääristynyt. On hyvin epätodennäköistä, että jonkun yksittäisen kerrostaloasunnon huonemäärä on 44 huonetta (rooms) tai velaton myyntihinta (Uncumbered Price) olisi 0 euroa.

Taulukko 7. Muuttujien maksimi- ja minimiarvoja

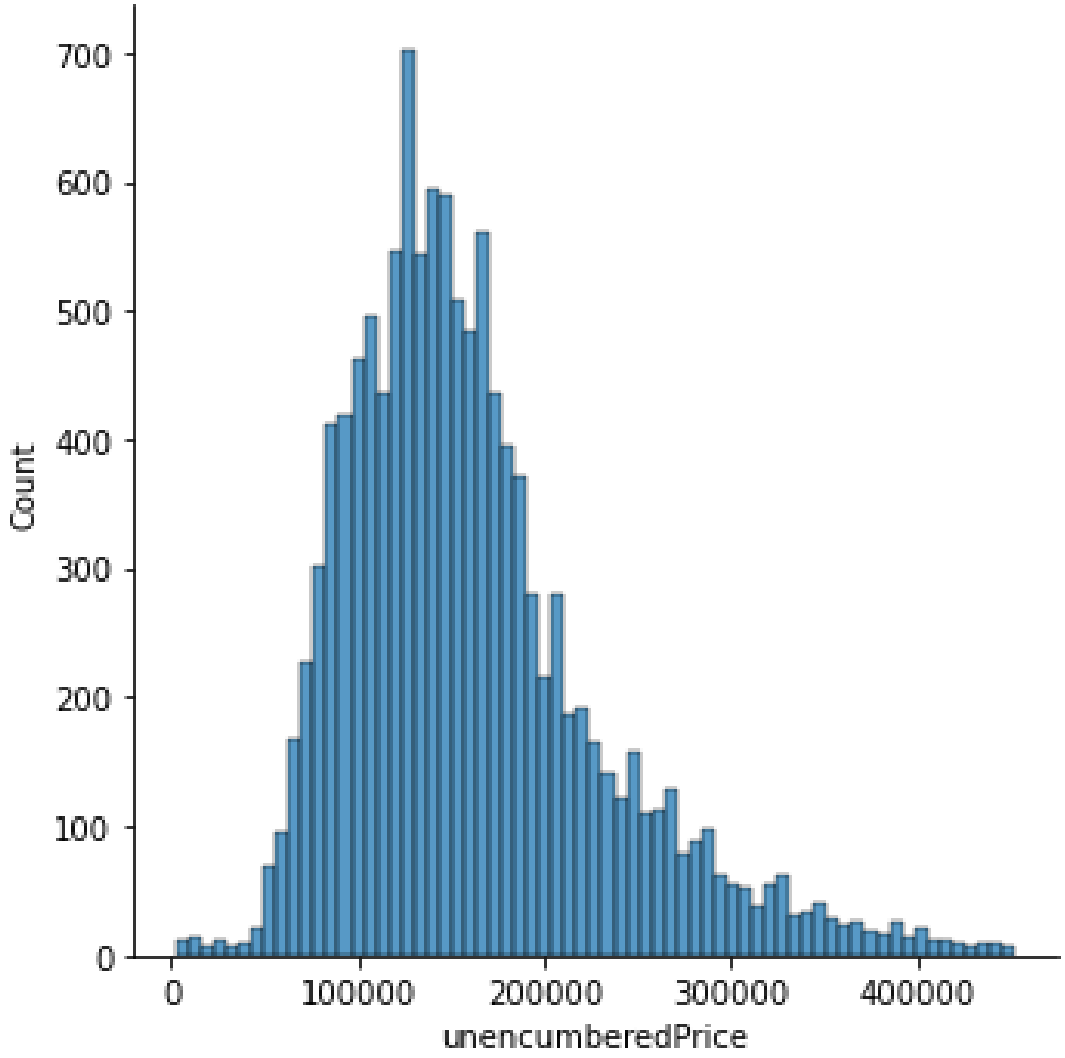
	Maintenance Charge	Uncumbered Price	PostCode	Rooms	Living Area	Built Year
Maksimi	17600	1650000	38850	44	550	2021
Minimi	0	0	12345	0	0	1855

Olisi tärkeätä, että normaalista selvästi poikkeavat arvot korjataan ennen niiden lataamista tietokantaan. Mikäli varsinaisia tietoja ei ole saatavilla, poikkeamien käsittely voidaan tehdä myös datan esikäsittelyssä. Data voidaan skaalata ja standardoida. Koneoppimismallin testaamiseen on poikkeaminen standardoiminen toteutettu työssämme käyttämällä Z-pisteen laskentamallia, joka ottaa huomioon normaalijakauman ja keskihajonnan. Tässä tavassa datalle lasketaan Z-pisteet tai standardipisteet, joiden perusteella lasketaan standardipoikkeamia, joissa datapisteiden arvo ylittää mitattavan keskiarvon. Poikkeamien käsittelyn jälkeen jakaumat noudattavat enemmän standardijakaumaa ja poikkeamat on pudotettu aineistosta pois. Kuvioista 26 ja 27 voidaan todeta, että poikkeamien

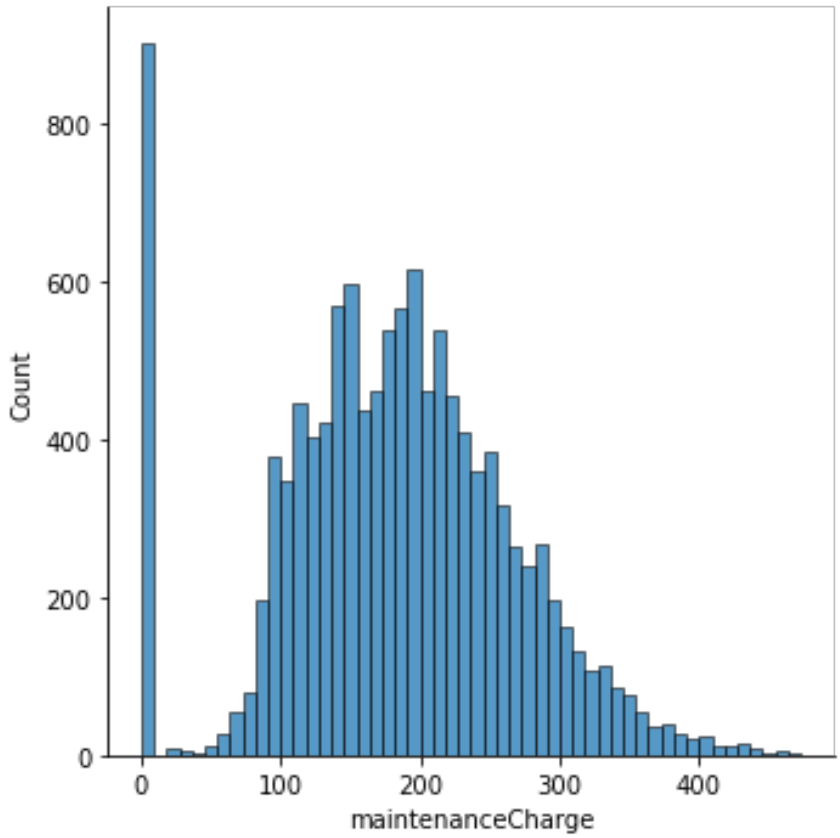
käsittelyn jälkeen jakaumat noudattavat enemmän standardijakaumaa ja poikkeamat on pudotettu aineistosta pois.



Kuvio 26. Asuinalan ja huoneiden lukumäärän jakauma puuttuvien arvojen poistamisen ja poikkeamien käsittelyn jälkeen.

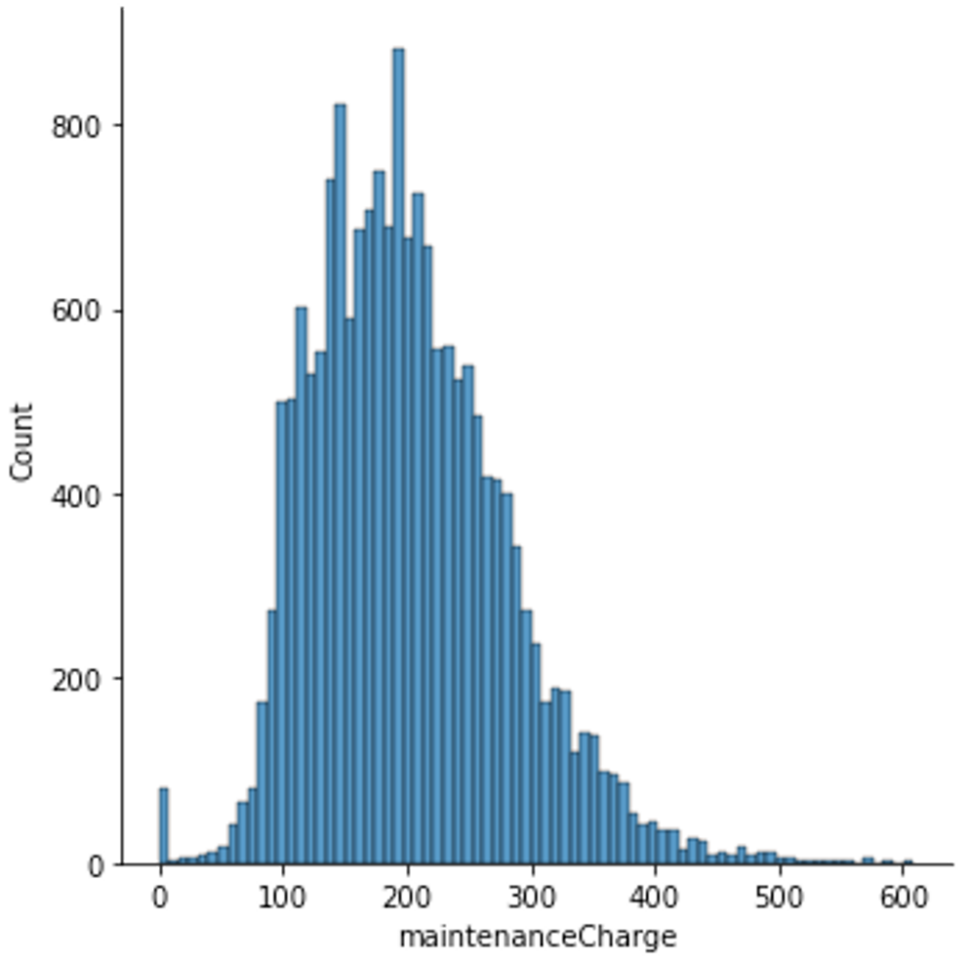


Kuvio 27. Velattoman myyntihinnan ja hoitovastikkeen jakauma puuttuvien arvojen poistamisen ja poikkeamien käsittelyn jälkeen.



Kuvio 28. Velattoman myyntihinnan ja hoitovastikkeen jakauma puuttuvien arvojen poistamisen ja poikkeamien käsittelyn jälkeen

Kuten visualisoinneista voidaan havaita, on data nyt standardoidumpi poikkeamien käsittelyn jälkeen. Kuitenkin hoitovastikkeen (maintenanceCharge) kuvaajasta huomaamme (kuvio 28) nolla-arvojen huomattavan määrän aineistossa. Todellisuudessa joidenkin asuntojen hoitovastike voi olla nolla euroa. Kuitenkin tässä data-aineistossa määrä on niin suuri, että nolla-arvot haluttiin korvata hoitovastikkeen laskennallisella keskiarvolla. Hoitovastikemuuttujan nolla-arvoja täydennetään laskemalla hoitovastikkeen määrä asuinneliötä kohden ja kertomalla se kohteen asuinneliöiden määrällä. Näin hoitovastikemuuttujan nolla-arvot eivät korostu liikaa aineistossa. Laskennassa hyödynnetään Tilastokeskuksen internetsivuilta saatua tietoa Tampereella sijaitsevien kerrostaloasuntojen keskimääräisestä hoitovastikkeesta. Laskentaan on käytetty vuoden 2017 arvoa kerrostaloasunnon hoitovastikkeesta, joka oli 4,81 euroa neliöltä (Tilastokeskus 2018). Kuviosta 29 nähdään hoitovastikkeen (maitenanceCharge) jakauma korjaustoimenpiteiden jälkeen.



Kuvio 29. Standardoidun hoitovastikkeen jakauma nolla-määräisten arvojen korvaamisen jälkeen.

9.3 Koneoppimismallin testaaminen

Koneoppimismalleina käytetään keinotekoista neuroverkkoa (ANN) sekä satunnaista metsää (RF), joita käsiteltiin kappaleissa 6.3.3. ja 6.3.4. Aineisto rajattiin koskemaan Tampereella vuosina 2014–2019 myytyjä kerrostaloasuntoja. Aineiston muuttujien määrä rajattiin kappaleessa 9.1 mukaan.

Koska tavoitteenamme on ymmärtää datan laatua ja näin ollen myös datan esikäsittelyn merkitystä, testataan rajattua tutkimusdataa kolmella eri tavalla. Ensimmäinen testi toteutetaan melko yksinkertaisilla datan laadun parantamisen esikäsittelytoimenpiteillä, jossa aineistosta poistetaan kokonaan puutteelliset rivit sekä kaikki aineiston rivit muunnetaan numeeriseen muotoon. Toinen testi toteu-

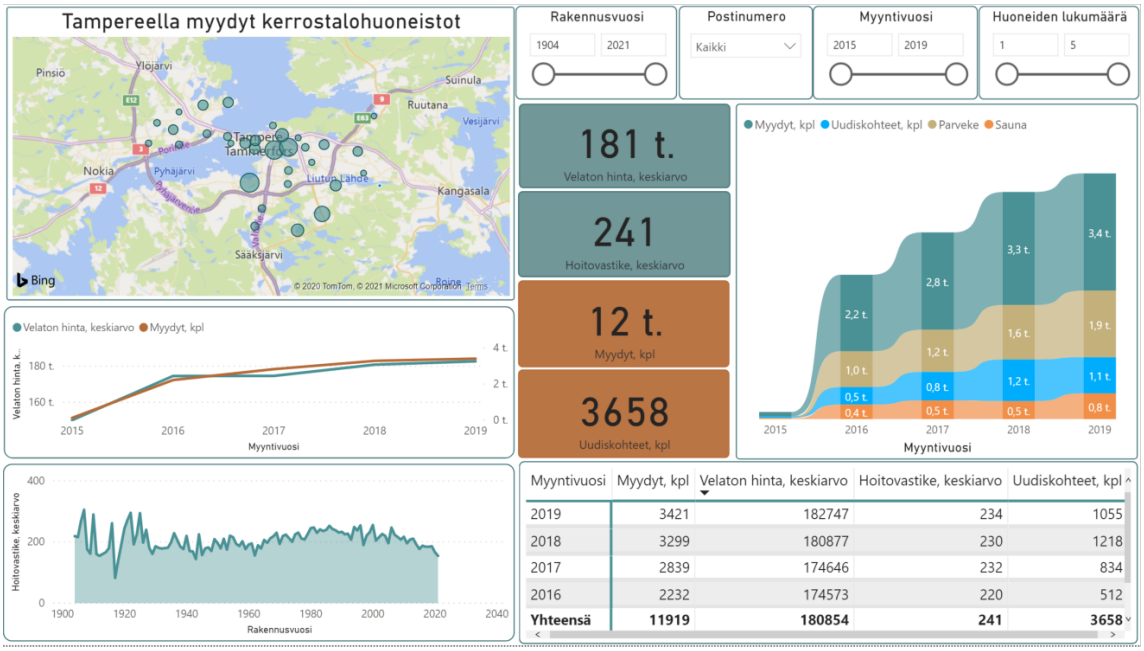
tetaan samalla aineistolla, mutta aineistoa esikäsitellään lisää siten, että aineistoa standardisoidaan ja toteutetaan poikkeavien arvojen käsittely. Esikäsittelymenetelmät on käyty tarkemmin läpi kappaleessa 9.2. Puuttuvien arvojen rivien poistaminen vähentää testiaineistoa merkittävästi. Koska puuttuvia arvoja emme pysty täydentämään todellisilla tiedoilla ja laskennallisen tiedon lisääminen voisi johtaa koneoppimismallia väärään lopputulokseen, toteutetaan kolmas testi siten, että muuttujien määrää muutetaan. Data-aineistosta eniten puuttuvia arvoja sisältävät muuttujat ovat sauna ja parveke (balcony). Näiden kahden muuttujan poistaminen aineistosta nostaa data-aineiston laajuuden 84,5 %:n alkuperäisestä aineistosta.

Esikäsittelyn aikana kaikki testiaineistot jaetaan kahteen osaan, jossa opetusdatan osuus on 75 % ja testiaineiston data 25 %. Taulukossa 8 on esitelty testiaineistot.

Taulukko 8. Testiaineistojen yhteenveto

	Esikäsittelytoimenpiteet	**Datan määrä**	**Aineiston jako**
Testi1	• Aineiston rajaus 2014–2019 • Muuttujien rajaus 13 kpl • Data muutettu numeeriseen muotoon • Puuttuvat arvot poistettu	12388, datan kattavuus 59 % alkuperäisestä aineistosta	Opetusdata 75 %, Testidata 25 %
Testi2	• Aineiston rajaus 2014–2019 • Muuttujien rajaus 13 kpl • Data muutettu numeeriseen muotoon • Puuttuvat arvot poistettu • Poikkeamien käsittely tehty	11919, 56,8 % alkuperäisestä aineistosta	Opetusdata 75 %, Testidata 25 %
Testi3	• Aineiston rajaus 2014–2019 • Muuttujien rajaus 11 kpl (sauna ja parveke jätetty pois) • Data muutettu numeeriseen muotoon • Puuttuvat arvot poistettu • Poikkeamien käsittely tehty	17751, 84,5 % alkuperäisestä aineistosta	Opetusdata 75 %, Testidata 25 %

Visualisoiminen on tehokas menetelmä datan analysointiin ja tulkintaan. Näin ollen eri testiaineistoja haluttiin vielä visualisoida monipuolisemmin, jotta saadaan selkeä kuva aineistoista ja niiden jakautumisesta. Kuviossa 30 on esitelty visualisointi, joka toteutettiin testi2:n datalla. Se on kattavuudeltaan siis 56,8 % koko tutkimusdatasta. Kuten kuvio paljastaa, testidata 2:sta on rajautunut esikäsittelytoimenpiteiden myötä lähes kokonaan vuosien 2014 ja 2015 myyntidata, koska kyseisinä vuosina ei kerätty tietoa siitä, onko kohteessa sauna tai parveke. Näin ollen kuva ei ilmennä kyseinä ajankohtana tapahtuneita myyntimääriä kokonaisuudessaan, vaan paljastaa esikäsittelyn vaikutuksen dataan. Esikäsittelyn vaikutukset onkin syytä huomioida, kun visualisointeja tutkitaan tai kun niillä lähdetään tekemään mallien opetusta.



Kuvio 30. Testidata 2 visualisoituna Power BI-työkalulla.

Sekä keinotekoisten neuroverkkojen, että satunnaisen metsän mallit sisältävät erilaisia parametrivaihtoehtoja, eli mallin koulutukseen käytettäviä asetuksia, joilla mallin oppimiseen voidaan vaikuttaa. Tämän työn tarkoituksena ei ole muodostaa tehokkainta koneoppimismallia, vaan kartoittaa datan laatua ja selvittää esikäsittelytoimenpiteiden vaikutuksia täysin erilaisten koneoppimismallien kouluttamiseen, joten työn ulkopuolelle on rajattu tarkempi läpikäynti asetuksien vaihtoehdoista. Taulukossa 9 on esitetty kuitenkin pääkohdat malleissa käytetyistä parametreista.

Taulukko 9. Pääkohdat koneoppimismallissa käytetyistä asetuksista

Keinotekoiset neuroverkot	Satunnainen metsä

• Keras – malli
• Kerrokset, 7 kpl
  ○ Input: ReLu
  ○ Hidden: 5 kerrosta / ReLu
  ○ Output: Linear
• Optimointi: ADAM
• 150 opetuskierrosta (epoc)
• 32 erän koko (batch size)
	• Satunnainen metsä regressiomalli sklearn.ensemble.RandomForestRegressor
	• Puiden lukumäärä = 200
	• Syvyys = Oletus
	• Näytteiden määrä haarassa = Oletus

9.4 Koneoppimismallin tulosten arviointi

Kuten yleisesti tiedetään, koneoppimisalgoritmit toimivat yleensä sitä paremmin, mitä enemmän dataa niillä on käytettävissä. Esikäsittelyssä on siis huomioitava toimenpiteiden vaikutukset datan kattavuuteen. Tässä testissä kuitenkin päästiin hyviin tuloksiin jo pelkästään pienillä laadun parantamisen toimenpiteillä, vaikka se tarkoittikin yli 40 % datan poistamista. Etenkin sellaisten asuntojen arvot pystyttiin ennustamaan melko tarkasti, jotka sijoittuvat keskihajonnan keskiosan lähettyville. Mitä erikoisemmasta asuntoyksilöstä on kyse, sitä haastavampaa on sille ennustaa velatonta myyntihintaa. Tulosten arvioinnissa käytetään kappaleessa 6.3.2 esiteltyjä R2 ja MAE-arvoja. Taulukossa 10 on tiivistetty käytettyjen koneoppimismallien tulokset eri testiaineistojen mukaan.

Taulukko 10. Koneoppimismallien tulokset

	Keinotekoiset neuroverkot			Satunnainen metsä		
	Testi1	Testi2	Testi3	Testi1	Testi2	Testi3
R2	0.8728	0.9069	0.9167	0.8986	0.9046	0.9059
MAE	19432 €	13269 €	13277 €	14017 €	12066 €	11689 €

Kaikissa testeissä R2 arvo oli lähellä 90 % (taulukko 10). Mitä enemmän aineistoa esikäsiteltiin ja datan laatua paranneltiin, sitä tarkemmin malli pystyi ennustamaan. Mitä lähempänä R2 on arvoa 1, sitä paremmin malli toimii. Jos R2 olisi

tasan 1, eli 100 % tarkoittaisi se ennusteen olevan aina oikein. Näin ollen tunnusluvun saavuttaessa yli 90 %, voidaan mallin todeta toimivan jo melko hyvin.

Toinen arviointiin käytettävä tunnusluku oli MAE eli keskimääräinen absoluuttinen virhe. Ensimmäisessä testiaineistossa saatiin heti melko hyvät tulokset. MAE jäi molemmissa tapauksissa alle 20 000 euron. Tämä tarkoittaa sitä, että asunnon ennustetun hinta-arvion ja toteutuneen myyntihinnan erotus on keinotekoisilla neuroverkoilla 19432 euroa, kun taas satunnainen metsä ennusti tarkemmin 14017 euron erotuksella (taulukko 10). Asuntokaupassa tätä voidaan pitää jo melko hyvänä tuloksena ottaen huomioon melko vähäiset datan parannus- sekä koneoppimismallin optimoinnin toimenpiteet.

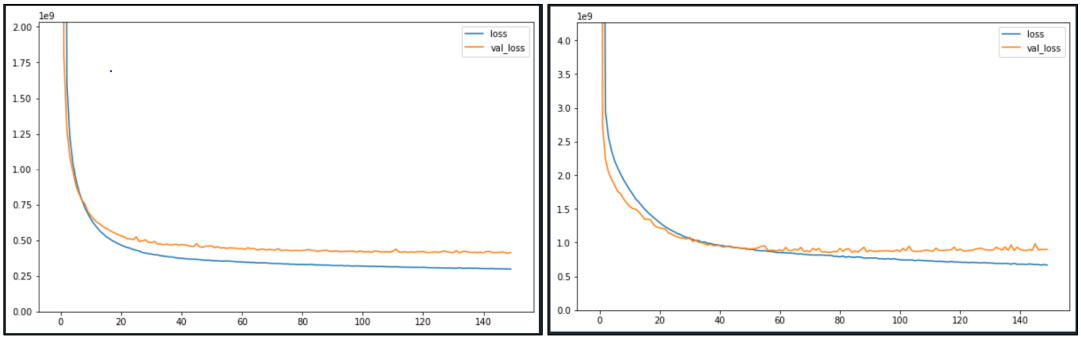
Tulos parani hiukan toisella testiaineistolla, jossa poikkeamien käsittely oli myös tehty. On huomioitava, että opetusdatan määrä putosi merkittävästi alkuperäisestä aineistosta, sillä esikäsittelyssä jouduttiin poistamaan merkittävä osuus datasta sen puutteellisuuden takia.

Aiemmin tutkimusaineistoa analysoitaessa huomattiin muuttujien sauna ja parveke (balcony) kohdalla niiden sisältävän merkittävän osuuden puuttuvia arvoja. Arvojen korvaamista ei haluttu toteuttaa, ettei data vääristyisi. Näin ollen kolmannessa testissä jätettiin nämä muuttuja-arvot pois kokonaan. Näin ollen kolmas testiaineisto oli huomattavasti aiempia testiaineistoja kattavampi vaikkakin muuttujien määrä oli pienempi. Aineistolle tehtiin myös poikkeamien käsittely. Tulosten perusteella tämän testiaineiston avulla päästiinkin parhaisiin tuloksiin (taulukko 10), vaikkei ero muihin testeihin ollutkaan huomattavasti aiemmista poikkeava.

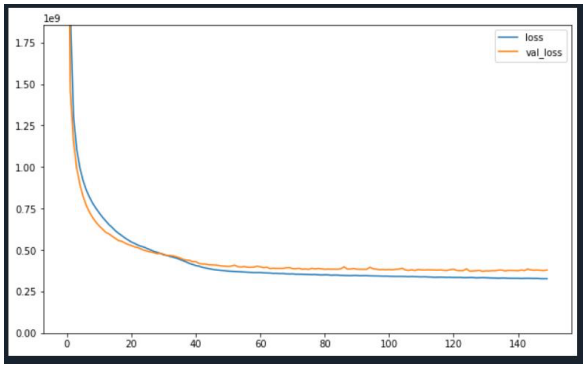
Molemmat koneoppimismallit näyttivät pääsevän hyviin tuloksiin, eikä mallien välillä ollut suurta hajontaa ennusteessa. Vaikka koneoppimismallin toimintalogiikka on hyvin erilainen, noudattavat molemmat ohjattua oppimista. Voidaan todeta, että molemmat koneoppimismallit ovat melko tehokkaita ennustamaan ainakin tällä data-aineistolla, esikäsittelytoimenpiteillä ja käytetyillä parametreilla.

Alla olevat kuviot 31 ja 32 esittävät neuroverkkomallin oppimiskäyriä, josta voidaan havaita kuinka nopeasti malli löytää suhteellisen tarkan ennustetason.

Testi2 – aineistolla mallin oppiminen on tasaisempaa ja tarkempaa ja testi3 aineistolla vieläkin parempaa. On kuitenkin huomioitava, että kuvien skaalaus on erilainen. Opetuskierroksen lisäämisellä ei näyttäisi olevan suurta merkitystä, joskin testi1 aineistossa malli alkaa jo hieman ylioppia n. 70 kierroksen jälkeen.



Kuvio 31: Opetuskäyrä testi1 (vasemmalla), opetuskäyrä testi2 (oikealla)



Kuvio 32. Opetuskäyrä testi3

MAE lukua eli keskimääräistä absoluuttista virhettä voidaan verrata todellisen ja ennustetun hinnan välillä. Taulukosta 11 voidaan havaita ennusteen olevan melko tarkka osalle asunnoista, kun taas osa asunnoista saa huomattavasti keskimääräistä absoluuttista virhettä korkeamman eron todelliseen myyntihintaan. Tarkastusvertailuun otetaan satunnaisesti kohteita aineistosta, joten vertailuun nousevat kohteet ovat poikkeavia sekä koneoppimismallien että tulosten tarkastamisen aikaan.

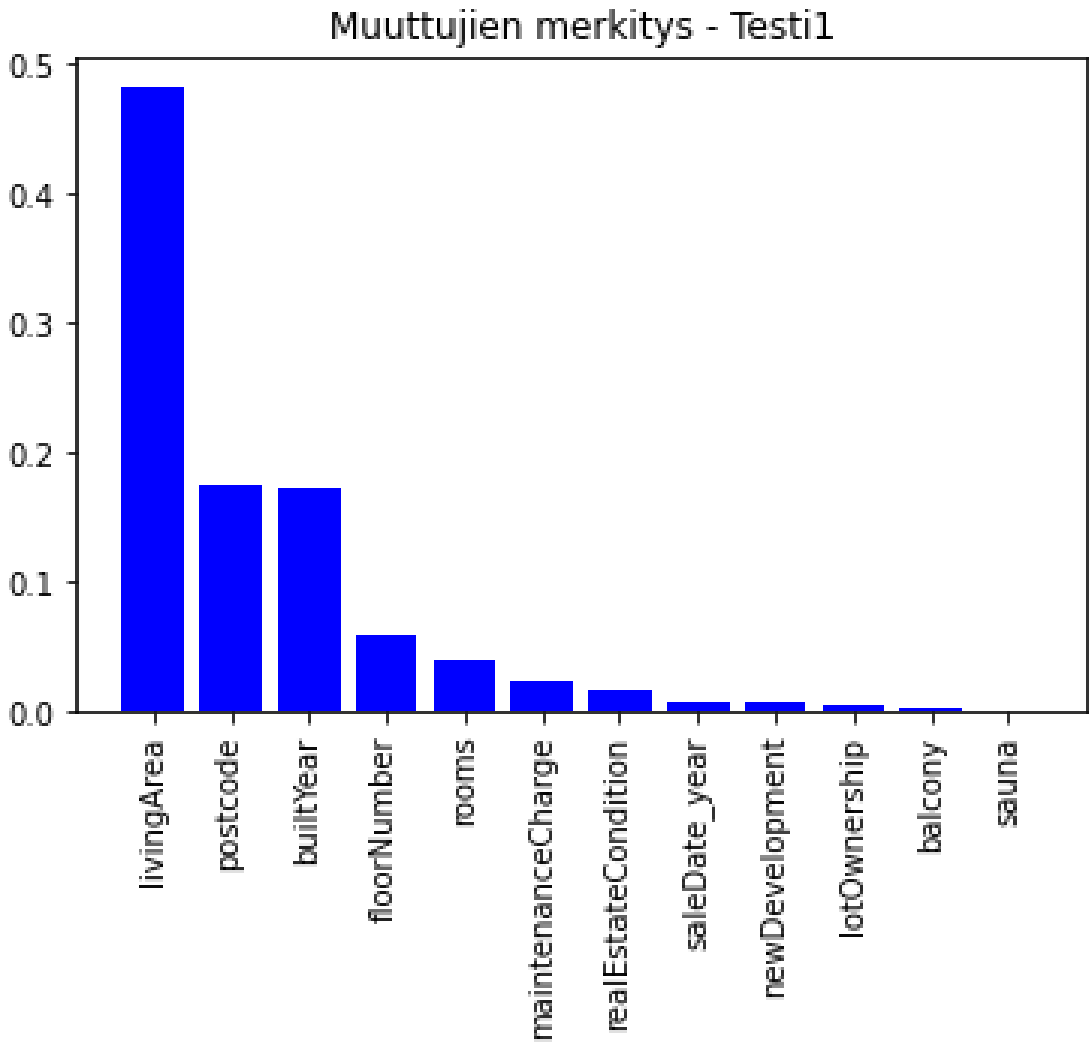
Taulukko 11. Vertailu todellisen ja ennustetun myyntihinnan välillä, satunnainen otanta 10 riviä (testi2 data)

Keinotekoiset neuroverkot			**Satunnainen metsä**		
Todellinen	Ennustettu	Erotus	Todellinen	Ennustettu	Erotus
140300	141084	-784	219750	210174	9575
275000	355916	-80916	136000	133651	2348

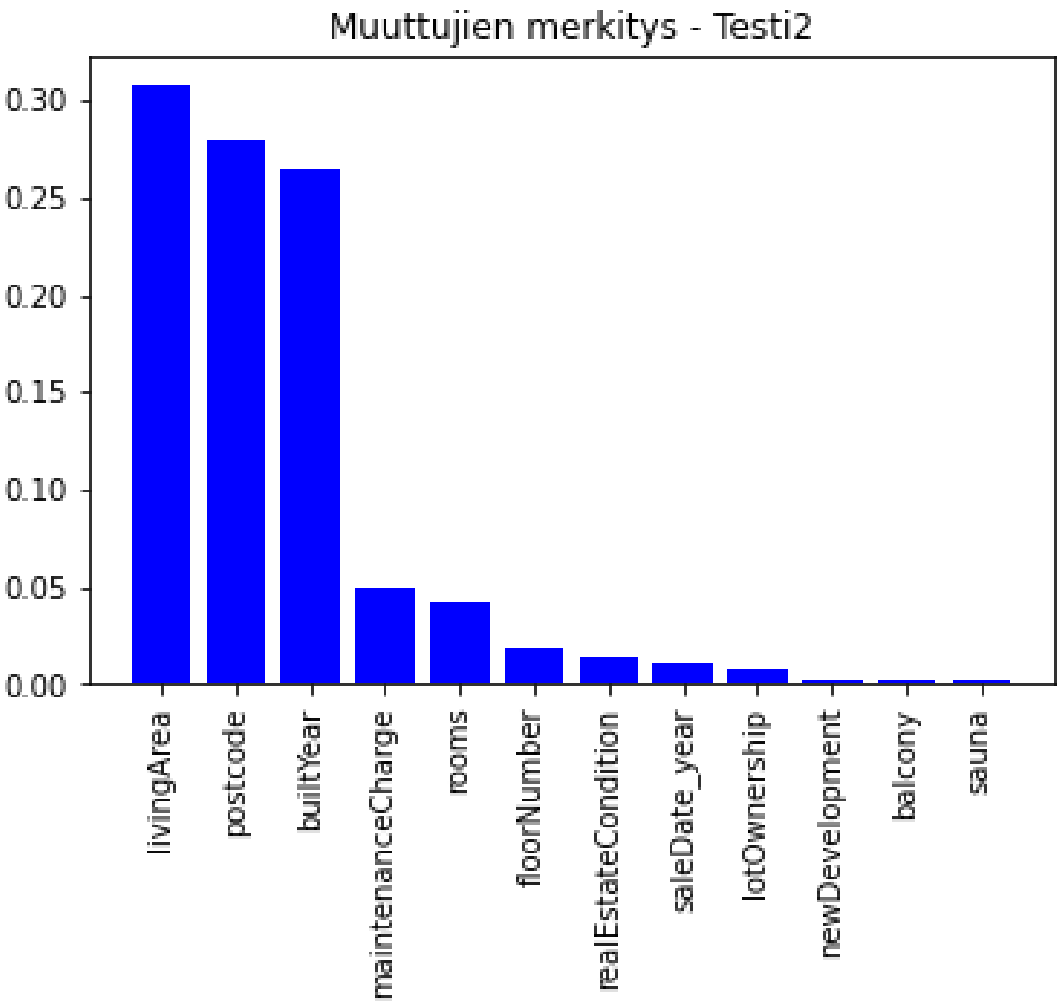
211347	218269	-6922	175000	178489	-3489
197500	190460	7040	166000	150073	15926
95000	86805	8195	106200	106763	-563
200700	201079	-379	74000	99435	-25435
108000	116589	-8589	208000	175396	32603
220000	192141	27859	367000	368022	-1022
190000	187273	2727	145000	123666	21333
108000	99873	8127	110500	101870	8629

Taulukosta 11 huomataan, että osa asuntojen hinnoista on ennustettu todella tarkasti, mutta osalle asunnoista on tullut jopa kymmenien tuhansien eurojen heitto. MAE luku ilmaiseekin tulosten keskimääräistä absoluuttista virhettä.

R2 ja MAE tunnuslukujen lisäksi tutkittiin muuttujien merkitystä ja painoarvoja asunnon arvon ennustamisessa satunnainen metsä koneoppimismallissa. Kuvioista 33 ja 34 on selkeästi huomattavissa, kuinka paljon muuttujien merkitys muuttuu testi1 ja testi2 aineistojen välillä. Vaikka R2 ja MAE arvot eivät poikenneet merkittävästi toisistaan, voidaan muuttujien merkityksen visualisoinneista huomata, että hinnan ennustaminen painottuu eri tavalla, riippuen siitä, miten aineistoa on esikäsitelty.



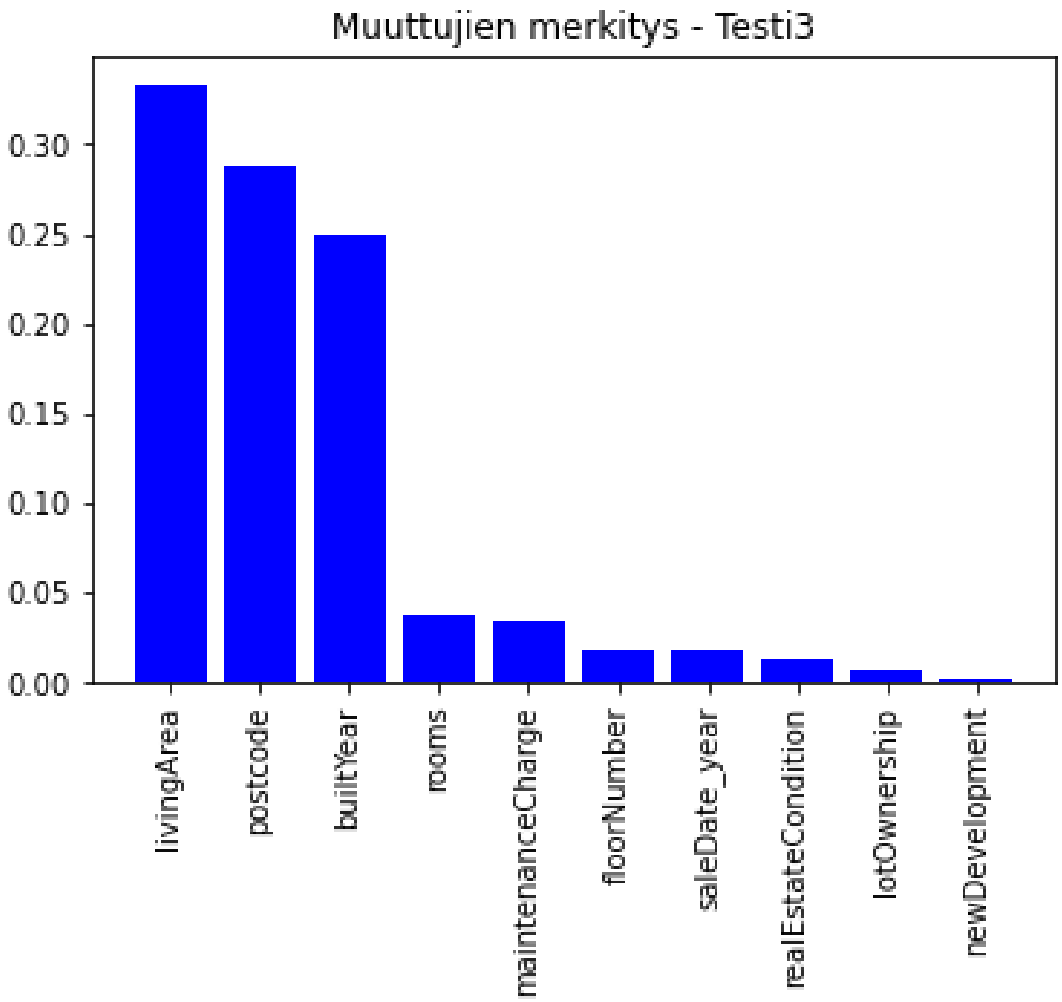
Kuvio 33. Testi1-aineiston muuttujien merkitys satunnainen metsä – koneoppimismallin ennusteessa.



Kuvio 34. Testi2-aineiston muuttujien merkitys satunnainen metsä – koneoppimismallin ennusteessa.

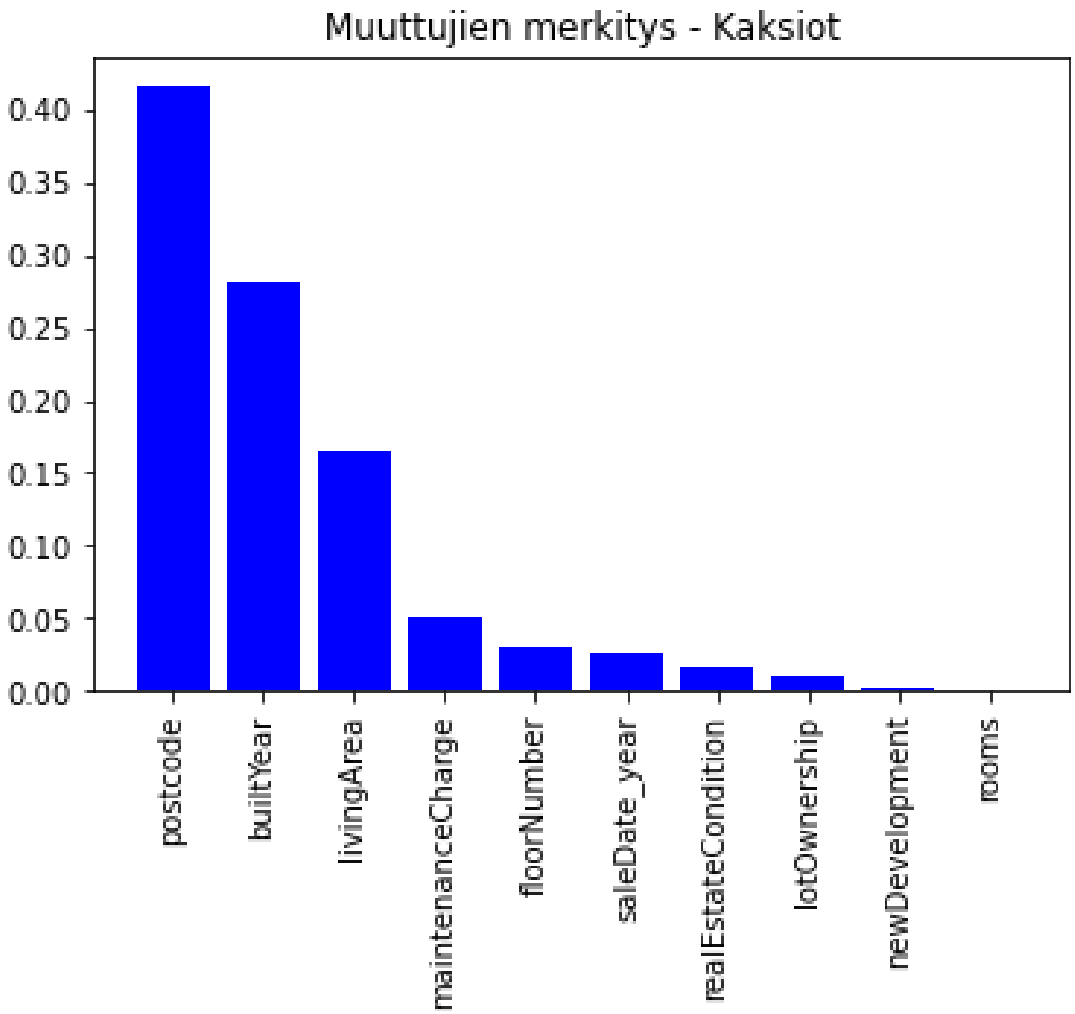
Testi1 -aineiston esikäsittelyssä poistettiin pelkät puutteelliset arvot, jolloin saatiin tulokseksi asuintilan olevan merkityksellisin muuttuja hinnan ennustamisessa (kuvio 33). Kun aineistoa standardisoitiin ja poikkeamat käsiteltiin, voitiin havaita, että postinumeroalueella ja rakennusvuodella on lähes yhtä merkityksellinen osuus (kuvio 34). Tässä vertailussa on kaikki Tampereen kerrostaloasunnot asunnon koosta riippumatta. Kuvioita 33 ja 34 tulkittaessa on syytä huomata, että skaala-asteikko on erilainen.

Testi3 toteutettiin ilman muuttujia sauna ja parveke (balcony). Kuten kuviosta 33 ja 34 huomataan, sauna ja parveke ovat hinnan ennustamisen painotuksen kannalta merkityksettömimmät muuttujat. Niiden painoarvo ei muuta juurikaan muuttujien painotusta kuten kuviossa 35 nähdään. Kuviossa 35 on visualisoitu muuttujien merkitys testi3 -aineistolla.



Kuvio 35. Testi3 – aineiston muuttujien merkitys satunnainen metsä - koneoppimismallin mukaan.

Testiaineistojen 1–3 lisäksi toteutettiin ylimääräisenä mallina pelkkien kerrostalokaksioiden muuttujien merkitysanalyysi. Näin ollen muuttujien painotus ja merkitys realisoituu vieläkin paremmin vastaamaan todellista tilannetta. Kuvasta 36 huomataan, että tärkeimmäksi muuttujaksi nouseekin postinumero eli asunnon sijainti, sen jälkeen rakennusvuosi ja vasta sitten asunnon koko. Tämä todennäköisesti vastaakin kaikkein todenmukaisinta tilannetta asunnon myyntihinnan määrittämisessä. Huoneiden lukumäärä (rooms) ei ole relevantti kuvion 36 tarkastelussa, sillä visualisointiin huomioitiin vain kohteet, joissa huoneiden lukumäärä on kaksi eli sama kaikissa.



Kuvio 36. Muuttujien merkitys satunnainen metsä - koneoppimismallin mukaan kerrostalokaksioiden osalta

Mielenkiintoisena huomiona nousi esiin uudiskohteen (newDevelopment) vähäinen merkitys. Tämän muuttujan merkitys oli kaikissa analysoinneissa (kuviot 33–36) vähäinen, lähes olematon. Toisaalta rakennusvuosi oli sitä vastoin painoarvoltaan korkealla merkityksellä. Voidaankin pohtia, onko uudiskohteen merkitsemisellä merkitystä koneoppimismallille ollenkaan vai onko tietokenttää käytetty väärin, kuten aiemmin kappaleessa 8.2 käytiin läpi.

Toinen mielenkiintoinen huomio kiinnittyy asunnon kuntoon (realEstateCondition). Myöskään tällä muuttujalla ei näyttäisi olevan merkittävää roolia koneoppimismallin ennusteessa tähän työhön rajatulla tutkimusaineistolla.

10 KEHITYSEHDOTUKSET

Tutkimuksessa syvennyttiin tutkimaan Kiinteistönvälitysalan Keskusliitto Ry:n KVKL Hintaseurantapalvelun datan laatua koneoppimisen hyödyntämisessä. Tarkempi otanta piti sisällään Tampereen alueen kerrostaloasuntojen myyntidatan. Tämä otanta osoitti, että datan laatu on pääsääntöisesti hyvällä tasolla, vaikka parannettavaakin löytyi. Tietokanta on kuitenkin niin kattava, että datan puhdistus- ja siivoustoimenpiteillä dataa oli edelleen jäljellä niin paljon, että koneoppimismallin muodostaminen on mahdollista.

Kiinteistönvälitysalan Keskusliitto Ry:n KVKL Hintaseurantapalvelun tärkein valtti onkin kattava tietokanta. Vaikka tietokannan rakenne ja laatu on muuttunut vuosien varrella, voidaan koko tietokantaa kuitenkin hyödyntää jo nykyisellään raportointitarpeisiin. Kaikesta olemassa olevasta datasta aina vuodesta 1999 alkaen voidaan tehdä visualisointeja, analysointia ja hakea ymmärrystä esimerkiksi alueellisten asuntojen hintojen kehitykseen. Näin ollen data ei jää hyödyntämättä, vaikka koneoppimiseen sitä ei käytettäisikään.

Koneoppimismallin kannalta voidaan ottaa huomioon datan objektiivisesti mitattavia ominaisuuksia, kuten datan kattavuus, oikeellisuus tai virheettömyys. Dataaineisto on kehittynyt merkittävästi, jolloin datan kattavuus on myös muuttunut. Näin ollen koneoppimismallin kouluttaminen vanhimpien vuosien datalla on haastavaa datan puutteellisuuden takia. Mikäli arvoja korvattaisiin laskennallisilla tai oletusarvoilla, voisi se vääristää koneoppimismallia, jolloin ennusteet muodostuvat myös virheellisiksi. Mikäli oikeaa tietoa puuttuvien arvojen korvaamiseksi ei ole saatavilla, eräs vaihtoehto koneoppimismallin kouluttamiseen on rajata muuttujien määrä niihin tietokenttiin, joissa tieto on täydennetty. Tärkeintä on kuitenkin pohtia, mikä on koneoppimismallin strategia, mihin kysymyksiin etsitään vastauksia ja minkälaisia muuttujia silloin tarvitaan.

Eräänä haasteena datan laadun osalta voidaan pitää manuaalisesti syötettyä dataa. Kuten tutkimuksessa huomattiin, näppäilyvirheet, väärät tiedot tai runsas vapaatekstikenttien käyttäminen aiheuttaa raportoinnissa ja koneoppimisessa ongelmia. Tätä pystytään kuitenkin ehkäisemään datan rakenteen kontrolleilla. Metadatan oikeanlainen määrittely on tärkeää, jotta data on oikeassa muodossa tai

oikeanlaista. Tämän lisäksi esimerkiksi pakollisten kenttien sekä tietotyypin määrittely on merkityksellistä. Näiden asioiden määritteleminen aloitetaan jo datastrategiaa suunniteltaessa: mitkä tiedot ovat oleellisen tärkeitä, mitä dataa halutaan kerätä ja millaisessa muodossa tieto annetaan. Vapaatekstikenttiä voidaan muuttaa pudotusvalikoiksi, kuten esimerkiksi postinumerolistaus tai lämmönlähde. Myös kentän merkkimäärää voidaan rajata, jolloin esimerkiksi päästään eroon ongelmasta, jossa käyttäjä syöttää samaan kenttään sekä postinumeron että postitoimipaikan.

Syötettyä dataa voidaan myös kontrolloida niin kutsutun mielekkyyden tarkastuksen (sanity check) avulla. Tämä tarkoittaa sitä, että kenttien välisiä yhteyksiä validoidaan esimerkiksi ETL-prosessissa, mikäli sitä ei tehdä jo järjestelmään syötettäessä. Esimerkiksi 28 neliömetrin asunto ei voi pitää sisällään seitsemää huonetta. Tällöin voidaan olettaa käyttäjän syöttäneen jotain väärin. Raja-arvojen, mittareiden ja tarkastettavien muuttujien yhteyden määritteleminen on tällöin oleellisessa osassa. Määrittely voi kuitenkin olla haastava, sillä asuntoja on hyvin erilaisia ja eriarvoisilla alueilla.

Johdonmukaisuutta ja yhdistettävyyttä pystytään kehittämään, mikäli dataan saadaan lisättyä esimerkiksi kiinteistötunnus. Tämä mahdollistaisi usean tietokannan yhdistämisen. Näin ollen myös datan oikeellisuutta voidaan tarkastaa sekä rikastuttaa esimerkiksi maanmittauslaitoksen tietokannasta löytyvällä tiedolla.

Datassa on jo nyt saatavilla kohteen sijaintitieto. Tätä sijaintitietoa on mahdollisuus hyödyntää analysoimalla esimerkiksi palvelujen tai keskusten etäisyyksiä ja sen vaikutusta asuntojen hintojen muodostumiseen. Tämä kuitenkin vaatii datajoukkojen yhdistämistä esimerkiksi kaupunkien tietokantoihin.

Koska data sisältää paljon vapaatekstikenttiä, näiden tietojen hyödyntäminen olisi myös tärkeää. Luonnollisen kielen analysointi mahdollistaa avainsanojen löytämisen, jolloin aineiston hyödyntäminen mahdollistuu uudella tavalla myös raportointi- ja koneoppimisen käyttöön.

Tutkimuksessa havaittiin, ettei tuplatietueita löytynyt viime vuosien ajalta. Näin ollen voidaan olettaa, että tuplatietueiden tarkastus on aineistolle jo olemassa. Kuitenkaan mikäli näin ei ole, voitaisiin tähänkin valjastaa kontrolli esimerkiksi ETL-prosessissa.

Taulukossa 12 on koottu yhteenveto kehitysehdotuksista verraten osa-alueita DAMA UK:n (2013) ulottuvuuksien mukaisesti. Koska data koostuu useasta eri lähteestä, on kehitysehdotukset riippuvaisia myös näiden lähdejärjestelmien toiminnoista. Useimpia muutoksia tarvittaisiin jo datan syöttämisvaiheeseen.

Taulukko 12. Kehitysehdotusaihiot

Aihealue	**Kehitysehdotus**
Datan kattavuus ja oikeellisuus	Puutteelliset arvot • Puuttuvien arvojen korvaaminen, jos tieto saatavilla tai noudettavista olemassa olevien tietojen perusteella • Uuden datan kontrollointi merkkimäärällä tai esimerkiksi rajatulla datajoukolla (pudotusvalikolla), kuten postinumerot ja lämmönlähde • Oletusarvoina puuttuville arvoille ”null”
Datan kattavuus	Luonnollisen kielen (NLP) hyödyntäminen vapaatekstikenttien analysoinnissa.
Oikeellisuus	Mielekkyyden tarkastus (sanity check) • Mittareiden asettaminen • Poikkeamien havainnointi • Metadatakontrolli
Datan kattavuus, oikeellisuus & johdonmukaisuus	Tietojen rikastuttaminen / oikeellisuuden tarkastaminen • Esim. maanmittauslaitos, rakennuksen perustiedot
Johdonmukaisuus	Lisähyödyntämisen mahdollisuus lisäämällä aineistoon etäisyyksiä esim. koulu, kirjasto, keskusta
Ainutlaatuisuus	Tuplatietuekontrollin varmistaminen

Ajankohtaisuus & oikea-aikaisuus	Data on tarpeeksi nopeasti ja oikea-aikaisesti muiden hyödynnettävissä. Käyttäjäkysely / sidosryhmien mielipiteen selvitys

Tämä tutkimus pohjautui DAMA UK:n objektiivisiin mittareihin. Jatkotutkimuksena voidaan suositella subjektiivisten ominaisuuksien selvittämistä, kuten kaipaavatko sidosryhmäläiset jotain dataa tarkemmin esitettynä tai monipuolisemmin kerättynä. Käyttäjätutkimuksen perusteella voidaan saada selville subjektiivista tietoa tietokannan hyödynnettävyydestä. Näin ollen kehitysaihiona voidaan pitää tällaisen käyttäjätutkimuksen teettämistä ja pohtia sen pohjalta tarvitaanko datalle muutoksia. Esimerkiksi voitaisiin tutkia, tarvitaanko muuttujien lisäystä, kuten ilmanvaihtojärjestelmän tai viilennyksen tietokenttiä, tai onko data tarpeeksi nopeasti käyttäjien saatavilla.

11 POHDINTA

Tämän kappaleen tarkoitus on koota yhteen tutkimuskokonaisuus, pohtia opinnäytetyöprosessin onnistumista ja arvioida saavutettuja tuloksia. Lisäksi tässä yhteydessä otetaan kantaa työn luotettavuuteen ja eettisiin näkökulmiin. Lopuksi vastataan asetettuihin tutkimuskysymyksiin.

11.1 Opinnäytetyön kokonaisuuden arviointi

Tämän työn tavoitteena oli lisätä ymmärrystä datan laadusta ja laatuun vaikuttavista tekijöistä. Tarkoitus oli antaa lukijalle tiivis ja selkeä kuva aihealueeseen ja syventää kokonaisuutta erityisesti koneoppimisen näkökulmasta. Tutkimuksen kautta oli tarkoitus osoittaa, mitkä ulottuvuuksista ovat koneoppimisen kannalta olennaisimpia ja miten data-aineistoja voidaan esikäsitellä, jotta koneoppimismallien ennustetarkkuus paranisi.

Tämä opinnäytetyö toteutettiin parityönä ja aihealueen tiedettiin olevan laaja. Molemmat tutkivat ovat olleet työelämän ja opintojen kautta tekemisissä erilaisten datahaasteiden parissa ja tällä työllä tutkijat halusivat syventää ymmärrystä näiden haasteiden takana. Datan laatuun haluttiin kytkeä koneoppiminen, sillä sen käyttö liiketoiminnassa tulee yleistymään ja tutkimuksia datan laadusta nimenomaan koneoppimisen näkökulmasta on tehty Suomessa vasta vähän. Työn toimeksiantajana toimi Kiinteistönvälitysalan Keskusliitto Ry, jonka KVKL Hintaseurantapalvelun tietokannan datan laatua tutkittiin asunnon hinnan ennustamisessa koneoppimisen näkökulmasta. Kumpikaan tutkijoista ei ole työsuhteessa toimeksiantajaan. Tästä syystä jo työn alussa päätettiin rajata tutkimusosuuden osalta subjektiiviset datan laadun ulottuvuudet työn ulkopuolelle, sillä niiden arvioiminen olisi vaatinut syvempää ymmärrystä organisaatiosta, sen asiakkaista ja heidän tarpeistaan. Toisaalta organisaation ulkopuolisina tutkijoina ennakko-oletuksien ja ennenaikaisten päätelmien tekemisen riski on pienempi, joka lisää tutkimuksen luotettavuutta.

Työn laajuus ja monimuotoisuus kuitenkin yllättivät työn edetessä ja ajoittain oli hankalaa rajata tutkimusaluetta oikeasta kohtaa. Työstä saatiin koottua hyvä kokonaisuus, vaikkakin tietyt osa-alueet datan laadun ympärillä olisivat tarvinneet laajemman käsittelyn niiden merkittävyyden kannalta. Esimerkiksi datan laadun arviointimallit ovat datastrategian kannalta hyödyllisiä ja tärkeitä, mutta niiden tarkempi käsittely oli rajattava tämän työn ulkopuolelle. Muuten kokonaisuus olisi kasvanut turhan laajaksi. Näin ollen kaikkia datan laatuun vaikuttavia osa-alueita ei pystytty ottamaan mukaan työhön, vaikka alun perin niin ajateltiin. Lisäksi haasteita ilmeni teoriaosuuden ja tutkimusosuuden kytkemisessä toisiinsa, sillä kaikkia teoriaosuudessa esille nostettuja datan laatuun vaikuttavia tekijöitä ei pystytty kytkemään toimeksiantajalle toteutettuun tutkimusosuuteen.

Kokonaisuus antaa kuitenkin lukijalle hyvän kuvan siitä, kuinka laajasta kokonaisuudesta on kyse ja siitä millaiset asiat vaikuttavat datan laatuun ja millaisilla toimenpiteillä laatuun voidaan vaikuttaa. Teoriaosuuteen saatiin myös selkeästi tiivistettyä tärkeimmät ulottuvuudet ja se, miten kannattaa mitata datan laadun jatkuvaa toteutumista. Tämä antaa lukijalle hyvän pohjan edetä syvemmälle datan laadun maailmaan.

11.2 Tutkimustulosten arviointi

Tutkimuksessa perehdyttiin dataan, sen esikäsittelyyn ja analysointiin. Koneoppimisen ennustemalleja arvoitiin R2 ja MAE-tunnuslukujen avulla. Lisäksi tutkittiin kuinka esikäsittelytoimenpiteet vaikuttavat eri muuttujien merkitsevyyteen asunnon hinnan ennustamisessa. Kuten itse tutkimusosuudessa kävi ilmi, ennustemallit ennustivat jopa 90 %:n tarkkuudella asunnon hinnan oikein. Tätä voidaan pitää hyvänä tuloksena. Ennustetarkkuutta saataisiin todennäköisesti parannettua, mikäli datan kattavuutta saataisiin nostettua sekä koneoppimismallien parametrointia optimoitua. Näin koneoppimismallia pystyttäisiin hyödyntämään myös hintahaarukan ääripäissä.

Muuttujien merkitsevyyden kannalta pystyttiin osoittamaan, että esikäsittelytoimenpiteillä on vaikutusta eri muuttujien painotukseen asunnon hinnan ennusta-

misessa. Tässä yhteydessä todettiin, ettei muuttujilla sauna ja parveke ollut juurikaan merkitystä hinnan muodostumiselle tehtyjen testien osalta. Tutkimuksessa testattiin koneoppimismalleja kolmella eri tavalla esikäsitellyllä tutkimusdatalla. Kolmannesta aineistosta jätettiin nämä muuttujat pois kokonaan. Mikäli testiä olisi jatkettu syvemmälle, olisi ollut mielenkiintoista toteuttaa vielä neljäs testiaineisto, joka olisi pitänyt sisällään rajatun aineiston Tampereella myydyistä kerrostalohuoneistoista vuodesta 2016 eteenpäin, jolloin myös parveke ja sauna -muuttujat olisi saatu mukaan aineistoon kattavasti käytettynä. Näin olisi pystytty osoittamaan tietokannan nykyisen rakenteen käytettävyys koneoppimismalleissa ja selvittämään, olisiko näillä muuttujilla ollut todellisuudessa suurempi painoarvo, kun nyt tehnyt testit antavat osoittaa. Tätä työtä toimeksiantaja voi jatkaa tämän opinnäytetyön pohjalta.

Tämä tutkimus osoitti myös, kuinka tärkeää on tuntea tutkimusdata ennen sen käyttämistä. Sen lisäksi tutkimus osoittaa, että muutokset tietokannan muuttujiin ja rakenteeseen tulee tehdä johdonmukaisesti ja harkitusti, sillä nämä muutokset vaikuttavat aineiston käyttöön. Myös aineiston kuvauksen yhteydessä pidettävää muutoslokia, joka kertoo aineiston käyttäjille tietokantaan tehdyistä muutoksista ja niiden ajankohdista, tulisi aina pitää ja tehdä se mahdollisimman selkeästi ja käyttäjää ohjaavasti. Kiinteistönvälitysalan Keskusliitto Ry:n KVKL Hintaseurantapalvelun kohdalla aineistonkuvauksen yhteydessä on pidetty aineiston käyttäjän saatavilla olevaa muutoslokia vuodesta 2019 alkaen, joka tulevaisuudessa ohjaa aineiston käyttäjää ja nopeuttaa aineiston omaksumista.

Tutkimusdata rajattiin koskemaan yhden kaupungin kerrostalohuoneistoja, joiden hinnan ennustaminen on ilmankin koneoppimismallin apua varsin suoraviivaista. Mikäli koneoppimismallia hyödynnettäisiin suuremman alueen ja erilaisten asuinmuotojen ennustamiseen, nousevat kohdetta yksilöivät muuttujat tärkeämmäksi. Näin ollen tämän tutkimuksen tuloksia ei voida pitää yleistettävinä. Niiden tarkoitus on nostaa esiin ymmärrys sen taakse, miten erilainen esikäsittely vaikuttaa aineistoon ja siitä saataviin tuloksiin. Tämä tutkimus kuitenkin osoitti, että jo varsin maltillisilla esikäsittely toimenpiteillä voidaan vaikuttaa koneoppimismallin tarkkuuteen.

11.3 Tutkimuskysymyksiin vastaaminen

Tämän opinnäytetyön ensimmäinen tutkimuskysymys oli: ”kuinka koneoppiminen muuttaa datan laatuvaatimuksia?” Kysymystä täsmentämään nostettiin kaksi alakysymystä: ”kuinka datan laatu vaikuttaa koneoppimismallien tarkkuuteen?” ja ”kuinka datan laadullisiin ominaisuuksiin voidaan vaikuttaa?”.

Koneoppimisessa ennustemalli on niin hyvä kuin sille syötettävä data on. Vääränlainen tai puutteellinen data voi vääristää ennustemalleja, jolloin muodostuva ennuste voi johtaa täysin virheellisiin lopputuloksiin. Datan laatuun vaikuttavien tekijöiden suunnittelu lähtee datastrategian noudattamisesta sekä kestävän datapohjan rakentamisesta. On tärkeää määritellä jo datastrategiasta lähtien datan tavoite ja tarkoitus. Myös tulevaisuuden tarpeet kannattaa huomioida mahdollisimman kattavasti. Toimenpiteet tähtäävät hyvälaatuiseen dataan, joka on mahdollisimman hyödyntämiskelpoista sellaisenaan. Laadullisiin ominaisuuksiin pystytään vaikuttamaan erilaisin keinoin, kuten metadatan avulla. Vapaatekstikenttien käyttäminen on koneelle haaste, mutta ei este. Kuitenkin mikäli konetta pystytään kouluttamaan määritellyllä aineistolla, vältytään massiivisilta datan esikäsittelytoimenpiteiltä. Tiivistettynä voidaan todeta, että on tärkeää kerätä oikeanlaista dataa oikeanmuotoisena oikeaan aikaan.

Toisena tutkimuskysymyksenä oli: ”minkälainen liiketoiminnallinen merkitys datan laadulla on?” Tämän opinnäytetyön ehkä tärkein sanoma on, että datan keräämiseen ja sen laatuun kannattaa aloittaa kiinnittämään entistä enemmän huomiota heti kun mahdollista. Lähes kaikki liiketoiminta on muuttunut ja muuttumassa suuntaa, jossa ne yritykset ja organisaatiot, jotka keskittyvät laadukkaan datan keräämiseen ja siitä saatavan tiedon hyödyntämiseen, tulevat tulevaisuudessa saamaan merkittävän liiketoiminta hyödyn. Datan merkitys korostuu entisestään tekoälyn aikakaudella ja se tulee koskemaan myös perinteisiä toimialoja. Ongelmat datan laadussa voivat johtaa merkittäviin ongelmiin ja tulonmenetyksiin liiketoiminnassa. Pahimmillaan huonolaatuinen data voi keskeyttää liiketoiminnan kokonaan.

LÄHTEET

Aaltio-Marjosola, I. 1999. Case tutkimus metodisena lähestymistapana. Metodix Oy. Luettu 7.1.2021. https://metodix.fi/2014/05/19/aaltio-marjosola-casetutkimus/

Ailisto, H., Heikkilä, E., Helaakoski H., Neuvonen, A., Seppälä, T. 2018. Tekoälyn kokonaiskuva ja osaamiskartoitus. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja. Valtioneuvoston kanslia. Luettu 25.11.2020. https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160925/46-2018-Tekoalyn%20kokonaiskuva.pdf

Amazon 2020. What is a data lake? Luettu 2.12.2020. https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/

Ari Hovi. 2020. Tietovarasto: paikallinen vai Enterprise Data Warehouse? Luettu 25.1.2020. https://www.arihovi.com/paikallinen-vai-enterprise-data-warehouse/

Azeroual, O., Saake, G., Schallen, E. 2018. Analyzing data quality issues in research information systems via data profiling. International Journal of Information Management (41), 50-56.

Bakshi, R. 2020. Random Forest Regression. GitConnected. Luettu 28.12.2020 https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Bhansali N. 2013. Data Governance. CRC Press, Auerbach Publications.

Brownlee J. 2019. How to use Learning Curves to Diagnose Machine Learning Model Performance. Machine Learning Mastery. Luettu 27.1.2021. https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/

Broända R., Maikkola K. & Rönkkö K. 2020 Kokemuksia datahankkeista & moderni analytiikka-alusta. IBM-webinaari 5.11.2020.

DAMA International, 2010.The DAMA Guide to the Data Management Body of Knowledge (DAMA DM-BOK). Ensimmäinen painos. Technics Publications.

DAMA UK, 2013. The six primary dimensions for data quality assessment. DAMA UK Working Group. Luettu 3.11.2020. https://damauk.wildapricot.org/resources/Documents/DAMA%20UK%20DQ%20Dimensions%20White%20Paper2020.pdf

Donges, N. 2019. Complete guide to the random forest algorithm. Luettu 7.1.2021 https://builtin.com/data-science/random-forest-algorithm.

Elements of AI, 2018. University of Helsinki and Reaktor. Luettu 25.11.2020 https://course.elementsofai.com

English, L. P., 1999. Improving Data Warehouse and Business Information Quality. New York: Wiley.

Gollapudi S. 2016. Practical Machine Learning, United Kingdom: Packt Publishing Ltd.

Hannila. H. 2020. Data-driven alkaa sanalla DATA. CGI Inc. Blogi. Luettu 25.4.2021. https://www.cgi.com/fi/fi/blogi/data-driven-alkaa-sanalla-data

Hannila H. 2019. Towards data-driven decision-making in product portfolio management. University of Oulu. Faculty of Technology. Luettu 3.11.2020. http://jultika.oulu.fi/files/isbn9789526224428.pdf

Hovi A., Hervonen H., Koistinen H., 2009. Tietovarastot ja business intelligence. Jyväskylä: WSOYpro/Docendo-tuotteet.

Hämäläinen M., 2021. Kehitysjohtaja. Haastattelu 25.1.2021. Haastattelijat Hulkkonen, P., Raunnos, E. Litteroitu. Etätapaaminen.

Hämäläinen, W. 2013. Tiedonlouhinta. Luettu 20.1.2021. http://cs.joensuu.fi/pages/whamalai/DM13/kalvot2_4per1.pdf

Kallio J. 2018. Neuroverkot analytiikan edistäjinä. Insta Blogi. Luettu 27.12.2020. https://www.insta.fi/ajankohtaista/neuroverkot-analytiikan-edist%C3%A4jin%C3%A4

Itewiki n.d. Digitalisoinnin opas – Tekoäly. Luettu 25.11.2020 https://www.ite-wiki.fi/opas/tekoaly/

Joutsijoki H. 2017. Koneoppiminen. Luettu 26.1.2021. https://coss.fi/wp-content/uploads/2017/12/4-Koneoppiminen.pdf

Kananen, H., Puolitaival H. 2019. Tekoäly – Bisneksen uudet työkalut. 1. painos. Helsinki: Alma Talent Oy

Kilkenny, M., Robinson, K. 2018. Data quality: “Garbage in – garbage out”. Health Information Management Journal 47(3), 103–105.

Kiinteistönvälitysalan Keskusliitto Ry n.d. a. Liiton tavoitteet. Luettu 15.2.2021. https://kvkl.fi/meista/toiminta-ja-tavoitteet/

Kiinteistönvälitysalan Keskusliitto Ry n.d. b. Hintaseurantapalvelu. Luettu 15.2.2021. https://kvkl.fi/tietopankki/hintaseurantapalvelu/

Kiinteistönvälitysalan Keskusliitto Ry n.d. c. Liiton tavoitteet. Luettu 23.5.2021. https://kiinteistonvalitysala.fi/meista/

Kiinteistönvälitysalan Keskusliitto Ry KVKL Hintaseurantapalvelu, n.d. Luettu 15.2.2021. https://www.hintaseurantapalvelu.fi/#/login

Korpela J. 2018. Mitä on tiedon laatu? Aureolis Oy. Webinaari 16.5.2018. Katsottu 7.12.2020. https://aureolis.com/bi-akatemia/kiitos-webinaaritallenne-tiedon-laatu/

Laatikainen, T. 2015. Master Data on monisyistä, monista syistä. Talent Base. Luettu 5.11.2020. https://www.talentbase.fi/blogi/master-data-on-monisyista-monista-syista/

Laihonen, H., Hannula, M., Helander, N., Iivonen, I., Jussila, J., Kukko, M., Kärkkäinen, H., Lönnqvist, A., Myllärniemi, J., Pekkola, S., Virtanen, P., Vuori, V., Yliniemi, T. 2013. Tietojohtaminen. Tampereen teknillinen yliopisto, Tietojohtamisen tutkimuskeskus Novi.

Laine M., Bamberg J., Jokinen P. 2007. Tapaustutkimuksen taito. Helsinki: Gaudeamus.

Mahanti, R. 2019. Data Quality. Quality Press.

McGilvray, D. 2008. Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information. Morgan Kaufmann Publishers.

Merilehto, A. 2018. Tekoäly – Matkaopas johtajalle. 1. painos. Helsinki: Alma Talent Oy

Merilehto, A. 2019. Tekoäly nyt - podcast. Jakso 4: Data on lähtökohta. Suomen podcastmedia.

Merilehto A., Hagman K. 2021. Tekoäly nyt – podcast. Jakso 24: Tekoälyprojektin sudenluopat. Suomen podcastmedia.

Microsoft, 2021. Microsoft AI Documentation. Luettu 19.1.2021. https://docs.microsoft.com/en-us/ai/

MKL, 2019. Making AI Simple. Data Preprocessing in Machine Learning – Complete Nutshell view for Beginners. Luettu 20.1.2021. https://machinelearningknowledge.ai/data-preprocessing-in-machine-learning/

Mueller J., Massaron L. 2016. Machine Learning for Dummies. For Dummies.

Niemi K. n.d. Leiska Contens Oy. Data-Suomi-Sanakirja. Termiviidakon selviytymisopas. https://www.slideshare.net/kalleniemi1/data-suomi

Redman, T. 2001. Data Quality. The Field Guide. Boston: Digital Press.

Redman, T. 2018. If Your Data Is Bad, Your Machine Learning Tools Are Useless. Luettu 26.10.2020. https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless

Ruokolainen L. 2018. Ylisovittaminen ja kuinka sen kanssa voi tulla toimeen. Bilot Group Blogi. Luettu 20.2.2021. https://bilot.group/articles/ylisovittaminen-ja-kuinka-sen-kanssa-voi-tulla-toimeen/

Saaranen-Kauppinen A. & Puusniekka A. 2006. KvaliMOTV - Menetelmäopetuksen tietovaranto. Tampere: Yhteiskuntatieteellinen tietoarkisto. Luettu 7.1.2020. https://www.fsd.tuni.fi/menetelmaopetus/

Sebastian-Coleman, L. 2013. Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework. Morgan Kaufmann.

Seppälä S. 2021. Mitä väliä on datan laadulla? CGI Inc. Blogi. Luettu 25.4.2021. https://www.cgi.com/fi/fi/blogi/mita-valia-on-datan-laadulla

Tamminen, S. 2019. Datan rooli koneoppimisessa ja tekoälyssä. Luettu 18.1.2021. https://www.slideshare.net/Solita_Oy/datan-rooli-koneoppimisessa-ja-tekolyss

Tilastokeskus, 2018. Asunto-osakeyhtiöiden hoitokulut laskivat vuonna 2017. Luettu 14.1.2021. http://www.tilastokeskus.fi/til/asyta/2017/asyta_2017_2018-09-11_tie_001_fi.html

Tilastokeskus, n.d. Käsitteet: Lämmitystapa. Luettu 11.1.2021. https://www.stat.fi/meta/kas/lam_tapa.html

Törmänen, A. 2017. Johdanto Tietovarastointiin. A. Törmänen.

Uncovering AI in Finland 2018. Field guide to AI. Microsoft & PwC. Luettu 8.12.2020. https://info.microsoft.com/WE-AzureDS-CNTNT-FY18-04Apr-17-UncoveringAIinFinland-MGC0002305_01Registration-ForminBody.html

Väre O. 2019. MASTER DATAN HALLINNAN KÄSIKIRJA liiketoiminnan kehittäjille ja päättäjille. Alma Talent Oy.

Willeams K. 2019. Keras Tutorial: Deep Learning in Python. Data Camp. Luettu 27.1.2021. https://www.datacamp.com/community/tutorials/deep-learning-python