Tampereen yliopisto

Mikko Pääkkönen

DATATIETEEN HYÖDYNTÄMINEN P2P-LAINAPÄÄTÖKSISSÄ

Kandidaatintyö
Johtamisen ja talouden tiedekunta
Tarkastaja: Ilona Ilvonen
Toukokuu 2022

TIIVISTELMÄ

Mikko Pääkkönen: Datatieteen hyödyntäminen P2P-lainapäätöksissä
(Data science in peer-to-peer lending decisions)
Kandidaatintyö
Tampereen yliopisto
Tietojohtamisen tutkinto-ohjelma
Toukokuu 2022

Lainamarkkinoilla on tapahtunut muutoksia edellisten vuosikymmenien aikana. Vuoden 2008 finanssikriisi herätti epäluottamusta pankkisektorille ja täten alalle synty uusia toimijoita kuten vertaislainapalveluita tuottavia verkkoalustoja. Vertaislainaaminen (Peer-to-peer lending) tarkoittaa suoraa lainaamista, joka tapahtuu yksilöiden välillä ilman virallisia instituutioita verkkopohjaisten alustojen kautta. P2P-laina-alustat kohtaavat toiminnassaan kuitenkin merkittäviä epäsymmetrisen informaation ongelmia, jotka voivat johtaa sijoittajien huonoihin lainapäätöksiin. Datatieteellä (Data science) on keskeinen rooli P2P-alustojen toiminnassa. Datatiede mahdollistaa datasta tiedon luomisen päätöksentekoa varten. Tässä kirjallisuuskatsauksessatutkimuksessa tarkastellaan, miten datatiedettä voidaan hyödyntää P2P-lainapäätöksissä.

Tutkimus toteutettiin kirjallisuuskatsauksena käyttäen aineistona alan julkaisuja. Tutkimuksen teoriaosuus jakautuu datatieteeseen ja P2P-lainaamiseen. Datatieteen osuudessa käsitellään datatiedettä käsitteenä ja kokonaisuutena sekä tarkastellaan dataa, datalähteitä ja koneoppimista osana datatiedettä. Toisessa osassa käsitellään P2P-lainaamisen toiminnallisuutta sekä keskeisiä ongelmia. Tutkimuksen viimeisessä osassa tutkitaan, miten datatiedettä voidaan hyödyntää P2P -lainapäätösten yhteydessä.

Tutkimuksessa havaittiin, että datatiedettä voidaan hyödyntää P2P-lainapäätöksissä. Tutkimuksen tuloksista huomattiin, että datatiedettä voidaan hyödyntää etenkin lainanhakijoiden ja laina-alustan välisen epäsymmetrisen informaation ongelman ratkaisemisessa. Datatieteen avulla laina-alusta voi arvioida lainanhakijoiden luottoriskiä, mikä tuottaa tietoa lainanmyöntäjien lainapäätöksentekoon. Lainanhakijoiden arvioinnissa voidaan hyödyntää laajasti eri rakenteellista ja rakenteetonta dataa. Rakenteellista dataa ovat etenkin lainanhakijan taloudellinen data kuten luottopisteytys ja tulot. Rakenteetonta dataa ovat puolestaan lainahakemusten tekstit sekä lainanhakijoiden sosiaalisen median julkaisut. Dataa lainanhakijoista voidaan saada sekä alustan sisäisistä datalähteistä kuten lainahakemuksista että ulkoisista datalähteistä kuten sosiaalisesta mediasta ja avoimista datalähteistä. Lainapäätöksissä voidaan hyödyntää etenkin ohjattua oppimista, jonka avulla voidaan ennustaa datasta lainanhakijoiden luottoriskiä. Lisäksi ohjaamattoman oppimisen avulla lainanhakijoiden teksteistä voidaan erotella luottoriskin arvioinnissa hyödyllisiä ominaisuuksia.

Avainsanat: Datatiede, data, datalähteet, koneoppiminen, P2P-lainaaminen, lainapäätös

SISÄLLYSLUETTELO

1. JOHDANTO 1
    1.1 Tutkimuksen taustoitus ja aihe 1
    1.2 Tutkimuksen rajaus ja tutkimuskysymykset 2
    1.3 Tutkimuksen rakenne 3
2. TUTKIMUKSEN TOTEUTUS 4
    2.1 Tutkimusmenetelmä 4
    2.2 Tutkimusaineisto 7
3. DATATIEDE 9
    3.1 Datatiede käsitteenä 9
    3.2 Datatiede prosessina 9
    3.3 Data ja datalähteet 11
    3.4 Koneoppiminen 15
4. P2P-LAINAAMINEN JA -ALUSTAT 18
    4.1 P2P-lainaamisen määritelmä 18
    4.2 P2P-lainaamisen ja -alustojen toiminta 18
    4.3 P2P-lainaamisen keskeiset ongelmat 20
5. DATATIEDE JA P2P-LAINAPÄÄTÖKSET 22
    5.1 Datatieteen käyttö P2P-lainaamisessa 22
    5.2 Data ja datalähteet P2P-lainapäätöksissä 23
    5.3 Koneoppiminen P2P-lainanhakijoiden arvioinnissa 26
6. YHTEENVETO 27
    6.1 Tulokset 27
    6.2 Tutkimuksen arviointi ja jatkotutkimusehdotukset 30
LÄHTEET 31

.

1. JOHDANTO

1.1 Tutkimuksen taustoitus ja aihe

Lainamarkkinoilla on tapahtunut muutoksia edellisten vuosikymmenien aikana. Zwilling et al. (2020, s. 854) vuoden 2008 finanssikriisi herätti epäluottamusta pankkisektorille ja täten alalle syntyi uusia toimijoita kuten vertaislainapalveluita tuottavia alustoja. Vertaislainaamisella (Peer-to-Peer lending) tarkoitetaan CFI:n (2022) mukaan lainaamista, joka tapahtuu suoraan eri yksilöiden välillä ilman virallisia finanssi-instituutioita verkkopohjaisten alustojen kautta. P2P-laina-alustat kohtaavat kuitenkin lainanhakijoita arvioidessaan merkittäviä epäsymmetrisen informaation ongelmia (Lenz 2016, s. 693). Lainanhakijoiden epäonnistunut arviointi voi johtaa lainanmyöntäjien huonoihin P2P-lainapäätöksiin.

Giudicin (2018, s. 161) mukaan datatiede (Data science) on keskeisessä roolissa osana P2P-alustojen toimintaa ja valtavaa kasvua. Datatieteelle on olemassa useita eri määritelmiä, mutta esimerkiksi Caon (2017, s. 8) mukaan se tarkoittaa alaa, joka yhdistää erilaisia tieteenaloja kuten tilastotiedettä, tietotekniikkaa, viestintää, johtamista sekä sosiologiaa luodakseen datasta hyödyllistä tietoa päätöksentekoa varten. Toisaalta datatiede tarkoittaa Kotun & Deshpanden (2018 luku 1) mukaan erilaisten tekniikkojen kokoelmaa, joiden avulla voidaan luoda datasta arvoa. Lisäksi Sarkerin (2021b, s. 3–5) mukaan datatiede on ikään kuin kattotermi monelle keinolle kuten koneoppimiselle, joiden tavoitteena on luoda datasta hyödyllistä tietoa liiketoiminnalliseen käyttöön. Lähes kaikissa datatieteen määritelmissä yhdistyy kuitenkin ajatus siitä, että sen avulla voidaan tuottaa datasta tietoa päätöksentekotilanteita varten.

P2P-alustat tekevät riskiarviot lainanhakijoista, minkä perusteella lainanmyöntäjät tekevät lainapäätöksiä. Lisäksi P2P-alustat eivät kykene muodostamaan tavallisten pankkien kaltaista luottamusta asiakkaisiinsa, joka johtaa alustojen tarpeeseen arvioida lainanhakijoita eri tavoin. (Lenz 2016, s. 693) Tässä tutkimuksessa tarkastellaan, miten datatiedettä voitaisiin hyödyntää P2P-lainapäätöksissä. Datatieteen kokonaisuuden avulla Caon (2017, s. 8) mukaan datasta voidaan tehdä esimerkiksi löydöksiä, ennusteita, suosituksia tai päätöksenteon ohjeistuksia. Valtavista datamassoista voidaan löytää erilaisia havaintoja käyttäen datatieteessä sovellettuja koneoppimismalleja (Kashyap 2017, luku. 1).

Datatieteen tutkiminen P2P-lainaamisen kontekstissa on siis hyvin keskeistä, sillä datatieteen avulla voisi olla mahdollista ratkaista esimerkiksi P2P -lainaamiseen liittyviä epäsymmetrisen informaation ongelmia. Tutkimuksen avulla voitaisiin myös saada selville, millaista dataa ja datalähteitä P2P-lainanhakijoista voitaisiin käyttää, jotta lainanmyöntäjät pystyisivät tekemään onnistuneita lainapäätöksiä. Lisäksi tutkimus voisi tarjota laajempaa näkymää datatieteen käytöstä myös muille lainasektorin toimijoille kuten tavallisille liikepankeille.

1.2 Tutkimuksen rajaus ja tutkimuskysymykset

Tutkimuksen aihe käsittelee sekä datatiedettä että P2P-lainaamista. Datatiede on kokonaisuutena äärimmäisen laaja käsite ja työssä käsitellään täten aihepiiriä pintapuolisesti. Kuitenkin aiheessa keskitytään juuri datan, datalähteiden ja koneoppimisen merkitykseen datatieteessä, joten kyseisiä aiheita tarkastellaan tarkemmin osana datatieteen kokonaisuutta. Koska tutkimus tehdään kandidaatintyön laajuudessa ja se keskittyy datatieteeseen päätöksenteon näkökulmasta, erilaisia teknisiä yksityiskohtia ei ole syytä tarkastella kovin yksityiskohtaisesti. Lisäksi vertaislainaamiseen liittyviä yksityiskohtia kuten sääntelyä ei tulla käsittelemään kovinkaan laajasti.

Vertaislainoissa keskitytään kuluttajiin eli yksityishenkilöille myönnettäviin lainoihin, mikä rajaa työstä yrityslainat pois rajaten aihekokonaisuutta kapeammaksi ja täten helpommin tutkittavaksi. Lisäksi lainapäätöstentekoa tarkastellaan lainaa myöntävien sijoittajien ja lainaamiseen suunniteltujen alustojen näkökulmasta. Kyseisessä näkökulmassa on keskeistä kuvailla, miten laina-alusta voi hyödyntää datatiedettä, jotta alustalla sijoittavat lainanmyöntäjät voivat tehdä lainapäätöksiä itselleen sopivalla riskitasolla. Aihekokonaisuutta käsitellään kokonaisuudessaan kandidaatintyön sopivassa laajuudessa, jolloin tietyt tekniset yksityiskohdat ja tarkennukset jätetään pois.

Tutkimuksen keskeisenä tavoitteena on pyrkiä tuomaan esiin, miten datatieteitä voidaan hyödyntää P2P-lainapäätöksissä. Kyseinen aihe on myös tutkimuksen keskeinen tutkimusongelma, johon pyritään vastaamaan keskeisessä tutkimuskysymyksessä:

- Miten datatiedettä voidaan hyödyntää P2P-lainapäätöksissä?

Lisäksi tutkimukselle on valittu alatutkimuskysymykset, jotka pyrkivät täydentämään ja tukemaan tutkimuksen pääkysymystä:

- Mitä datatiede tarkoittaa?
- Millaista dataa, datalähteitä ja koneoppimista voidaan käyttää datatieteessä?
- Mitä P2P-lainaaminen tarkoittaa ja miten se toimii?

- Millaista dataa, datalähteitä ja koneoppimismalleja voidaan hyödyntää P2P-lainapäätöksissä?

Tutkimuksessa on tärkeää vastata myös alatutkimuskysymyksiin, sillä ne pohjustavat päätutkimuskysymykseen liittyviä tärkeitä käsitteitä kuten datatiedettä ja P2P-lainaamista ja alustoja. Viimeisin alatutkimuskysymyksen avulla voidaan myös tukea ja konkretisoida päätutkimuskysymykseen vastaamista.

1.3 Tutkimuksen rakenne

Tutkimus jaettiin kolmeen keskeiseen osaan. Tutkimuksen ensimmäisessä osassa eli luvuissa yksi ja kaksi pyritään täsmentämään tutkimusta ja sen taustaa. Luvussa yksi kuvaillaan tutkimuksen aihepiirien taustaa ja perustellaan tutkimuksen tutkimuskohde. Lisäksi kyseisessä luvussa esitellään tutkimuksen pää- ja alakysymykset. Luvussa kaksi täsmennetään tutkimukseen valittu kirjallisuuskatsausmalli ja tarkastellaan tutkimusaineistoa.

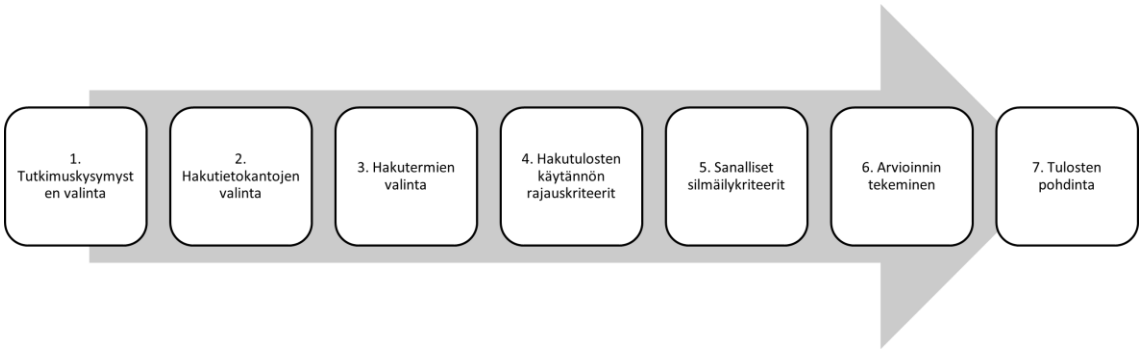
Tutkimuksen toisessa osassa taustoitetaan tutkimuksen keskeistä teoriaa ja vastataan ensimmäiseen ja toiseen alatutkimuskysymykseen. Luvussa kolme määritellään datatiede käsitteenä ja sitä kuvaillaan prosessina. Lisäksi kyseisessä luvussa määritellään datan, datalähteiden ja koneoppimisen käyttö datatieteessä. Tutkimuksen toinen osuus jatkuu luvussa neljä, jossa vastataan kolmanteen alatutkimuskysymykseen, mikä täsmentää P2P-lainaamista ja sen yleisiä ongelmia.

Tutkimuksen kolmannessa osassa eli luvussa viisi vastataan tutkimuksen pääkysymykseen sekä viimeiseen alatutkimuskysymykseen. Luvussa viisi esitetään tutkimuksen kirjallisuuskatsausmallin avulla saatuja vastauksia. Kyseinen luku yhdistää siis datatieteen ja P2P-lainaamisen kokonaisuutta vastaamalla edellä mainittuihin kysymyksiin. Lisäksi tutkimuksen ja täten etenkin luvun viisi tulokset esitetään työn yhteenvedossa. Yhteenvedossa tarkastellaan myös tutkimuksen aiheen ja tuloksien kannalta oleellisia jatkotutkimuskysymyksiä ja vaihtoehtoisiä näkökulmia tutkimuksen aihepiirin kannalta.

2. TUTKIMUKSEN TOTEUTUS

2.1 Tutkimusmenetelmä

Tutkimus toteutetaan oheisessa kuvassa 1 esitellyn Finkin (2019, s. 5) systemaattisen kirjallisuuskatsaus mallin mukaisesti. Kyseisen menetelmän tavoitteena on parantaa tutkimuksen toistettavuutta ja lisäksi se tekee keskeiseen tutkimuskysymykseen vastaamisesta järjestelmällistä ja suoraviivaista. Menetelmää hyödynnetään tarkasti viidennessä luvussa, jossa vastataan päätutkimuskysymykseen.



Kuva 1: *Finkin (2019) systemaattisen kirjallisuuskatsauksen malli.*

Työn keskeiset tutkimuskysymykset on esitetty edellisessä luvussa. Hakutietokannoiksi on valikoitunut Andor ja Web of Science, sillä kyseisistä palveluista löytyy runsaasti erilaisia artikkeleita, kirjoja ja muita tieteellisiä julkaisuja. Etenkin Andor yhdistelee eri tietokantoja laajasti, mikä hajauttaa hakutuloksia eri tietokantoihin. Puolestaan Web of Science tarjoaa mahdollisuuden tehdä täsmällisiä hakuja juuri kyseiseen palveluun. Lisäksi tiedonhaussa käytetään myös relevantteja nettisivuja kuten esimerkiksi finanssialan termejä määrittelevää Investopediaa tukemaan teoriaosuutta.

Hakutermit kuvan 1 kohdassa kolme valittiin työn keskeisten kysymysten ja käsitteiden perusteella. Hakutermit koostuivat erilaisista käsitteiden yhdistelmistä käyttäen apuna Boolen -operaattoreita ”AND” ja ”OR”. Aiheanalyysin ja tutkimussuunnitelman alustavan analyysin perusteella aihepiirin hakusanoja ja -lausekkeita oli syytä hakea vain englanniksi, sillä esimerkiksi suomeksi hakutuloksia ei saatu merkittäviä määriä. Oheisessa taulukossa 1 esitetään hakutulosten määrä eri tutkimuksen keskeisille käsitteille ja joillekin käsitteiden yhdistelmille, ilman minkäänlaisia hakutietokantojen rajauskriteerejä kuten vuosilukuja tai aineistotyyppejä. Kyseinen taulukko antaa kuvaa yleisesti kyseisten käsitteiden ja aihepiirien esiintymisestä hakutietokannoissa.

***Taulukko 1**: Hakutulosten määrä eri käsitteillä ja hakulausekkeilla*

Hakutietokanta	Hakutermi/lauseke	Tulokset (kpl)
Andor / Web of Science	“Data science”	490 713 / 21 632
Andor	Data AND “data sources” AND “data science”	28 779
Andor	“Machine learning”	2 026 487
Andor	“Peer-to-peer lending” OR “P2P lending”	28 176
Andor	“P2P lending platform”	3033
Andor	“Data science” AND (“Peer-to-peer lending” OR “P2P lending”)	378

Taulukosta 1 nähdään, että etenkin ”data science” ja ”machine learning” tuottivat merkittävän määrän tuloksia. Puolestaan vertaislainaamiseen liittyvät käsitteet ”Peer-to-peer lending” ja “P2P lending platform” eivät tuottaneet edellisiin aihepiireihin verrattuna lähellekään niin paljon tuloksia. Viimeinen haku, joka yhdistää datatiedettä ja vertaislainaamista, tuotti alle 400 tulosta. Tutkimuksessa todettiin hyväksi yhdistää myös muita hakusanoja kuten ”machine learning”, ”data” ja ”data sources” yhdessä vertaislainaamisten käsitteiden kanssa, jotta hakutuloksia saatiin laajennettua. Hakutermeille asetettiin kuitenkin myös erilaisia kriteerejä, joiden avulla hakutuloksia rajattiin tutkimuksen kannalta paremmiksi.

Käytännön hakukriteereissä kohdassa neljä valittiin tarkasteltaviksi kriteereiksi aineiston saatavuus, vertaisarviointi, aineistotyypit, julkaisuvuodet ja haettujen aineistojen tarkasteluun valittava määrä. Tutkimuksessa käytettiin verkossa saatavilla olevia ja vertaisarvioituja julkaisuja. Aineistotyypeiksi valittiin artikkelit, lehdet, kirjat sekä konferenssijulkaisut. Esimerkiksi konferenssijulkaisut voivat sisältää erillisiä artikkeleita, joita ei löydy erillisinä julkaisuina, joten nekin ovat aiheellista sisällyttää mukaan.

Etenkin 2000 -luvun informaatioteknologian merkittävän kehityksen työssä ei käytetty kovin vanhoja lähdeaineistoja. Lisäksi tutkinnan kohteena olevat alat ovat myös kehittyneet viimeisten vuosien aikana hurjasti, joten lähdeaineistoja valittiin tarkasteluun maksimissaan viiden vuoden takaa eli väliltä 2017–2022. Tutkimusta taustoittavassa teoriassa voitiin kuitenkin hyödyntää myös aineistoja, jotka olivat hieman vanhempia.

Koska hakutuloksia löytyi joillekin hakulausekkeille useita satoja, valittiin jokaisen hakulausekkeen hakutuloksista tarkasteluun vain 50 ensimmäistä teosta. Tämä mahdollisti tutkimukselle paremman toistettavuuden mahdollisuuden sekä rajasi tarkasteltavien te-

oksien määrän sopivaksi. Edellä mainittuja käytännön hakukriteerejä hyödynnettiin tutkimuksen viidennen osion, eli päätutkimuskysymyksen ja viimeisen alatutkimuskysymyksen vastaamisen yhteydessä. Teoriaosuuksissa kuten kappaleissa neljä ja viisi oli tarpeellista käyttää kriteereistä poikkeavia aineistoja kuten relevantteja verkkosivuja esimerkiksi juuri yksittäisten käsitteiden määrittelyyn.

Seuraavassa taulukossa 2 on esitettynä hakutulosten määriä eri hakulausekkeille valittuja käytännön rajauskriteerejä käyttäen. Etenkin kolme ensimmäistä taulukossa esitettyä lauseketta liittyy juuri päätutkimuskysymykseen sekä kahteen ensimmäiseen alatutkimuskysymykseen. Taulukosta 2 havaitaan, että hakutulokset hakulausekkeille tiivistyivät huomattavasti käyttäen rajauskriteerejä. Kyseisillä hakutermeillä näyttää löytyvän riittävästi hakutuloksia, jotta tutkimus voidaan toteuttaa.

Taulukko 2: *Hakutuloksia eri hakulausekkeilla eri hakutietokannoissa.*

Hakutietokanta	**Hakutermi/lauseke**	**Tulokset (kpl)**
Andor	“Data science” AND (“Peer-to-peer lending” OR “P2P lending” OR “P2P lending decision”)	56
Andor / Web of Science	“Machine learning” AND (“Peer-to-peer lending” OR “P2P lending” OR “P2P lending decision”)	286 / 35
Andor	(data OR “data sources”) AND (“Peer-to-peer lending” OR “P2P lending” OR “P2P lending decision”)	1253
Andor	“Data science” AND (“Peer-to-peer lending platform” OR “P2P lending platform”)	25

Selkeästi tutkimuksen aihepiiriin liittymättömät teokset karsittiin ensin pois tarkemmasta tarkastelusta teoksien otsikoiden perusteella. Tällaisia olivat esimerkiksi teokset, jotka eivät otsikkonsa perusteella viitanneet P2P-lainaamisen ja datatieteen aihepiireihin. Hakutuloksien materiaaleja tarkasteltiin myös sanallisten silmäilykriteerien perusteella. Nämä kriteerit tarkoittivat tutkimuksen suorittamisessa sitä, että otsikkonsa perusteella hyviä teoksia silmäiltiin esimerkiksi johdanto tasolla, joiden perusteella karsittiin pois aiheeseen kuulumattomat aineistot. Tarkastellaan seuraavaksi tutkimuksen toteuttamiseen valikoitunutta tutkimusaineistoa.

2.2 Tutkimusaineisto

Kuten edellisestä kuvasta yksi huomataan, tutkimusaineiston valinnassa hyödynnettiin sekä käytännön rajauskriteerejä että sanallisia silmäilykriteerejä. Tässä osassa tutustutaan ja arvioidaan lyhyesti tutkimuksen keskeistä tutkimusaineistoa. Systemaattisella kirjallisuuskatsauksella saatua aineistoa tarkastellaan luvussa viisi. Alhaalla olevassa taulukossa 3 esitellään tutkimuksen pääkysymyksen ja viimeiseen alakysymykseen vastaamiseen valikoitunut tutkimusaineisto.

Taulukko 3: *Tutkimuksen pääkysymykseen sekä sitä tukevaan viimeiseen alakysymykseen valikoitunut tutkimusaineisto*

Teos	Lyhyt kuvaus
Aleksandrova, Y. (2021). Comparing Performance of Machine Learning Algorithms for Default Risk Prediction in Peer to Peer Lending. TEM Journal. 10 (1), 133–143.	Vertailee eri koneoppimisalgoritmien suoriutumista luottoriskin ennustamisessa P2P-lainaamisessa.
Cao, L., Yang, Q. & Yu, P.S. (2021). Data science and AI in FinTech: an overview. International Journal of Data Science and Analytics. 12 (2), 81–99.	Listaa yleisesti datatieteen ja tekoälyn käyttöä osana finanssiteknologiaa.
Ge, R., Feng, J., Gu, B. & Zhang, P. (2017). Predicting and Deterring Default with Social Media Information in Peer-to-Peer Lending. Journal of Management Information Systems, 34 (2), 401–424.	Tarkastelee sosiaalisen median tietojen merkitystä luottoriskin ennustamisessa P2P-lainaamisessa.
Giudici, P. (2018). Financial data science. Statistics & probability letters. 136, 160–164.	Lyhyt kuvaus datatieteestä finanssialalla. Tutkii, miten korrelaatiomallit voivat parantaa luottoriskiarvioita.
Jagtiani, J. & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. Financial Management. 48 (4), 1009–1029.	Vertailee LendingClub alustan lainoja vastaaviin pankkilainoihin. Tarkastelee vaihtoehtoisen datan merkitystä P2P-lainaamisessa.
Jiang, C., Wang, Z., Wang, R. & Ding, Y. (2017). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. Annals of Operations Research. 266(1-2), 511–529.	Esittelee mallin, joka yhdistää luottoriskin arviointiin lainahakemusten tekstejä.
Lee, J.Y. (2020). Prediction of Default Risk in Peer-to-Peer Lending Using Structured and Unstructured Data. Financial counseling and planning. 31 (1), 115–129.	Luottoriskin ennustamista sekä lainadatalla että lainanhakijan tiedoilla. Tarkastelee lainahakemusten tekstejä.
Mi, J.J, Hu, T. & Deer, L. (2018). User Data Can Tell Defaulters in P2P Lending. Annals of Data Science. 5 (1), 59–67.	Tarkastelee luottoluokituksen merkitystä P2P-lainaamisessa. Esittelee mallin, jota vertaillaan luottoluokituksen onnistumiseen lainanhakijoiden arvioinnissa.
Niu, B., Ren, J. & Li, X. (2019). Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending. Information (Basel). 10 (12), 397–.	Tarkastelee sosiaalisen median verkoston merkitystä lainanhakijoiden luottoriskin ennustamisessa.

Systemaattisen kirjallisuuskatsausmallin eri rajaus- ja silmäilykriteerien avulla onnistuttiin löytämään riittävä tutkimusaineisto tutkimuksen toteuttamiseen. Kaikki tutkimusaineiston tekstit käsittelevät datatieteen ja P2P-lainaamisen aihepiirejä ja ovat vertaisarvioituja tieteellisiä artikkeleita. Suurin osa tutkimukseen valituista teksteistä tarkastelee yksittäisiä esimerkkejä datatieteen käytöstä P2P-lainaamisessa esimerkiksi datan, datalähteiden tai koneoppimisen käsitteiden avulla. Toisaalta esimerkiksi Cao et al. (2021) on luonteeltaan puolestaan laaja ja finanssiteknologian käyttöä kokoava teos.

3. DATATIEDE

3.1 Datatiede käsitteenä

Datatieteelle (Data Science) on olemassa useita eri määritelmiä. Yhden määritelmän mukaan datatiede on tieteenala datalle tai dataa koskevaa tiedettä. Toisen määritelmän perusteella datatiede tarkoittaa alaa, joka yhdistää erilaisia tieteenaloja kuten tilastotiedettä, tietotekniikkaa, viestintää, johtamista sekä sosiologiaa luodakseen datasta hyödyllistä tietoa päätöksentekoa varten. Caon (2017, s. 8) Tiivistetysti ilmaistuna datatiede on Kotun & Deshpanden (2018 luku 1) mukaan kokoelma erilaisia tekniikoita, joiden avulla voidaan luoda arvoa datasta. Toisaalta datatiedettä voidaan tarkastella myös datatuotteiden näkökulmasta. Datatieteen avulla datasta voidaan tehdä tuotteita, jotka voivat olla esimerkiksi löydöksiä, ennusteita, suosituksia tai esimerkiksi päätöksenteon ohjeistuksia (Cao 2017, s. 8). Edellä mainittuja määritelmiä täydentää myös Sarkerin (2021b, s. 3–5) ajatus siitä, että datatiede on ikään kuin kattotermi monelle keinolle, joiden tavoitteena on luoda datasta hyödyllistä tietoa liiketoiminnalliseen käyttöön. Näitä keinoja ovat esimerkiksi edistyneen analytiikan keinot kuten koneoppiminen (Sarker 2021b, s. 3–5).

Cao (2017), Kotu & Deshpande (2018) ja Sarker (2021b) esittämissä määritelmissä yhdistyy ajatus siitä, että datatieteen avulla datasta voidaan tuottaa tietoa erilaisia päätöksentekotilanteita varten. Täten datatiede määritellään kokonaisuudeksi, jonka tavoitteena on luoda datasta tietoa päätöksentekoa varten käyttäen hyväksi erilaisia tekniikoita kuten koneoppimista. Pyritään seuraavaksi konkretisoimaan kyseistä näkökulmaa, jotta datatieteen hyödyntämisestä päätöksenteossa voidaan ymmärtää selkeämmin.

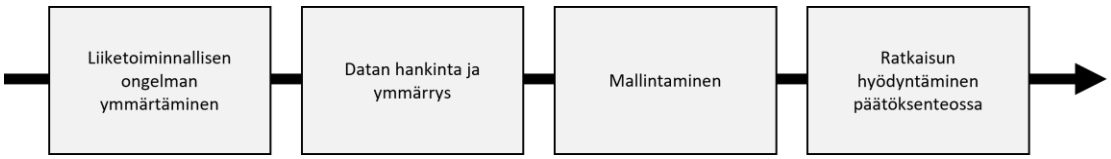
3.2 Datatiede prosessina

Datatieteen kokonaisuutta päätöksenteon mahdollistamisessa voidaan hahmottaa selkeämmin erilaisten prosessimallien avulla. Vaikka datatieteen keinot ja teknologia ovat kehittyneet viimeisten vuosien aikana, eivät Kotun ja Deshpanden (2018, luku 1) mukaan perustavaa laatua olevat datatieteen prosessit ole toisaalta muuttuneet tai ole muuttumassa lähitulevaisuudessa.

Datatiedettä voidaan hahmottaa useiden eri prosessimallien avulla. Yksi yleisimmin käytetyistä datatieteen prosessimalleista on Kotun ja Deshpanden (2018, luku 2) mukaan

Chapmanin et al. (2000) esittelemä CRISP-DM. Malli sisältää Chapman et al. (2000, s. 13) perusteella vaiheet, joita ovat tarkasteltavana olevan liiketoiminallisen ongelman ymmärrys, datan ymmärtäminen ja valmistelu, ratkaisun arviointi, ja ratkaisun hyödyntäminen. Lisäksi esimerkiksi Microsoftin (2022a) esittelemässä TDSP elinkaarimallissa kuvaillaan datatieteen projektien keskeisiksi vaiheiksi samoja kohtia kuten liiketoiminnallinen ymmärrys, datan hankinta -ja ymmärrys, mallintaminen, hyödyntäminen sekä ratkaisun hyväksyntä. Myös Sarker (2021b, s. 5) esittelee teoksessaan datatieteen prosessimallille aiemmin listattuja kohtia kuten datan hankintaa ja hyödyntämistä ratkaistavassa ongelmassa.

Chapman et al. (2000, s. 13), Kotu & Deshpande (2018, luku 2) ja Sarker (2021b, s. 5) perusteella datatieteen prosessina mallintamisessa keskeisiä vaiheita ovat liiketoiminnallisen ongelman ymmärtäminen, datan hankinta ja ymmärtäminen, mallintaminen ja ratkaisun hyödyntäminen. Kuvassa 2 on esitetty datatiede prosessina kyseisen päätelmän perusteella. Esimerkiksi Chapman et al. (2000, s. 13) ja Sarker (2021b, s. 5) esittämien prosessimallien perusteella vaiheet eivät todellisuudessa ole niin suoraviivaisia kuin kuvassa 2, sillä kyseisissä teoksissa esitellään myös esimerkiksi datan valmistelun ja siivoamisen vaiheet. Kyseiset kohdat on jätetty pois, sillä suoraviivaisempi esitys auttaa kuitenkin tarkastelemaan datatiedettä selkeämmin päätöksenteon näkökulmasta. Tarkastellaan seuraavaksi oheisen kuvan 2 eri vaiheita hieman tarkemmin.



Kuva 2: *Datatieteen prosessimalli*

Ensimmäisessä vaiheessa on olennaista ymmärtää ongelma, jota ollaan ratkaisemassa. Se auttaa asettamaan oikeita kysymyksiä, joita datatieteellä pyritään ratkaisemaan. (Kampakis 2020, luku 5) Kyseinen vaihe on hyvin tärkeä, sillä se määrittää pohjan koko liiketoiminnallisen ongelman ratkaisemiseksi. Ilman sen määrittämistä olennaisen datan kerääminen ja täten hyödyllisen tiedon luominen on huomattavan hankalaa. (Sarker 2021b, s. 6) Kampakis (2020, luku 5) ja Sarker (2021b, s. 6) havaintojen perusteella vaihe luo pohjan koko datatieteen käytölle erilaisten ongelmien ratkaisemissa. Tässä vaiheessa voidaan miettiä, millaista dataa ongelman ratkaisemissa voitaisiin käyttää ja mistä sitä saadaan.

Kuten todettiin ensimmäisen vaihe luo pohjan datan hankinnalle. Toisen vaiheen tavoitteena on muodostaa datamassa, joka mahdollistaa ongelmaan vastaamisen. Tässä vai-

heessa on tärkeää pohtia esimerkiksi datan saatavuutta ja sen sopivuutta ongelman ratkaisemiseen. (Kotu & Deshpande 2018, luku 2) Tämän perusteella vaiheessa on siis syytä kiinnittää huomiota oleellisiin datalähteisiin ja täten relevantin datan saantiin. Datalle on yleensä hyvin tyypillistä, että se ei ole laadultaan hyvää ja se sisältää erilaisia virheitä, jotka huonontavat esimerkiksi ennustavien mallien suorituskykyä (Sarker, 2019, s. 3–4). Datatieteen tuottamat ratkaisut ovat Kotun ja Deshpanden (2018, luku 2) mukaan juuri yhtä hyviä, kun niissä hyödynnetty data. Kotu & Deshpande (2018, luku 2 ja Sarker (2019, s. 3–4) päätelmien mukaan voidaan todeta, että itse datalla ja sen laadulla on hyvin keskeinen rooli datatieteen hyödyntämisessä.

Mallintamisen vaiheessa on oleellista valita saadusta datasta keskeiset muuttujat, joita hyödynnetään koneoppimismalleissa (Microsoft 2022b). Koneoppiminen on itsessään hyvin tärkeä osa datatiedettä liiketoiminnallisten ongelmien ja päätöksenteon kannalta, sillä Sarkerin (2021a s. 7) mukaan juuri koneoppimismallit mahdollistavat päätöksenteossa käytettävien ennusteiden ja suositusten tekemistä. Mallintamisen vaiheessa Sarker (2021a s. 7) ja (Microsoft 2022b) perusteella on siis olennaista keskittyä hyödyntämään koneoppimismalleja päätöksentekoa edistävien löydösten ja havaintojen tekoon.

Viimeisessä vaiheessa muodostettua datatieteen ratkaisua käytetään päätöksenteossa. Kuten Caon (2017, s. 8) mukaan aiemmin todettiin, datatieteen avulla voidaan tuottaa datasta tuotteita, jotka voivat olla esimerkiksi löydöksiä, ennusteita, suosituksia tai esimerkiksi päätöksenteon ohjeistuksia. Lisäksi loppuratkaisu yhdistetään usein reaaliaikaisiin päätöksentekoprosesseihin, kuten verkkosivujen palveluihin. Kyseisten ratkaisujen hyödyntäminen vaatii yleensä jatkuvaa huoltoa ja kehittämistä. (Chapman et al. 2000, s. 14) Jos dataa kerätään ratkaisun hyödyntämisen aikana lisää, voisi olla oleellista tarkastella kokonaisuuden toimivuutta uudelleen.

Kuten Kampakisin (2020) ja Sarkerin (2021b) perusteella huomattiin, datatieteen ratkaisuissa olennaista on syytä keskittyä ratkaistavaan ongelmaan. Lisäksi Kotu & Deshpande (2018) ja Sarker (2019) perusteella huomattiin, että datalla ja datalähteillä on suuri merkitys datatieteessä. Lisäksi Sarkerin (2021a) ja Microsoftin (2022b) mukaan koneoppimismallit ovat keskeisessä roolissa erilaisten löydösten ja täten päätösten tekemisessä. Kaikkien näiden havaintojen perusteella voidaan pohtia, voitaisiinko datatieteen tuottamia ratkaisuja hyödyntää esimerkiksi osana P2P-alustojen toimintaa.

3.3 Data ja datalähteet

Kuten Sarkerin (2021a, s. 6) mukaan aikaisemmin todettiin, on liiketoiminnallisen ongelman määritteleminen tärkeää datan ja sen keräämisen kannalta. Datatieteen prosessin

toisessa vaiheessa pyritään hankkimaan data ja tutustumaan sen sisältöön (Microsoft 2022). Dataan liittyvät ominaisuudet kuten sen rakenne määrittelevät datan analysointiin tarvittavat keinot ja kertovat ylipäätään ongelmasta, mitä tarkastellaan (Ozdemir 2016, luku 2). Tarkastellaan seuraavaksi, millaista dataa datatieteessä voidaan hyödyntää ja mistä tätä dataa voidaan saada.

Yksi tapa tarkastella dataa on keskittyä sen rakenteeseen. Data voi olla rakenteellista tai rakenteetonta. Rakenteellinen data on selkeästi havainnoitavaa ja sen käsittely ja analysointi on yleensä huomattavan helppoa. (Ozdemir 2016, luku 2) Rakenteellinen data on Sarkerin (2021b, s. 3) mukaan yleensä myös taulukkomaisessa muodossa. Datan selkeä ja tarkasti määritelty rakenne sopii erityisesti koneoppimisen algoritmien käyttöön. Esimerkkejä rakenteellisesta datasta ovat päivämäärät, henkilöiden nimet, osoitteet, luottokorttien numerot, paikkatiedot, osakkeiden hinnat ja muu rahassa mitattavissa oleva data (Hurwitz et al. 2013, s. 73; IBM Cloud Education, 2021; Sarker 2021b, s. 4). Edellä mainittujen tekstien esittelemien esimerkkien perusteella näyttää siltä, että rakenteellinen data kuvailee esimerkiksi henkilöjen perustietoja, jotka voidaan ilmoittaa selkeästi kuten syntymäaika tai osoite. Lisäksi tähän voisi kuulua yksilöiden varallisuuteen liittyvät tiedot, jotka ovat mitattavissa rahassa ja siten myös helposti taulukkomaisessa ja selkeässä muodossa esitettävissä. Täten rakenteellista dataa voitaisiin hyödyntää osana P2P-lainojen lainanhakijoiden arviointia.

Toisaalta voidaan tarkastella, onko data sellaista, että sitä voidaan laskea numeroilla vai ei. Rakenteellinen data voidaan IBM Cloud Education:n (2021) mukaan yhdistää myös kvantitatiiviseen dataan eli dataan, joka on Ozdemirin (2016, luku 2) mukaan helposti kuvattavissa numeroilla ja sitä voidaan täten käsitellä matemaattisten mallien avulla. Aikaisemmin mainittujen tekstien IBM Cloud Education (2021) ja Sarker (2021b, s. 4) perusteella kaikki rakenteellinen data kuten nimet ja osoitteet, eivät selkeästi ole kuitenkaan kvantitatiivisia. Kuitenkin esimerkiksi Hurwitzin et al. (2013, s. 73) kuvailema taloudellinen data, kuten osakkeiden hinnat, on puolestaan helposti mitattavissa. Tämän perusteella voidaan todeta, että esimerkiksi henkilöitä koskeva varallisuus ja tulot ovat yhtä lailla helposti tarkasteltavissa olevaa dataa.

Rakenteeton data ei puolestaan noudata tiettyä rakennetta, joten sitä on hankalampi ymmärtää kuin rakenteellista dataa. Suurin osa maailmassa luodusta datasta on rakenteetonta. (Ozdemir 2016, luku 2) Epämääräisen rakenteensa vuoksi rakenteeton data vaatii IBM Cloud Educationin (2021) mukaan huomattavasti enemmän erilaisia työkaluja ja analysointia, jotta sitä voidaan käyttää hyväksi. Rakenteetonta dataa ovat esimerkiksi sosiaalisen median julkaisut, muiden nettisivujen sisältö, sähköpostit ja mobiililaitteiden tuottama data kuten tekstiviestit ja puhelut (Hurwitz et al. 2013, s. 79; Ozdemir 2016,

luku 2; IBM Cloud Education, 2021). Näiden havaintojen perusteella näyttää siltä, että rakenteeton data on yleensä ihmisen kirjoittamaa tekstimuodossa olevaa dataa. Tekstidataa voidaan analysoida erilaisilla luonnollisen kielen analysointiin tarkoitetuilla koneoppimismalleilla. Esimerkiksi erilaisista sosiaalisen median julkaisuista voidaan tuoda esiin julkaisujen tunnetiloja. (IBM Cloud Educationin 2020) Täten voisi olla mahdollista käyttää myös lainahakemusten tekstejä datana, joista voitaisiin esimerkiksi IBM Cloud Education (2020) mainituin keinoin tuoda esiin erilaisia lainapäätökselle oleellisia ominaisuuksia.

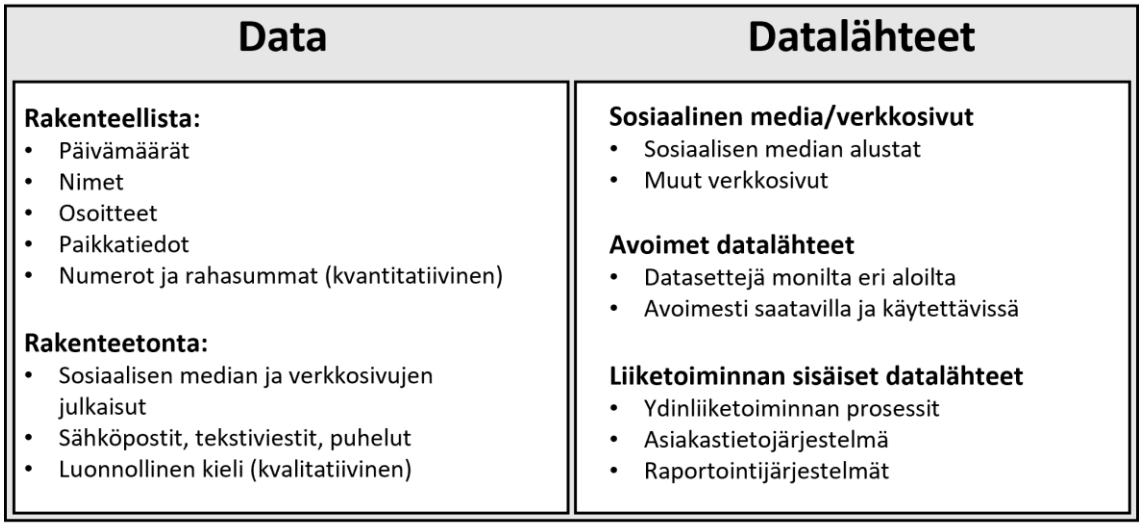
Rakenteetonta dataa kuvaillaan usein myös kvalitatiivisena datana (IBM Cloud Education, 2021). Kvalitatiivinen data on yleensä tekstimuotoista luonnollista kieltä, joten sitä on hankalampi analysoida matemaattisesti (Ozdemir 2016, luku 2). Edellisessä kappaleessa huomattiin, että rakenteeton data on sekä Ozdemir (2016, luku 2) että IBM Cloud Education (2021) perusteella usein tekstidataa. Täten voidaan siis todeta, että rakenteeton data on yleensä myös kvalitatiivista dataa. Tarkastellaan seuraavaksi tarkemmin datalähteitä.

Relevanttien datalähteiden määrittely on datatieteessä keskeistä tarkasteltavan ongelman ratkaisemiseksi (Sarker 2021b, s. 5). Kuten esimerkiksi IBM Cloud Education (2021) ja Ozdemir (2016, luku 2) perusteella huomattiin, eri muodoissa olevaa dataa löytyy yleensä digitaalisessa muodossa esimerkiksi sosiaalisen median julkaisuina. Yksilöiden digitaalisilla alustoilla tuottamaa dataa voidaan hyödyntää ihmisten käyttäytymisen ennustamiseen ja täten määritellä esimerkiksi luottoluokitusta, vakuutusriskiä tai muita henkilökohtaisia asioita. Lisäksi alun perin eri käyttöön kerättyä dataa voidaan hyödyntää täysin toisiin tarkoituksiin. (Grossi et al. 2021, s. 268) Grossi et al. (2021) perusteella voidaan siis todeta, että sosiaalinen media ja internet ovat keskeisiä datalähteitä datatieteessä, kun tutkitaan esimerkiksi ihmisjoukkojen käyttäytymistä. Ottamatta huomioon datankäytön oikeudellisia ja eettisiä näkökulmia, voisivat sosiaalinen media ja internet yleensä olla hyviä datalähteitä myös P2P-lainapäätösten kontekstissa.

Datatieteessä on mahdollista hyödyntää avoimia datalähteitä, jotka tarjoavat vapaasti käytössä olevaa dataa eri aloilta. Kyseisiä datalähteitä on kehitetty etenkin tieteen ja tutkimuksen tekemiseen. (Cao 2017, s 17) Avoimella datalla viitataan yleensä datalähteisiin, joka tarjoaa kaikkien vapaasti käytettävissä ja täten analysoitavissa olevaa dataa (Patel 2019). Useita eri avoimia datalähteitä on avattu esimerkiksi tieteellisten toimijoiden, valtioiden ja kaupallisten tahojen puolesta. Lisäksi kyseisiä datalähteitä voidaan hyödyntää tutkimuksen lisäksi myös kaupallisissa sovelluksissa. (Lahti 2018, s. 33) Avoimista datalähteistä saa yleensä valmiiksi rakenteellisessa muodossa olevaa dataa (Patel

2019). Kuten Ozdemir (2016, luku 2) totesi, on rakenteellisen datan käsittely ja analysointi vaivatonta. Tämän perusteella voidaan todeta, että avoimet datalähteet voivat olla hyviä datalähteitä datatieteen sovelluksissa, sillä ne tarjoavat datan jo valmiiksi hyvässä muodossa. Avointen datalähteiden datan avulla voitaisiin myös ymmärtää tarkastelun kohteena olevia ilmiöitä, jos ilmiöiden ja datalähteen data sisältävät samankaltaisia muuttujia.

Esimerkiksi Lahti (2018, s. 33) ja Ozdemir (2016, luku 2) perusteella huomataan, että iso osa datalähteistä on verkossa olevia julkaisuja tai avoimiin datalähteisiin varastoituja datasettejä. Mainitut datalähteet ovat yritysten liiketoiminnan ulkopuolella (Hayes, 2014). Vaikka käsittelimme aikaisemmin ulkoisia datalähteitä, on Hayesin (2014) mukaan myös sisäisiä datalähteitä, joita ovat esimerkiksi yritysten asiakastieto- ja raportointijärjestelmät. Esimerkiksi lainahakemuksia vastaanottava P2P-laina-alusta voisi pitää sisäisinä datalähteinään juuri aikaisempia lainahakemuksia. Lisäksi Hayesin (2014) mukaan sisäiset datalähteet ovat hyvin liiketoimintaa harjoittavan tahon hallussa, joten niissä kerättävä data voidaan määritellä juuri halutun mukaiseksi. Oheisessa kuvassa 2 on esitetty edellä mainittuja havaintoja datasta ja datalähteistä.



Kuva 3: *Tiivistelmä datatieteessä hyödynnettävästä datasta ja eri datalähteistä*

Kuten kuvasta 3 nähdään, data voidaan jaotella rakenteelliseen ja rakenteettomaan dataan. Rakenteellinen data on kuvan perusteella usein laskettavissa olevaa numeerista dataa, kun taas rakenteeton on tekstisisältöä. Lisäksi kuvasta on huomattavissa, että dataa voidaan saada useista erilaisista datalähteistä, joita ovat sosiaalinen media ja verkkosivut, avoimet datalähteet ja liiketoiminnan sisäiset datalähteet.

3.4 Koneoppiminen

Kuten Sarker (2021b, s. 6) mainitsee, mallintamisen vaiheessa osana datatieteen prosessia rakennetaan malli, jossa käytetään hyväksi koneoppimista. Koneoppimisen avulla voidaan löytää datasta erilaisia ominaisuuksia tai havaintoja. Kyseisiä löydöksiä pystytään löytämään myös valtavista datamassoista. (Kashyap 2017, luku. 1) Koneoppiminen on tekoälyn osa-alue, joka tarkastelee koneiden kykyä rakentaa analyyttisiä malleja, jotka oppivat ja osaavat kehittää itseään ilman, että niitä erikseen ohjelmoidaan uudelleen. Oppiminen tapahtuu toistuvien ja toiminnallisten takaisinkytkentäsilmukoiden avulla, joita suorittaessa koneoppismalli kehittyy. (Boobier 2018, s. 46; Sas 2022)

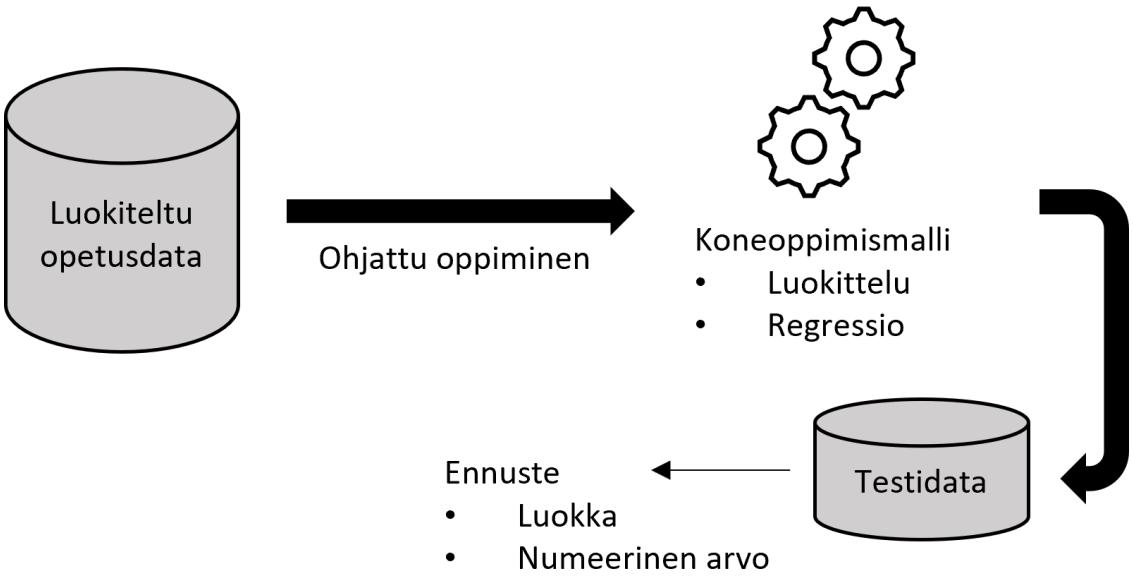
Malleihin valitut koneoppimismallit ovat riippuvaisia siitä, millaista ongelmaa ollaan ratkaisemassa. Ongelmaa voidaan tarkastella esimerkiksi siten, onko ongelmassa kyse esimerkiksi jonkin asian ennustamisesta vai ei. (Sarker 2021b, s. 6) Tarkastellaan seuraavaksi koneoppimisen yleisiä osa-alueita ja niiden käyttökohteita. Täten voidaan ymmärtää, miten koneoppimista voidaan hyödyntää datatieteessä osana päätöksentekoa.

Koneoppiminen voidaan perinteisesti jakaa kahteen eri osa-alueeseen, joita ovat ohjattuja ohjaamaton oppiminen. Lisäksi kolmanneksi osa-alueeksi määritellään toisaalta myös vahvistettu oppiminen. (Boobier 2018, s. 46; Chandramouli et al. 2018, luku 1) Ohjatussa oppimisessa käytetään Boobierin (2018, s. 46) mukaan opetusdataa, jonka perusteella malli oppii datasta tiettyjä sääntöjä. Kun koneoppimismalli on opetettu, voidaan sitä käyttää opetusdataa vastaavan datan analysointiin ja täten ennusteiden tekemiseen. Ohjatun oppimiset keinot voidaan yleensä jakaa kahteen osaan, joita ovat luokittelu- ja regressiomallit. (Chandramouli et al. 2018, luku 1) Luokittelun ja regression merkittävin ero perustuu siihen, että klassifioinnissa ennustetaan tiettyä datan luokkaa, kun taas regressio mahdollistaa jatkuvan muuttujan ennustamista (Sarker 2021a, s. 8).

Luokittelumallien avulla voidaan Chandramoulin et al. (2018, luku 1) mukaan ennustaa datasta ominaisuuksia eri luokkiin. Esimerkiksi pankkimaailmassa luokittelua voidaan käyttää erottelemaan petoksia rahan siirrossa. Lisäksi Delua (2021) toteaa, että sitä voidaan käyttää esimerkiksi roskapostin suodattamiseen. Puolestaan regressiomalleilla voidaan Chandramoulin et al. (2018, luku 1) mukaan ennustaa numeerisia ominaisuuksia. Kuten myös Sarker (2021a, s. 8) aiemmin perusteli, ennusteet eivät siis kerro yhtä luokkaa vaan jatkuvan muuttujan arvon. Regressiota voidaan yleisesti käyttää esimerkiksi myynnin ennustamisessa, trendien analysoinnissa ja aikasarja-analyysissa (Delua 2021, Sarker 2021a, s. 8).

Chandramouli et al. (2018, luku 1), Delua (2021) ja Sarker (2021a, s. 8) perusteella voidaan todeta, että ohjatulla oppimisella on erilaisia laajoja käyttökohteita. Täten voidaan

pohtia, olisiko ohjatulla oppimisella käyttökohteita myös lainapäätösten yhteydessä. Mahdollisuuksia voisi olla esimerkiksi lainan maksukyvyn ennustamisessa. Oheisessa kuvassa 4 on tiivistettynä ohjatun oppimisen periaate.



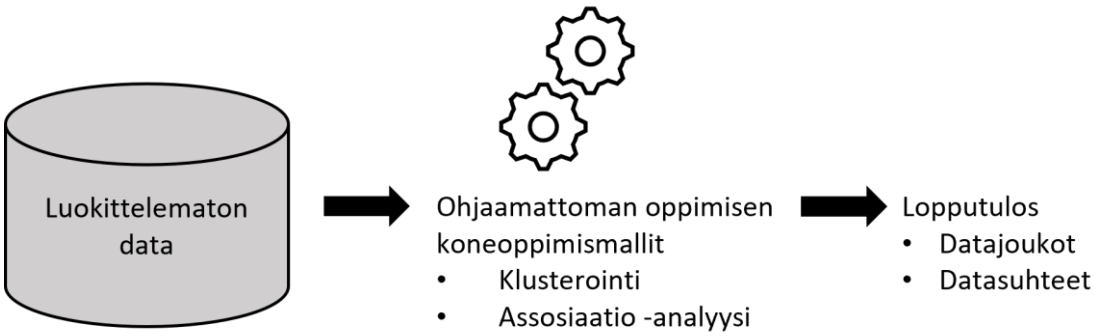
Kuva 4: *Ohjattu oppiminen (mukaillen lähdettä Chandramouli et al. (2018, luku 1)).*

Kuvassa 4 koneoppimismalli opetetaan ensin luokitellulla opetusdatalla, minkä jälkeen ennustavaa mallia voidaan hyödyntää testidatan arviointiin. Kun mallia hyödynnetään testidatan tarkasteluun, saadaan muodostettua ennusteita. Ennusteet ovat luokittelumalleille datan luokkia, ja regressiomalleille usein numeerisia arvoja.

Ohjaamattomassa oppimisessa Chandramoulin et al. (2018, luvun 1) mukaan koneoppimismalli etsii itsenäisesti rakenteettomasta datasta toistuvia ominaisuuksia. Sekä Chandramoulin et al. (2018, luvun 1) että Delua (2021) mukaan ohjaamattomassa oppimisessa on selkeä ero ohjattuun oppimiseen, sillä siinä ei käytetä valmiiksi määriteltyjä luokkia datalle, vaan mallit pyrkivät löytämään annetusta datasta piilossa olevia ominaisuuksia, mahdollisia pistejoukkoja ja toistuvia kaavoja. Ohjaamaton oppiminen voidaan perinteisesti jakaa klusterointiin ja assosisaatio-analyysiin (Chandramouli et al. 2018, luku 1).

Klusteroinnissa tavoitteena on jaotella eri datapisteitä toisistaan ja järjestää toisiaan muistuttavia datapisteitä samaan pistejoukkoon (Chandramouli et al. 2018, luku 1). Klusterointia voidaan yleisesti käyttää esimerkiksi asiakkaiden jakamiseen erilaisiin eri ryhmiin niiden käyttäytymisen perusteella (Kashyap 2017, luku 1; Sarker 2021a, s. 9). Klusterointia voidaan tämän lisäksi hyödyntää Deluan (2021) mukaan esimerkiksi anomalioiden eli selittämättömien epäsäännöllisyyksien tunnistamisessa. Assosiointi -analyysi tar-

kastelee puolestaan eri datapisteiden suhdetta toisiinsa, josta yksi esimerkki on ostoskorianalyysi. (Chandramouli et al. 2018, luku 1). Deluan (2021) mukaan ostajille voidaan tarjota ostoskorissa olevien tuotteiden lisäksi tuotteita, joita muut saman tuotteen ostaneet asiakkaat ovat ostaneet. Vastaaville assosisaatio -analyysin mahdollistaville suosittelujärjestelmille, kyseisten lähteiden perusteella olla myös hyötyä muissa päätöksentekoa edistävissä käyttökohteissa. Alhaalla olevassa kuvassa 5 tiivistettynä ohjaamaton oppiminen.



Kuva 5: *Ohjaamaton oppiminen (mukaillen lähdettä Chandramouli et al. (2018, luku 1)).*

Kuvassa 5 luokittelematonta dataa syötetään ohjaamattoman oppimisen koneoppimismallille. Kuvasta nähdään, että ohjaamattomassa oppimisessa ei vaadita opetusdataa kuten ohjatussa oppimisessa puolestaan vaadittiin. Klusterointimallien avulla data voidaan jakaa eri joukkoihin, ja assosisaatio -analyysin avulla voidaan selvittää datasta erilaisia suhteita.

Kuten aiemmin oppiminen Boobier (2018, s. 46) ja Chandramouli et al. (2018, luku 1) perusteella todettiin, koneoppimiselle määritellään myös kolmas osa-alue, joka on vahvistettu oppiminen. Vahvistettu oppiminen perustuu palkitsemisjärjestelmään, jossa koneoppimismalli oppii erilaisten positiivisten- tai negatiivisten pisteiden avulla. Vahvistettua oppimista käytetään yleensä roboteissa ja esimerkiksi itse ajavissa autoissa. (Chandramouli et al. 2018, luku 1).

4. P2P-LAINAAMINEN JA -ALUSTAT

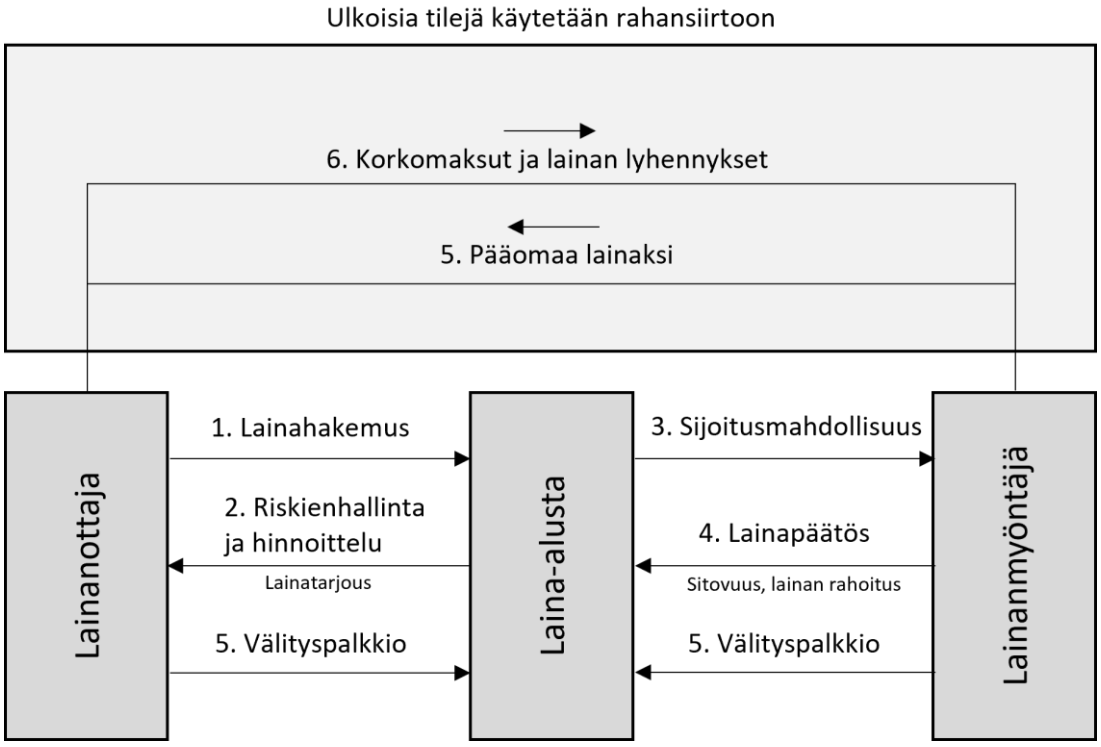
4.1 P2P-lainaamisen määritelmä

Vuoden 2008 finanssikriisi loi epäluotettavan kuvan pankkisektoria kohtaan. Ihmisten taloudellinen tilanne heikkeni ja pankkisektoria kohtaan asetettiin uusia vaatimuksia ja entistä kovempaa säätelyä. Täten tavallisten pankkien ohelle syntyi uusia toimijoita kuten vertaislainapalveluita tuottavia alustoja. (Zwilling et al. 2020, s. 854) Vertaislainaamisella (Peer-to-Peer lending, P2P) tarkoitetaan suoraa lainaamista, joka tapahtuu eri yksilöiden tai yritysten välillä ilman virallisen finanssi-instituution osallistumista itse lainaamisen prosessiin. P2P-lainaamisessa lainaaminen tapahtuu verkkopohjaisilla alustoilla, joiden tehtävänä on saattaa lainanhakijat ja lainanmyöntäjät yhteen. (CFI 2022) P2P-alustat eivät itse lainaa rahaa ja niillä ei ole samanlaisia pääomavaatimuksia. Lisäksi alustat eivät ota myöskään luottoriskiä, vaan riskit ovat alustan käyttäjillä. (Ferretti 2021, s. 121)

Vertaislainaaminen tarjoaa erilaisia mahdollisuuksia sekä lainaajille että lainaa hakeville. Yleensä lainanmyöntäjä eli sijoittaja saa rahalleen hyvää tuottoa. Lainanhakijalle palvelussa on kyse puolestaan vaihtoehtoisen rahoituksen saamisesta, sillä P2P-lainat ovat mahdollinen lainaamisen myös niille, joilla on esimerkiksi huono luottoluokitus tai luottopisteytys (CFI 2022). Rahan lainaaminen perinteisistä instituutioista kuten pankeista voi olla siis jokseenkin hankalaa, jos lainan hakijan taloudelliset edellytykset eivät ole täysin kunnossa.

4.2 P2P-lainaamisen ja -alustojen toiminta

P2P-lainaamisen toiminnallisuutta esitellään Lenz (2016, s. 691) mukaisesti oheisessa kuvassa 6. Kuva mahdollistaa selkeän ymmärryksen esimerkiksi siitä, millainen lainaalustojen rooli on lainaamisen prosessissa ja miten lainaamisen eri osapuolet erottuvat toisistaan. Tarkastellaan seuraavaksi kuvan eri vaiheita.



Kuva 6: *P2P-lainaamisen kokonaisuus prosessina (mukaillen lähdettä Lenz (2016, s.691)).*

Kuvan ensimmäisessä kohdassa Lenzin (2016, s. 691) mukaan lainanhakija lähettää lainahakemuksen laina-alustalle, jossa on Schneiderin (2022) mukaan syytä kertoa lainan tarve, hakijan taloudellinen historia sekä jotain muuta lainan hakemisen kannalta hyödyllistä tietoa. CFI:n (2022) mukaan suurin osa P2P-lainoista ovat vakuudettomia kulutusluottoja. Vaikka lainat ovat tyyliltään hyvin samanlaisia kuten normaalit kulutusluotot, on asiakkaan mahdollista luovuttaa tietoa, joka ei suoraan liity heidän taloudelliseen suorituskykyynsä. Tätä tietoa voisi olla esimerkiksi lainanhakijan sosiaalisen mediaan jättämä digitaalinen informaatio. Tähän liittyy kuitenkin avoimia kysymyksiä esimerkiksi tiedon käytön sääntelystä ja eettisyydestä. (Lenz 2016, s. 693)

Toisessa vaiheessa alusta arvioi lainahakemuksen ja määrittää asiakasta koskevan luottoriskin. Jos luottoriski on alustan standardeihin sopiva, voidaan lainanhakijalle esittää lainatarjous riskin mukaisella korkotasolla. (Lenz 2016, s. 691; CFI 2022) Luottoriski tarkoittaa mahdollisuutta, jolla lainanottaja ei kykene maksamaan ottamaansa lainaa tai maksamaan lainaan liittyviä kuluja lainan myöntäjälle (Investopedia 2021). Lenzin (2016, s. 693) mukaan P2P-alustoilla riskiarviot tehdään yleensä automatisoiduilla tietokoneohjelmilla ilman ihmisen osallistumisen tarvetta. Täten lainanhakija joko hyväksyy tai hylkää saamansa lainatarjouksen.

Kolmannessa vaiheessa, kun lainanhakija on hyväksynyt lainan ehdot, annetaan tieto lainasta alustalla lainanmyöntäjille, jolloin lainoihin on mahdollista sijoittaa. Esimerkiksi

Zwilling et al. (2020, s. 854) mukaan, joillain P2P-alustoilla lainan myöntäjien on mahdollistaa vaikuttaa, millaisia ominaisuuksia he vaativat lainan hakijoilta, kun taas joissain palveluissa päätökset tehdään palvelun puolesta automaattisesti. Aiemmin esiteltyjen Lenzin (2016, s. 691) ja CFI:n (2022) havaintojen perusteella alustan tekemät luottoriskiarviot ovat keskeinen informaation lähde lainanmyöntäjille, sillä sijoitettavat lainat jaetaan lainanmyöntäjille niiden perusteella. Neljännessä vaiheessa lainanmyöntäjät sitoutuvat rahoittamaan lainan eli tekevät lainapäätöksen. Yhden lainoittajan ei tarvitse myöntää yksittäistä lainaa yksin, vaan laina voidaan hajauttaa usealle sijoittajalle. Lisäksi lainanmyöntäjät sitoutuvat laina-alustan kanssa tiettyihin ehtoihin ja esimerkiksi heidän tietonsa tutkitaan esimerkiksi rahanpesun varalta. (Lenz, 2016, s. 692)

Viidennessä vaiheessa sekä lainanhakijat että -myöntäjät maksavat laina-alustalle palkkion sen tuomasta palvelusta molemmille osapuolille, joka maksetaan ennen varsinaisen lainapääoman siirtoa (Lenz 2016, s. 692). Kyseisessä vaiheessa alusta saa siis palkkion siitä, että se yhdistää eri osapuolet toisiinsa ja toimii koko prosessin ohjaajana. Tämän jälkeen alusta kerää lainanmyöntäjiltä kerätyn lainasumman ja siirtää sen lainanhakijan tilille (Lenz 2016, s. 692).

Kuvan 6 viimeisessä eli kuudennessa vaiheessa lainanhakija maksaa lainanmyöntäjälle korkoa saamastaan lainasta. Itse alustalle ei ole vastuuta taata sijoittajille, jos lainan ottaja ei kykene maksamaan korkojaan (Zwilling et al. 2020, s. 857). Tässä on siis nähtävissä selkeä ero esimerkiksi pankkeihin, joissa sääntely on erilaista P2P-alustoihin verrattuna. Lisäksi tässä vaiheessa lainanhakija maksaa lainasta lyhennyksiä lainanmyöntäjälle kunnes laina on kokonaan maksettu (Lenz 2016, s. 692).

4.3 P2P-lainaamisen keskeiset ongelmat

Vertaislainaamiseen liittyy tiettyjä ongelmia, jotka liittyvät yleisesti alan riskeihin ja sääntelyyn. P2P-lainat altistuvat usein korkealle luottoriskille. Usein kyseisiä lainoja hakevilla henkilöillä on huono luottopisteytys tai -luokitus, ja he eivät saa otettua lainoja perinteisistä pankeista. (CFI 2022) Koska P2P-lainat muistuttavat usein tavallisia kulutusluottoja eikä niissä yleensä tarvitse selventää mihin lainarahat käytetään, voivat lainan ottajat käyttää rahojaan vastuuttomasti. Täten on mahdollista, että suurella osalla laina-alustojen lainanhakijoista on korkea luottoriski. Ilmiötä kutsutaan yleensä haitalliseksi valikoitumiseksi. (Lenz 2016, s. 697)

Lenzin (2016, s. 693) mukaan P2P-alustat kohtaavat epäsymmetrisen informaation ongelmia. Epäsymmetrisen informaation ongelmalla tarkoitetaan tilannetta, jossa taloudel-

listen osapuolien välinen tieto ei ole samalla tasolla eli toinen osapuoli tietää käsiteltävästä asiasta enemmän kuin toinen. Kyseinen ongelma on hyvin yleistä taloudellisten transaktioiden kontekstissa. (Bloomenthal 2021) Esimerkki epäsymmetrisen informaation ongelmasta on Lenzin (2016, s. 693) mukaan laina-alustojen kyvyttömyys muodostaa samanlaisia pitkäaikaisia asiakassuhteita kuin tavalliset pankit, mikä voi johtaa lainanhakijan todellisen taloudellisen tilanteen virhearviointiin. Lenzin (2016, s. 693) havaintojen ja Bloomenthal (2021) määritelmän perusteella alustojen tulisi siis kyetä ymmärtämään paremmin lainanhakijoiden todellisia riskejä, jotta lainanmyöntäjät saavat palvelusta hyötyä. Toisaalta myöskään lainanmyöntäjät eivät välty epäsymmetrisen informaation ongelmalta, koska alustojen lainanhakijoiden luottoriskien arviointiin liittyvät kriteerit voivat olla epäselviä (Ferretti 2021, s. 121). Tähän asiaan vaikuttaa kuitenkin kirjottajan mukaan alan sääntely ja eri maiden lainaamista koskevat lait.

Kuten tavalliset pankit, P2P-alustat eivät itsessään lainaa rahaa, joten samat sääntelyt eivät päde alustoille esimerkiksi pääomavaatimusten kannalta. Itse alustat eivät myöskään ota luottoriskiä, vaan riskit ovat alustojen käyttäjillä. (Ferretti 2021, s. 121) Kirjottajan mukaan sääntelyn puute P2P-alustoilla voi johtaa ongelmiin sekä lainanhakijoiden että -myöntäjien välillä. Sääntelyn puute voi siis aiheuttaa ongelmia esimerkiksi kiistatilanteiden ratkaisemisessa, joissa lainaa ei ole maksettu takaisin.

5. DATATIEDE JA P2P-LAINAPÄÄTÖKSET

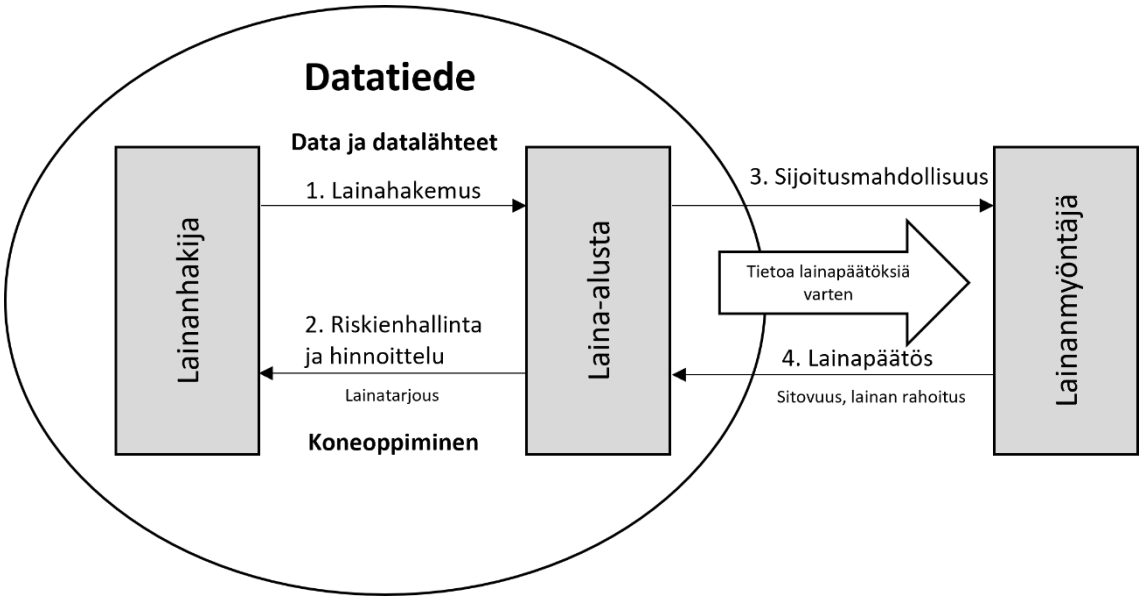
5.1 Datatieteen käyttö P2P-lainaamisessa

Kuten luvussa kolme Cao (2017), Kotu & Deshpande (2018) ja Sarker (2021b) perusteella todettiin, datatieteen avulla datasta voidaan tuottaa tietoa päätöksentekoa varten käyttäen hyväksi erilaisia tekniikoita kuten koneoppimista. Puolestaan luvussa neljä Lenz (2016, s. 693) perusteella huomattiin, että P2P-lainaamisessa esiintyy merkittävä epäsymmetrisen informaation ongelma, joka koskee etenkin lainanhakijan ja laina-alustan välistä suhdetta. Lisäksi Lenzin (2016, s. 693) tekstistä huomattiin, että epäsymmetrisen informaation ongelma liittyy selkeästi lainanhakijoiden todellisen luottoriskin ymmärtämiseen.

Tekstissään Cao et al. (2021, s. 85) kokoavat erilaisia finanssiteknologian käyttökohteita ja listaavat P2P-lainaamisen olevan yksi esimerkki datatieteen ja tekoälyn käyttökohteista. Datatiedettä ja tekoälyä voidaan hyödyntää vertaislainoissa esimerkiksi lainatarjouksien optimointiin sekä epäsymmetrisen informaation ongelmien ennustamiseen ja hallitsemiseen (Cao et al. 2021, s. 85). Lisäksi Giudici (2018, s. 161) mukaan datatiede on keskeisessä roolissa osana P2P-alustojen toimintaa ja valtavaa kasvua, sillä datatieteen mahdollistamat keinot ovat muuttaneet sitä, miten dataa kerätään, prosessoidaan ja arvioidaan. Lisäksi tämä on johtanut siihen, että luottoarvioiden kannalta hyödyllistä tietoa saadaan tuotettua edullisemmin (Giudici 2018, s. 161). Giudici (2018, s. 161) näkemyksiä tukee Jagtianianin ja Lemieuxin (2019, s. 1027) havainnot, joiden mukaan P2P-alustat käyttävät vaihtoehtoisia datalähteitä ja erilaisia algoritmeja, jotka mahdollistavat nopeiden ja edullisten luottoarvioiden tekemisen.

Kuten aikaisemmin luvussa neljä Lenz (2016, s. 691) perusteella huomattiin, laina-alustojen tuottamat luottoriskien arviot ovat keskeinen informaation lähde lainanmyöntäjien lainapäätöksiin. Lisäksi Giudici (2018) , Cao et al. (2021) ja Jagtianiani & Lemieux (2019) havaintojen perusteella huomataan, että datatiedettä voidaan hyödyntää osana lainanhakijan ja laina-alustan välisen epäsymmetrisen informaation ongelman ratkaisua. Täten voidaan todeta, että P2P-alustojen datatieteen mahdollistamien luottoriskin arvioinnin keinojen avulla voidaan tuottaa tietoa lainanhakijoista lainanmyöntäjien lainapäätöksentekoa varten.

Aikaisemmin luvussa kolme esitetyssä kuvassa 2 havainnollistettiin datatiedettä prosessina, joka koostui neljästä eri vaiheesta: liiketoiminnallisen ongelman ymmärtäminen, datan hankinta ja ymmärrys, mallintaminen ja ratkaisun hyödyntäminen päätöksenteossa. Aikaisempien Giudici (2018) , Cao et al. (2021) ja Jagtianiani & Lemieux (2019) havaintojen perusteella P2P-lainaamisen kontekstissa liiketoiminnallisena ongelmana on selvittää mitkä lainanhakijat ovat hyviä ja mitkä eivät. Tätä voidaan arvioida tarkastelemalla lainanhakijoiden luottoriskiä, jonka arviointiin tulee olla relevanttia dataa. Tätä dataa voitaisiin analysoida erilaisten koneoppimismallien avulla. Lisäksi kokonaisuuden tuottamaa ratkaisua voitaisiin täten hyödyntää tuottamaan tietoa lainanhakijoista lainanmyöntäjille lainapäätöksiä varten. Oheisessa kuvassa 7 pyritään kuvailemaan datatieteen käyttöä osana kuvassa 6 esiteltyä P2P-lainaamisen kokonaisuutta.



Kuva 7: *Datatieteen hyödyntäminen osana P2P-lainaamisen kokonaisuutta.*

Kyseisessä kuvassa oleva kehä kuvailee P2P-lainaamisen prosessin vaiheita, joissa datatiedettä voidaan hyödyntää. Kuvasta nähdään, että datatiedettä voidaan hyödyntää lainanhakijoiden ja laina-alustan välisissä kohdissa. Kuvassa oleva nuoli puolestaan kuvaa sitä, että datatiede tarjoaa tietoa lainanmyöntäjien lainapäätöksentekoa varten. Tarkastellaan seuraavaksi datatieteen kehässä esitettyjen datan, datalähteiden ja koneoppimisen merkitystä P2P -lainapäätöksissä.

5.2 Data ja datalähteet P2P-lainapäätöksissä

Kuten aikaisemmin luvussa kolme todettiin, liiketoiminnallisen ongelman määrittely on Sarkerin (2021b, s. 6) mukaan datatieteen vaiheista hyvin tärkeä, sillä se määrittää pohjan koko ongelman ratkaisemiseksi ja täten olennaisen datan keräämiselle. Lisäksi Kotu & Deshpande (2018, luku 2) ja (Sarker 2019, s. 3-4) päätelmien mukaan todettiin, että

itse datalla ja sen laadulla on hyvin keskeisen rooli datatieteen hyödyntämisessä eri ongelmien ratkaisemisessa. Kuten aiemmin Lenz (2016, s. 691) perusteella huomattiin, laina-alustojen tuottamat luottoriskien arviot ovat keskeinen informaation lähde lainanmyöntäjien lainapäätöksiin. Tarkastellaan seuraavaksi, millaista dataa ja datalähteitä voidaan hyödyntää osana lainapäätösten arviointia.

Sekä rakenteellista että rakenteetonta dataa voidaan hyödyntää osana P2P-alustojen lainanhakijoiden maksukyvyn tarkastelussa. Asiakkaiden luottoriskiä voidaan arvioida tavanomaisesti rakenteellisella datalla kuten asiakkaan hakeman lainasumman ja vuosittaisten tulojen perusteella. (Lee 2020, s. 126) Artikkelin mukaan lainanhakijat, joilla on hyvä taloudellinen historia, esimerkiksi hyvä luottopisteytys, saavat lainoilleen usein matalammat korot. Taloudellisella rakenteellisella datalla on Lee (2020, s. 126) havaintojen mukaan siis hyötyä P2P-lainapäätösten arvioinnissa. Kuitenkin aikaisemmin Lenz (2016, s. 693) perusteella huomattiin, että P2P-alustojen voi olla hankala arvioida lainanhakijoita perinteisen taloudellisen informaation avulla sen saatavuuden puutteen takia. Tätä tukee Niu et al. (2019, s. 12) havainto siitä, että esimerkiksi kehittyvissä maissa P2P-alustojen lainanhakijoilta puuttuu usein lainanmaksuun ja taloudelliseen tietoon perustuvaa historiallista dataa.

P2P-lainaamisessa esimerkiksi luottoluokitus ei itsessään ole riittävä tapa erotella lainanhakijoita toisistaan. Luottoluokitusten tarjoama tieto ei anna sijoittajille tarpeeksi selkeää kuvaa asiakkaista, mikä luo sijoittajille riskejä. (Mi et al. 2018, s. 66) Tekstissään Mi et al. (2018, s. 66) painottavat täten, että tavallisen taloudellisen informaation lisäksi voisi olla hyödyllistä käyttää myös lainanhakijoiden henkilökohtaisia tietoja ja edeltävää lainaamisen käyttäytymistä osana luottoriskin arviointia.

Luottoriskiä voidaan ennustaa myös rakenteettomalla datalla, jota saadaan lainahakemusten vapaista tekstikentistä. Lainahakemusten teksteistä voidaan tarkastella ja pyrkiä arvioimaan lainanhakijan lainaamiseen liittyviä tunnetiloja. (Lee 2020, s. 126) Tutkimuksessaan Lee (2020, s. 124) toteaa, että positiivinen ja paljon sanoja sisältävä yksityiskohtainen lainahakemus auttaa P2P-alustoja ymmärtämään lainanhakijan riskejä ja täten asettamaan lainan koron oikealle tasolle. Lee (2020) näkemystä lainahakemusten tunnetilojen tarkastelusta tukee myös Jiang et al. (2017, s. 527) havainnot siitä, että lainatekstien sisältöä voidaan käyttää taloudellisen datan tukena lainanhakijoiden luottoriskin arvioinnissa. Perinteisen rakenteellisen datan lisäksi tekstien sisältöjä käyttämällä voidaan saada aikaan tarkempia luottoriskin arvioita, mistä hyötyvät sekä P2P-alustat että lainanmyöntäjät (Jiang et al. 2017, s. 527).

Lee (2020) ja Jiang et al. (2017) havaintojen perusteella voidaan siis todeta, että sekä rakenteellista usein taloudellista tietoa sisältävää dataa sekä rakenteetonta dataa kuten asiakkaiden lainahakemusten tekstejä voidaan hyödyntää osana P2P-lainapäätöksiä. Tästä voidaan päätellä, että itse lainahakemukset ovat yksi tärkeä datalähde osana lainapäätösten arviointia, joka tarjoaa dataa sekä asiakkaan taloudellisesta tilanteesta että henkilökohtaisesta käyttäytymisestä. Tämä tukee aiemmin luvussa neljä Lenz (2016, s. 693) pohdintoja siitä, että P2P-alustat voivat käyttää vaihtoehtoista dataa perinteisin taloudellisen tiedon lisäksi. Lisäksi havainnot vahvistivat myös Mi et al. (2018, s. 66) havaintoja datan käytön mahdollisuuksista P2P-lainaamisessa.

Lainapäätöksissä hyödynnettävää dataa on mahdollista saada myös P2P-alustan ulkopuolisista datalähteistä, sillä Ge et al. (2017, s. 420) havaintojen mukaan myös sosiaalisten median tietojen perusteella on mahdollisuuksia tarkastella lainanhakijoiden luottoriskiä. Lainahakemuksista saatava data on luvun kolme Hayes (2014) määritelmien ja Lenz (2016, s. 691) P2P-alustan toiminnan havaintojen perusteella laina-alustojen sisäisten prosessien hallinnassa olevaa dataa. Kuitenkin Ge et al. (2017, s. 420) perusteella dataa voidaan saada muualta kuin alustalle keskeisistä asiakkaiden lainahakemuksista, joten lainapäätöksien kannalta hyödyllistä tietoa voidaan siis saada myös ulkoisista datalähteistä. Jos lainanhakija suostuu luovuttamaan tarkasteluun sosiaalisen median tietonsa, voidaan näistä tiedoista arvioida asiakkaan luottoriskiä. Näitä tietoja ovat esimerkiksi lainanhakijan sosiaalisen median viestien määrä ja sosiaalisen verkoston tiedot. (Ge et al. 2017, s. 420) Lisäksi Ge et al. (2017, s. 420) näkemyksiä tukee Niu et al. (2019, s. 12) havainnot, joiden mukaan lainanhakijoiden puhelimesta saadun sosiaalisen verkoston tietojen avulla voitaisiin tuottaa tarkempia luottoriskien arvioita. Sosiaalisen verkoston tarjoamia tietoja voitaisiin täten hyödyntää osana P2P-laina-alustojen toimintaa (Niu et al. 2019, s. 12). Siten Ge et al. (2017) ja Niu et al. (2019) perusteella voidaan todeta, että P2P-lainapäätöksissä on mahdollista hyödyntää sisäisten datalähteiden lisäksi myös ulkoisia datalähteitä.

Teoksessa Lee (2020, s. 118) käytetään avointa dataa LendingClub:sta, joka on yksi maailman suurimmista P2P-alustoista. Kyseisen alustan sivut tarjoavat historiallista dataa vertaislainoista. (Lee 2020, s. 118). Lisäksi tutkimuksessaan Niu et al. (2019, s. 5) käyttävät dataa kiinalaisesta P2P-alustasta. Etenkin Lee (2020) ja toisaalta myös Niu et al. (2019) perusteella voidaan siis todeta, että P2P-laina-alustat voisivat hyödyntää toisten alustojen tarjoamaa avointa dataa osana lainanhakijoiden arviointia. Täten P2P-lainapäätöksissä voitaisiin hyödyntää myös muitakin ulkoisia datalähteitä kuin sosiaalista mediaa.

5.3 Koneoppiminen P2P-lainanhakijoiden arvioinnissa

Kuten Sarker (2021b, s. 6) mukaan luvussa kolme huomattiin, mallintamisen vaiheessa osana datatieteen prosessimallia rakennetaan malli, jossa käytetään hyväksi koneoppimista. Lisäksi Kashyap (2017, luku. 1) perusteella huomattiin, että koneoppimisen avulla voidaan löytää suurista datamassoista erilaisia löydöksiä ja ominaisuuksia. Tarkastellaan seuraavaksi, miten koneoppimista voidaan käyttää lainanhakijoihin liittyvän datan ymmärtämisessä ja täten P2P-lainapäätösten mahdollistamisessa.

Koneoppimismalleja voidaan käyttää onnistuneesti osana P2P-alustojen lainanhakijoiden luottoriskin arviointia (Aleksandrova 2021, s. 141). Artikkelin perusteella huomataan, että luottoriskin arvioinnissa voidaan hyödyntää luokittelumalleja, jotka kuuluvat luvun kolme Chandramouli et al. (2018, luku 1) määritelmien mukaan ohjattuun oppimiseen. Lisäksi Niu et al. (2019, s. 10) käyttivät tutkimuksessaan kolmea eri koneoppimismallia, jossa tarkasteltiin sosiaalisen verkoston merkitystä luottoriskin arviointiin. Niu et al. (2019, s. 10) hyödynsivät tutkimuksessaan myös juuri ohjatun oppimisen luokittelumalleja. Sekä Aleksandrova (2021, s. 135) että Niu & Li (2019, s. 10) teksteissä käytetään luokittelumallien datana rakenteellisessa muodossa olevaa dataa.

Sekä Aleksandrova (2021, s. 141) että Niu et al. (2019, s. 10) tutkimuksista huomataan, että luottoriskin arvioinnissa voidaan hyödyntää etenkin ohjatun oppimisen luokittelumalleja, joiden avulla pyritään luokittelemaan riskialttiita lainanhakijoita hyvistä lainanhakijoista. Koska luokittelumalleilla voidaan jakaa lainanhakijoita hyviin ja huonoihin, voidaan sen avulla ratkaista siis lainanhakijoiden ja P2P-alustan välistä epäsymmetrisen informaation ongelmaa. Ohjatun oppimisen koneoppimismallien avulla voidaan siis luoda hyödyllistä tietoa lainapäätöksentekoa varten.

Aikaisemmin Jiang et al. (2017) ja Lee (2020) avulla huomattiin, että lainahakemusten tekstien sisältöä voidaan käyttää lainanhakijoiden luottoriskin arvioinnissa. Tutkimuksessaan Jiang et al. (2017, 515–518) käyttävät ohjaamattoman oppimisen koneoppimismallia, jonka avulla voidaan erotella tekstistä erilaisia teemoja. Tutkimuksessa pystyttiin tunnistamaan rakenteettomasta lainateksteistä lainanhakijan taloudelliseen tilaan liittyviä aiheita, jotka pystyttiin täten muuttamaan rakenteelliseen muotoon. Täten tunnistettuja aiheita pystyttiin hyödyntämään luottoriskin arvioinnissa. Jiang et al. (2017) havaintojen perusteella voidaan todeta, että ohjaamattomalla oppimisella voidaan analysoida lainatekstien sisältöä ja täten tuottaa lisää tietoa lainapäätöksentekoa varten.

6. YHTEENVETO

6.1 Tulokset

Työssä tutkittiin datatieteen hyödyntämistä P2P-lainapäätöksissä. Työ suoritettiin kirjallisuuskatsauksena alan julkaisuja käyttäen. Tutkimusongelmaa, miten datatiedettä voidaan hyödyntää P2P-lainapäätöksissä, tarkasteltiin ensin määrittelemällä datatieteen sekä P2P-lainaamisen kokonaisuus. Kyseisiä aihepiirejä lähestyttiin kandidaatintyön laajuudessa. Datatieteessä keskeisiksi aiheiksi tunnistettiin data, datalähteet ja koneoppiminen. Puolestaan P2P-lainaamisessa keskityttiin lainanhakijoiden, laina-alustan sekä lainanmyöntäjien välisen prosessin kokonaisuuteen sekä yleisiin ongelmiin. Datatieteen käyttöä P2P-lainaamisessa pyrittiin hahmottamaan ensin kokonaisuutena, minkä avulla voitiin ymmärtää, missä vaiheessa P2P-lainaamisen kokonaisuutta datatiedettä voidaan hyödyntää. Tunnistettuja käyttökohteita pyrittiin tarkentamaan edellisessä luvussa määriteltyjen datatieteen aiheiden eli datan, datalähteiden ja koneoppimisen näkökulmasta.

Ensimmäinen alatutkimuskysymys oli: Mitä datatiede tarkoittaa? Cao (2017), Kotu & Deshpande (2018) ja Sarker (2021b) määritelmien avulla datatiede määriteltiin kokonaisuudeksi, jonka tavoitteena on luoda datasta tietoa päätöksentekoa varten käyttäen hyväksi erilaisia tekniikoita kuten koneoppimista. Toisessa tutkimuskysymys oli: Millaista dataa, datalähteitä ja koneoppimista voidaan käyttää datatieteessä? Tutkimuksessa huomattiin, että datatieteessä voidaan hyödyntää sekä rakenteellista että rakenteetonta dataa ja dataa voidaan hankkia sisäisistä ja ulkoisista datalähteistä. Lisäksi huomattiin, että esimerkiksi Grossi et al. (2021) perusteella yksilöistä on tarjolla valtavasti dataa esimerkiksi sosiaalisen median kautta, jolla voisi olla käyttöä myös P2P-lainaamisen kontekstissa. Puolestaan Sarker (2021b) perusteella koneoppimismallit ovat riippuvaisia siitä, millaista ongelmaa ollaan ratkaisemassa. Tekstissä tunnistettiin kolme keskeistä koneoppimisen osa-aluetta, joita ovat ohjattu, ohjaamaton ja vahvistettu oppiminen. Etenkin ohjatulle ja -ohjaamattomalle oppimiselle tunnistettiin mahdollisia käyttökohteita P2P-lainapäätöksissä.

Tutkimuksen kolmas alatutkimuskysymys oli: Mitä tarkoittaa P2P-lainaaminen ja miten se toimii? Aihetta tarkasteltiin sekä käsitteenä että toimivuuden kokonaisuutta kuvaavan kuvan avulla. P2P-lainaaminen määriteltiin CFI (2022) mukaan suoraksi lainaamiseksi,

joka tapahtuu yksilöiden välillä yleensä verkkoalustoilla ilman virallisen finanssi-instituution osallistumista lainaamisen prosessiin. Lenz (2016) avulla tarkasteltiin P2P-lainaamisen kokonaisuutta prosessina ja lisäksi tunnistettiin P2P-lainaamisen keskeiset ongelmat. Keskeiseksi ongelmaksi tunnistettiin lainanhakijoiden ja laina-alustojen välinen epäsymmetrisen informaation ongelma.

Tutkimuksen viimeisessä osassa vastattiin sekä keskeiseen tutkimuskysymykseen, että viimeiseen alatutkimuskysymykseen. Tutkimuksen pääkysymys oli: Miten datatiedettä voidaan hyödyntää P2P-lainapäätöksissä? Tutkimuksessa huomattiin, että datatieteen avulla voidaan ratkaista lainanhakijan ja laina-alustan välistä epäsymmetrisen informaation ongelmaa. Datatieteen mahdollistamien keinojen avulla voidaan arvioida lainanhakijoiden luottoriskiä, minkä perusteella lainanmyöntäjät pystyvät tekemään lainapäätöksiä. Datatieteen avulla P2P-alusta voi siis tuottaa tietoa lainanhakijoista lainanmyöntäjien päätöksentekoa varten.

Viimeinen alatutkimuskysymys, joka tuki suoraan päätutkimuskysymystä oli: Millaista dataa, datalähteitä ja koneoppimismalleja voidaan hyödyntää P2P-lainapäätöksissä? Tutkimuksessa huomattiin, että lainapäätöksenteossa voidaan hyödyntää laajasta erilaista dataa. Etenkin Lee (2020) perusteella huomattiin, että lainanhakijoiden taloudellisella ja rakenteellisella datalla kuten tuloilla on merkitystä luottoriskin arvioinnissa. Toisaalta tutkimuksessa tunnistettiin laaja mahdollisuus käyttää vaihtoehtoisesti rakenteetonta dataa lainanhakijoiden arvioinnissa. Lee (2020) ja Jiang et al. (2017) havainnoista huomattiin, että lainanhakijoiden lainahakemusten tekstejä voidaan hyödyntää luottoriskin arvioinnissa. Rakenteettomasta datasta nousi esiin myös Ge et al. (2017) esiin tuomat havainnot, joiden perusteella myös lainanhakijoiden sosiaalisen median tietoja voidaan hyödyntää luottoriskien arvioinnissa. Tutkimuksessa siis huomattiin, että P2P-lainapäätöksissä käytettävä data voi olla sekä rakenteellista että rakenteetonta. Lisäksi dataa voidaan saada sisäisistä datalähteistä kuten lainahakemuksista, mutta toisaalta myös ulkoisista lähteistä kuten sosiaalisesta mediasta ja avoimista datalähteistä.

Tutkimuksessa havaittiin, että koneoppimismallien avulla voidaan ennustaa lainanhakijoiden riskejä. Sekä Aleksandrova (2021) että Niu et al. (2019) havaintojen perusteella huomattiin, että luottoriskin arvioinnissa voidaan hyödyntää etenkin ohjatun oppimisen luokittelumalleja. Koneoppimisen avulla voidaan tuottaa lainanhakijoiden datasta ymmärrystä ja täten hyödyntää sitä lainapäätöksiä varten. Lisäksi Jiang et al. (2017) perusteella huomattiin, että ohjaamatonta oppimista voidaan käyttää lainahakemusten tekstien analysointiin. Vahvistetulle oppimiselle ei tutkimuksen perustella löydetty käyttöä P2P-lainapäätöksissä kuten luvussa kolme osattiin olettaa. Alla olevassa taulukossa 4

on esitetty yhteenveto tutkimuksen pääkysymyksen sekä viimeisen alatutkimuskysymyksen tuloksista.

Taulukko 4: *Tutkimuksen keskeiset tulokset.*

Datatiede	**Hyödyntäminen P2P-lainapäätöksissä**
Yleisesti	- Lainanhakijan ja laina-alustan epäsymmetrisen informaation ongelman ratkaiseminen - Lainanhakijoiden luottoriskiarviot - Tuottaa tietoa lainapäätöksiä varten
Data	
Rakenteellinen	- Lainanhakijan taloudellinen data - Luottopisteytys - Tulot, lainasumma
Rakenteeton	- Lainahakemusten tekstikentät - Sosiaalisen median julkaisut - Sosiaalisen verkoston tiedot
Datalähteet	
Sisäiset	- Laina-alustan itse hallinnoima data - Lainahakemukset
Ulkoiset	- Sosiaalinen media - Avoimet datalähteet
Koneoppiminen	
Ohjattu	- Luokittelumallit luottoriskin ennustamisessa
Ohjaamaton	- Lainatekstien sisältöjen analysointi, rakenteettomasta datasta rakenteellista

Taulukon 4 vasemmassa datatieteen sarakkeessa esitellään datatiede yleisesti ja siihen kuuluvat keskeiset termit. Oikeanpuolisessa sarakkeessa kuvaillaan tutkimuksen keskeisiä tuloksia kunkin datatieteen termin osalta. Ensimmäisessä kohdassa esitellään yleisesti datatieteen hyödyntäminen P2P-lainapäätöksissä eli tiivistetään päätutkimuskysymykseen vastaaminen. Seuraavissa kohdissa tiivistetään havainnot viimeiseen alatutkimuskysymykseen saaduista tuloksista kunkin termin osalta.

6.2 Tutkimuksen arviointi ja jatkotutkimusehdotukset

Tutkimus toteutettiin käyttäen hyväksi Finkin (2019) systemaattista kirjallisuuskatsausmallia. Kyseisen mallin tarkoituksena oli parantaa tutkimuksen toistettavuutta ja suoraviivaistaa päätutkimuskysymykseen sekä viimeiseen alatutkimuskysymykseen vastaamista. Kirjallisuuskatsausmallista huolimatta joku toinen tutkimuksen toteuttaja olisi voinut valita saaduista hakutuloksista eri teoksia tutkimukseen, jolloin tutkimuksessa olisi voitu päätyä eri tuloksiin. On siis mahdollista, että tutkimuksen ulkopuolelle jäi joitain tutkimuksen kannalta merkittäviä näkökulmia tai havaintoja.

Tutkimuksen keskeisenä tavoitteena oli selvittää, miten datatiedettä voidaan hyödyntää P2P-lainapäätöksissä. Päätutkimuskysymykseen onnistuttiin vastaamaan työssä käytetyn aineiston perusteella. Kolmeen ensimmäiseen alatutkimuskysymykseen pystyttiin vastaamaan tutkimuksen teoriassa siten, että niiden tuomat havainnot tukivat itse päätutkimuskysymykseen vastaamista. Lisäksi viimeiseen alatutkimuskysymykseen onnistuttiin vastaamaan, joka konkretisoi ja täydensi päätutkimuskysymyksen havaintoja. Tutkimuskysymyksiin vastaamisessa on otettava huomioon kandidaatintyön laajuus.

Tutkimuksessa käsiteltiin datatiedettä lainapäätöksissä juuri P2P-lainojen kontekstissa. Lisäksi havaittiin, että P2P-alustoille ei päde sama sääntely kuin tavanomaisille pankeille. Työssä ei kiinnitetty huomioita lainanhakijoita koskevan datan hyödyntämisen lainsäädäntöön tai eettisyyteen, jolloin tutkimuksessa oli mahdollista käsitellä datatieteen käyttöä lainaamisessa ilman esimerkiksi lainsäädännön merkittäviä rajoituksia. Jatkotutkimuksessa olisi täten mahdollista tutkia, miten datatieteen menetelmät ja innovatiivinen datankäyttö sopisivat esimerkiksi laajemmin säänneltyyn liikepankkitoimintaan. Täten olisi mahdollista vertailla, miten datatieteen käyttö eroaisi P2P-lainaamisesta laajemmin säännellyllä pankkisektorilla. Toisaalta tutkimuksessa jätettiin tietoisesti pois yrityksiä koskevat vertaislainat. Yrityslainoja voitaisiin käsitellä jatkotutkimuksissa, mikä mahdollistaisi datatieteen hyödyntämisen vertailun yksityislainojen ja yrityslainojen lainapäätösten välillä.

Tutkimuksessa datatieteeseen keskityttiin vahvasti päätöksenteon näkökulmasta, jolloin datatieteen kokonaisuutta yksinkertaistettiin eikä sitä käsitelty kovin teknisesti. Datatieteelle saatu määritelmä oli selkeästi riippuvainen kyseisestä näkökulmasta. Jatkotutkimuksissa aihepiirejä voitaisiin käsitellä päätöksenteon lisäksi myös teknisemmällä tasolla, mikä voisi tuottaa tutkimukseen lisää konkretiaa.

LÄHTEET

Aleksandrova, Y. (2021). Comparing Performance of Machine Learning Algorithms for Default Risk Prediction in Peer to Peer Lending. TEM Journal. 10 (1), 133–143.

Bloomenthal, A. (2021). Asymmetric Information. Investopedia. Saatavilla www-osoitteessa: <https://www.investopedia.com/terms/a/asymmetricinformation.asp>, (Viitattu 28.3.2022).

Boobier, T. (2018). Advanced Analytics and AI: Impact, Implementation, and the Future of Work. Advanced Analytics and AI. Newark: John Wiley & Sons, Incorporated.

Cao, L. (2017). Data Science: A Comprehensive Overview. ACM computing surveys. 50 (3), 1–42.

Cao, L., Yang, Q. & Yu, P.S. (2021). Data science and AI in FinTech: an overview. International Journal of Data Science and Analytics. 12 (2), 81–99.

CFI. (2022). What is Peer-to-Peer (P2P) Lending? Saatavilla www-osoitteessa: <https://corporatefinanceinstitute.com/resources/knowledge/finance/peer-to-peer-lending/>, (Viitattu 27.3.2022).

Chandramouli, S., Das, A.K. & Dutt, S. (2018). Machine learning. 1st edition. Pearson Education India.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. The CRISP-DM consortium.

Delua, J. (2021). Supervised vs. Unsupervised Learning: What's the Difference? Saatavilla www-osoitteessa: <https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning>, (Viitattu 7.4.2022).

Ferretti, F. (2021). Peer-to-Peer lending and EU credit laws: A creditworthiness assessment, credit-risk analysis or... neither of the two? *German law journal*. 22 (1), 102–121.

Fink, A. (2019). Conducting Research Literature Reviews: From the Internet to Paper. 5th ed. SAGE Publications. 1–18.

Ge, R., Feng, J., Gu, B. & Zhang, P. (2017). Predicting and Deterring Default with Social Media Information in Peer-to-Peer Lending. Journal of Management Information Systems, 34 (2), 401–424.

Giudici, P. (2018). Financial data science. Statistics & probability letters. 136, 160–164.

Grossi, V., Giannotti, F., Pedreschi, D., Manghi, P., Pagano, P. & Assante, M. (2021). Data science: a game changer for science and innovation. *International journal of data science and analytics*. 11 (4), 263–278.

Hayes, B. (2014). The What and Where of Big Data: A Data Definition Framework. Business Over Broadway. Saatavilla www-osoitteessa: <https://businessoverbroadway.com/2014/07/30/the-what-and-where-of-big-data-a-data-definition-framework/>, (Viitattu 4.3.2022).

Hurwitz, J., Nugent, A., Halper, F. & Kaufman, M. (2013). Big Data for Dummies. John Wiley & Sons, Incorporated, Somerset.

IBM Cloud Education. (2020). Natural Language Processing. Saatavilla www-osoitteessa: <https://www.ibm.com/cloud/learn/natural-language-processing>, (Viitattu 8.4.2022).

IBM Cloud Education. (2021). Structured vs. Unstructured Data: What's the Difference? Saatavilla www-osoitteessa: <https://www.ibm.com/cloud/blog/structured-vs-unstructured-data>, (Viitattu 8.4.2022).

Investopedia. (2021). Credit Risk. Saatavilla www-osoitteessa: <https://www.investopedia.com/terms/c/creditrisk.asp> (Viitattu 15.2.2022).

Jagtiani, J. & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. Financial Management. 48 (4), 1009–1029.

Jiang, C., Wang, Z., Wang, R. & Ding, Y. (2017). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. Annals of Operations Research. 266(1-2), 511–529.

Kampakis, S. (2020) *The Decision Maker's Handbook to Data Science A Guide for Non-Technical Executives, Managers, and Founders*. 2nd ed. Berkeley, CA: Apress L. P.

Kashyap, P. (2017). *Machine Learning for Decision Makers Cognitive Computing Fundamentals for Better Decision Making*. 1st ed. Berkeley, CA: Apress L. P.

Kotu, V. & Deshpande, B. (2018) *Data Science: Concepts and Practice*. 2nd Edition. Morgan Kaufman.

Lahti, L. (2018). Open Data Science. Advances in intelligent Data Analysis XVII. Springer International Publishing. 31–39.

Lee, J.Y. (2020). Prediction of Default Risk in Peer-to-Peer Lending Using Structured and Unstructured Data. Financial counseling and planning. 31 (1), 115–129.

Lenz, R. (2016) Peer-to-Peer Lending: Opportunities and Risks. *European journal of risk regulation*. 7 (4), 688–700.

Mi, J.J, Hu, T. & Deer, L. (2018). User Data Can Tell Defaulters in P2P Lending. Annals of Data Science. 5 (1), 59–67.

Microsoft. (2022a). The Team Data Science Process lifecycle. Saatavilla www-osoitteessa: <https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview>, (Viitattu 2.3.2022).

Microsoft. (2022b). Modeling stage of the Team Data Science Process lifecycle Saatavilla www-osoitteessa: <https://docs.microsoft.com/en-us/azure/architecture/data-science-process/lifecycle-modeling>, (Viitattu 8.4.2022).

Niu, B., Ren, J. & Li, X. (2019). Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending. Information (Basel). 10 (12), 397–.

Ozdemir, S. (2016). Principles of Data Science.1st edition. Packt Publishing.

Patel, H. (2019). These Are The Best Free Open Data Sources Anyone Can Use. Saatavilla www-osoitteessa: <https://www.freecodecamp.org/news/https-medium-freecodecamp-org-best-free-open-data-sources-anyone-can-use-a65b514b0f2d/>, (Viitattu: 9.4.2022).

Sarker, I. H. (2019). A machine learning based robust prediction model for real-life mobile phone data. Internet of Things: Engineering Cyber Physical Human Systems. (5), 180–193.

Sarker, I. H. (2021a) Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN computer science*. 2 (3), 160.

Sarker, I. H. (2021b). Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. SN Computer Science. 2 (5), 377.

Sas. (2022). Machine learning: What it is and why it matters. Saatavilla www-osoitteessa: <https://www.sas.com/en_us/insights/analytics/machine-learning.html>, (Viitattu 9.3.2022).

Schneider, B. (2022). Peer-to-Peer Lending Breaks Down Financial Borders. Investopedia. Saatavilla www-osoitteessa: <https://www.investopedia.com/articles/financial-theory/08/peer-to-peer-lending.asp>, (Viitattu 27.3.2022).

Zwilling, M., Klein, G. & Shtudiner, Z. (2020) Peer-to-peer lending platforms' legitimacy in the eyes of the general public and lenders. *Israel affairs*. 26 (6), 854–874.