Koneoppimisen hyödyntäminen sijoituskohteiden tunnistamisessa: pääomasijoitusyhtiöiden näkökulma

Turun yliopisto
Tietotekniikan laitos
TkK-tutkielma
Tietotekniikka
Toukokuu 2026
Tom Pippuri

TURUN YLIOPISTO
Tietotekniikan laitos

Tom Pippuri: Koneoppimisen hyödyntäminen sijoituskohteiden tunnistamisessa: pääomasijoitusyhtiöiden näkökulma

TkK-tutkielma, 27 s.
Tietotekniikka
Toukokuu 2026

Riskipääomasijoittaminen on noussut keskusteluihin merkittävänä talouskasvun kiihdyttäjänä ja sen tavoitteena on rahoittaa varhaisen vaiheen kasvuyrityksiä. Tässä tutkielmassa tarkastellaan koneoppimisen hyödyntämistä sijoituskohteiden tunnistamisessa pääomasijoitusyhtiöiden näkökulmasta.
Tutkielman tavoitteena on selvittää, millaisia mahdollisuuksia ja rajoitteita koneoppimiseen perustuvilla menetelmillä on erityisesti sijoituskohteiden esiseulonnassa. Tutkielma on toteutettu kirjallisuuskatsauksena, jossa tarkastellaan aiempaa tutkimusta koneoppimisen hyödyntämisestä sijoituskohteiden tunnistamisessa riskipääomasijoitustoiminnassa.
Tutkielmassa voidaan havaita, että koneoppimisella on potentiaalia tehostaa sijoituskohteiden tunnistamista, sillä menetelmien avulla voidaan suodattaa potentiaalisia sijoituskohteita ihmisen arviointia varten sekä niiden avulla voidaan tunnistaa päätöksenteon kannalta relevantteja signaaleja. Toisaalta koneoppimisen hyödyntämistä rajoittavat aineistojen epätasapainoisuus, datan vaihteleva laatu sekä mallien selitettävyyden haasteet.
Kirjallisuudessa korostuu ihmisen asiantuntija-arvioinnin merkitys, minkä vuoksi koneoppimista ei voida pitää itsenäisenä ratkaisuna sijoituspäätösten tekemiseen. Tutkielman perusteella koneoppiminen näyttäytyy lupaavana työkaluna sijoituskohteiden esiseulonnassa, mutta sen laajamittainen hyödyntäminen edellyttää lisää tutkimusta erityisesti mallien suorituskyvyn yleistettävyyden, tulkittavuuden ja käytännön sovellettavuuden näkökulmasta.

Asiasanat: koneoppiminen, koneoppimismenetelmät, riskipääomasijoittaminen, esiseulonta, päätöksenteko

Sisällys

Kuvat

Taulukot

1 Johdanto

Sijoituskohteiden tunnistaminen on keskeinen osa riskipääomasijoittamista, sillä sijoittajien on kyettävä erottamaan potentiaalisimmat yritykset laajasta kohdejoukosta ja kohdentamaan rahoitus niihin riittävän varhaisessa vaiheessa [1, 2]. Viime aikoina tekoälyn kehitys on tehostanut monien toimialojen prosesseja sekä parantanut dataan perustuvaa päätöksentekoa, jonka seurauksena sijoittajien kiinnostus tekoälyn hyödyntämiseen sijoituskohteiden tunnistamisessa on kasvanut. Röhm et al. [3] havaitsivat, että vuonna 2022 suurin osa pääomasijoitusyhtiöistä ei vielä hyödyntänyt tekoälyä. Ne pääomasijoitusyhtiöt, jotka hyödynsivät tekoälyä, pyrkivät pääasiassa tehostamaan sijoituskohteiden hankintaa [3].

Aihe on kuitenkin vielä suhteellisen uusi, ja sitä koskeva tutkimus on toistaiseksi rajallista, mistä syystä tekoälyn tarjoamia mahdollisuuksia ei ole vielä hyödynnetty täysimääräisesti. Tekoäly kattaa kuitenkin laajan kokonaisuuden erilaisia menetelmiä, minkä vuoksi tässä tutkielmassa tarkastelu rajataan koneoppimiseen, joka muodostaa keskeisen osan tekoälyä.

Olen havainnut kirjallisuudessa tutkimusaukon, joka liittyy siihen, että sijoituskohteiden tunnistamista riskipääomasijoittamisen esiseulontavaiheessa ei ole aiemmassa tutkimuksessa tarkasteltu riittävän kokonaisvaltaisesti siten, että samassa tarkastelussa huomioitaisiin sekä käytännön toimintaan liittyvät haasteet että koneoppimismenetelmien tarjoamat hyödyt, mahdollisuudet ja rajoitteet. Aiheen tutkimi-

nen on tärkeää, koska riskipääomasijoittamisella on merkittävä rooli yhteiskunnassa talouskasvun sekä uusien innovaatioiden edistämisessä [4].

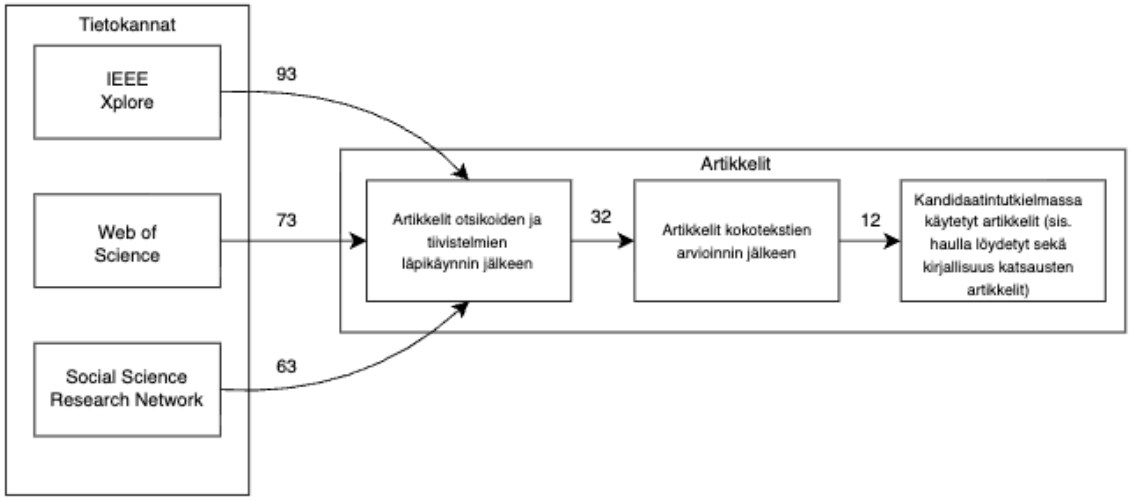
Tämän tutkielman tavoitteena on vastata edellä kuvattuun tutkimusaukkoon kokoamalla yhteen aiempaa tutkimusta sijoituskohteiden tunnistamiseen liittyvistä haasteista sekä koneoppimisen hyödyntämisestä näiden haasteiden ratkaisemisessa riskipääomasijoittamisen esiseulontavaiheessa. Tutkielmassa tarkastellaan aihetta kokonaisvaltaisesti yhdistämällä käytännön toimintaan liittyvien haasteiden tarkastelu koneoppimismenetelmien tarjoamien hyötyjen, mahdollisuuksien ja rajoitteiden arviointiin. Tutkielma on toteutettu kirjallisuuskatsauksena eikä sisällä omaa empiiristä aineistoa. Tutkielman tavoitteet pyritään saavuttamaan seuraavien tutkimuskysymysten avulla (TK) avulla:

TK 1: Miten koneoppimista voidaan hyödyntää riskipääomasijoittamisen esiseulontavaiheessa sijoituskohteiden tunnistamisessa?

TK 2: Mitä hyötyjä ja rajoitteita koneoppimisen käyttöön liittyy riskipääomasijoittamisen esiseulontavaiheessa?

Tutkielmassa hyödynnetyt artikkelit on kerätty kolmesta tietokannasta: **Web of Science**, **IEEE Xplore** ja **Social Science Research Network**. Lisäksi osa lähteistä on otettu aiemmista kirjallisuuskatsauksista sekä Turun yliopiston Volter-tietokannasta. Aiheen suomenkielinen terminologia ei ole vakiintunut, minkä vuoksi kaikki haut toteutettiin englannin kielellä. Löydettyjen artikkelien pohjalta hakua täydennettiin lisäämällä uusia hakusanoja, ja osa artikkeleista on tunnistettu kirjallisuuskatsausten lähdeluetteloiden kautta. Tiedonhaun prosessi on esitetty kuvassa 1.1. Hakutuloksista löytyi 73 artikkelia Web of Sciencestä haulla "venture capital" AND "machine learning". Social Science Research Network:stä vastaavalla haulla löytyi 63 artikkelia. IEEE Xplore:stä vastaavilla hakusanoilla löytyi 93 artikkelia.

Tulokset käytiin aluksi läpi otsikoiden ja tiivistelmien perusteella. Tämän jälkeen tarkempaan tarkasteluun valittiin tutkimukset, jotka käsittelivät koneoppimisen hyödyntämistä sijoituskohteiden tunnistamisessa erityisesti varhaisen vaiheen yritysten menestyksen ennustamisen ja niitä koskevan pääomasijoittajien päätöksenteon tukemisen näkökulmasta. Lopulliset 12 artikkelia valikoituivat tutkielmaan niiden tutkimuskysymysten, sisällöllisen osuvuuden, aiheellisen rajauksen ja menetelmällisen relevanssin perusteella suhteessa esiseulontavaiheen sijoituskohteiden tunnistamiseen.



Kuva 1.1: Tiedonhaun vaiheet.

Tutkielman luvussa 2 avataan riskipääomasijoittamisen ominaispiirteitä sekä sijoituskohteiden tunnistamiseen liittyviä keskeisiä haasteita. Tämän jälkeen luvussa 3 perehdytään siihen, miten koneoppimista voidaan hyödyntää näiden haasteiden ratkaisemisessa ja sijoituskohteiden esiseulonnan tukemisessa. Luku 4 tarkastelee koneoppimisen tarjoamia hyötyjä ja siihen liittyviä rajoitteita sekä arvioi nykyisten menetelmien soveltuvuutta käsiteltävän ongelman ratkaisemiseen. Lopuksi luvut 5 ja 6 kokoavat tutkielman keskeiset havainnot pohdinnan ja yhteenvedon muodossa.

2 Sijoituskohteiden tunnistaminen ja sen haasteet

Jotta voidaan tarkastella koneoppimisen hyödyntämistä sijoituskohteiden tunnistamisessa, on aluksi olennaista ymmärtää pääomasijoitusyhtiön ominaispiirteet ja toimintatavat. Tässä luvussa tarkastellaan pääomasijoitusyhtiön rakennetta ja päätöksentekoa sekä sijoituskohteiden tunnistamista osana hankevirran hallintaa. Luvun tarkoituksena on kuvata, miten potentiaalisia sijoituskohteita perinteisesti etsitään ja arvioidaan sekä millaisia haasteita tähän prosessiin liittyy.

2.1 Riskipääomasijoittaminen

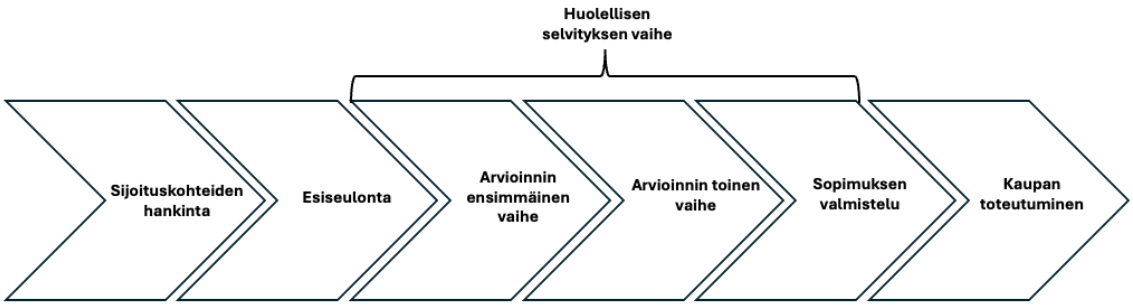
Pääomasijoittaminen (engl. *private equity*) tarkoittaa ammattimaista sijoitustoimintaa, jossa sijoittaja tarjoaa pääomaa pörssilistaamattomille yrityksille vastineeksi omistusosuudesta [5]. Vaikka pääomasijoittaminen kattaa useita eri alasegmenttejä, tässä tutkielmassa keskitytään riskipääomasijoittamiseen (engl. *venture capital*), joka kohdistuu varhaisen vaiheen yrityksiin. Riskipääomasijoittamisen tavoitteena on kasvattaa yrityksen arvoa aktiivisen omistajuuden keinoin ja realisoida arvonnousu irtautumalla sijoituksesta yrityskaupan (engl. *exit*) tai pörssiin listautumisen (engl. *initial public offering*, IPO) kautta [6].

Pääomasijoitusyhtiö on asiantuntijaorganisaatio, joka harjoittaa riskipääomasijoittamista ja keskittyy varhaisen vaiheen yrityksiin. Näissä yhtiöissä pääomasi-

joittajat (engl. *venture capitalists*) vastaavat sijoituskohteiden hankinnasta ja alkuvaiheen arvioinnista. Yhtiötä hallinnoivat vastuunalaiset yhtiömiehet (engl. *general partners*), jotka tekevät lopulliset sijoituspäätökset [7].

2.2 Hankevirta

Pääomasijoitusyhtiön sijoitusprosessi muodostaa laajan kokonaisuuden, minkä vuoksi tarkastelu rajataan hankevirtaan (engl. *deal flow*), jossa sijoituskohteiden tunnistaminen ja arviointi ovat keskeisiä. Hankevirta jäsennetään kirjallisuudessa yleisimmin kuusivaiheisen mallin avulla (ks. kuva 2.1), joka mahdollistaa prosessin tarkastelun yhtenäisenä ja loogisesti etenevänä kokonaisuutena aina sijoituskohteiden hankinnasta kaupan toteutumiseen [1].



Kuva 2.1: Hankevirta. Lähde: Mukailtu [1, 6]

Mallia täydennetään huolellisen selvityksen vaiheella (engl. *due diligence phase*), jossa sijoittaja arvioi kohteen taloudellisia, oikeudellisia sekä operatiivisia riskejä ennen lopullisen sijoituspäätöksen tekemistä. Tiedon saatavuuden ja teknisen kehityksen myötä huolellisesta selvityksestä on tullut keskeinen osa arviointiprosessia erityisesti riskienhallinnan näkökulmasta [8].

Hankevirta alkaa sijoituskohteiden hankinnalla (engl. *deal sourcing*), jossa pääomasijoitusyhtiö rakentaa ja ylläpitää jatkuvaa virtaa sijoituskohteista. Sijoituskohteita hankitaan perinteisesti suositusten ja verkostojen välityksellä sekä suorien yhteydenottojen avulla [9]. Hankintavaihetta seuraa esiseulonta (engl. *pre-screening*), jossa laajaa yritysmassaa suodatetaan pääomasijoitusyhtiön sijoitusstrategian mukaisesti. Suodatuksen tukena hyödynnetään manuaalisia pistekortteja, joilla yritykset pisteytetään valintakriteerien mukaan [10]. Näiden avulla sijoituskohteet voidaan luokitella niin, että sijoittajat käyttävät aikansa kaikkein lupaavimpiin kohteisiin. Varhaisen vaiheen yrityksissä sijoittajat reagoivat voimakkaasti perustajatiimiä koskevaan informaatioon, kuten operatiiviseen kyvykkyyteen, akateemiseen taustaan sekä pitkäjänteiseen sitoutumiseen [11]. Tämän lisäksi sijoittajat arvioivat muun muassa markkinan viehättävyyttä, kilpailua, myynnillistä näyttöä, investointivaihetta sekä tuotetta tai palvelua [1, 2]. Esiseulonnan tavoitteena on tunnistaa potentiaalisimmat kohteet mahdollisimman nopeasti, sillä vain pieni osa yrityksistä etenee tarkempaan arviointivaiheeseen. Tuloksena syntyy rajattu joukko potentiaalisia sijoituskohteita.

Arvioinnin ensimmäisessä vaiheessa (engl. *first phase*) sijoittajat valitsevat suodatetusta joukosta potentiaalisia sijoituskohteita ja arvioivat perustajatiimin osaamista sekä liiketoimintasuunnitelman uskottavuutta hyödyntäen sidosryhmien ja asiakkaiden referenssejä. Usein tässä vaiheessa päätöksenteon tukena hyödynnetään myös teknisiä arvioita ja asiantuntijalausuntoja [1]. Ensimmäistä vaihetta voidaan pitää keskeisenä, sillä siinä päätetään, eteneekö sijoituskohde arvioinnin toiseen vaiheeseen (engl. *second phase*) investointikomitealle (engl. *investment committee*). Investointikomitea vastaa lopullisen sijoituspäätöksen tekemisestä ja koostuu tyypillisesti vastuunalaisista yhtiömiehistä, jotka määrittävät sijoitusstrategian ja valintakriteerit. Laadukkaan hankevirran takaamiseksi esiseulonnan ja alkuvaiheen arvioinnin tulee olla linjassa strategisten painotusten kanssa.

Kun arvioinnin toisessa vaiheessa investointikomitea on tehnyt lopullisen sijoituspäätöksen, alkaa sopimuksen valmistelu (engl. *deal structuring*), johon kuuluu yrityksen arvon (engl. *valuation*) määrittäminen sekä pääomasijoitusyhtiön ja yrityksen välisten sopimusehtojen neuvotteleminen. Sopimusehdoissa neuvotellaan tyypillisesti määräysvallasta ja likviditaatio-oikeuksista [2]. Kun osapuolet ovat päässeet yksimielisyyteen sopimuksen ehdoista, sopimus allekirjoitetaan ja kauppa toteutetaan (engl. *closing*).

2.3 Haasteet sijoituskohteiden tunnistamisessa

Aiemman tutkimuksen mukaan pääomasijoitusyhtiöiden hankevirta on merkittävästi saapuvapainotteinen (engl. *inbound*), eli yritykset ottavat yhteyttä sijoittajiin [12]. Tutkimuksessa korostuu sijoituskohteiden merkittävä kasvu, mikä on haastanut sijoittajien sisäisiä resursseja saapuvien yhteydenottojen käsittelyssä. Manuaalisen esiseulonnan vuoksi sijoittajat eivät kykene käsittelemään kasvavaa sijoituskohteiden määrää, mikä pakottaa heidät joko jättämään potentiaalisia sijoituskohteita arvioimatta tai käymään saatavilla olevan tiedon nopeasti läpi. Tämän on huomattu kasvattavan väärän luokittelun riskiä [9]. Lisäksi on havaittu, että sijoittajat ovat ylioptimistisia omista seulontapäätöksistään, mikä voi johtaa subjektiivisiin ja siten epäoptimaalisiin valintoihin [3, 10].

Sijoituskohteiden tunnistamiseen liittyy paljon epävarmuustekijöitä erityisesti varhaisen vaiheen yrityksissä, joista ei tyypillisesti ole ollut saatavilla riittävästi kvantitatiivista historiatietoa laaja-alaiseen arviointiin [12]. Tämä viittaa rahoitusmarkkinoilla yleiseen ilmiöön, informaatioasymmetriaan, jossa eri osapuolilla on eritasoinen määrä tietoa kohdeyrityksen liiketoiminnasta [13]. Varhaisen vaiheen yrityksissä informaatioasymmetria korostuu, sillä perustajatiimillä on tyypillisesti enemmän tietoa yrityksen teknologiasta, liiketoimintamallista ja tulevaisuuden näkymistä kuin ulkopuolisilla sijoittajilla. Ilmiön seurauksena sijoittajien on vaikeampi

arvioida yrityksen todellista potentiaalia ja riskiä, minkä vuoksi he ovat joutuneet tekemään päätöksiä vähäisen tiedon ja intuition pohjalta.

Kuitenkin kirjallisuudessa on havaittu, että useat kaupalliset datan tarjoajat ovat tunnistaneet tämän epäsymmetrian liiketoimintamahdollisuutena ja alkaneet kerätä sekä tarjota tietoa listaamattomista yrityksistä [9]. Tiedon määrän kasvaessa pääomasijoittajat ovat alkaneet hyödyntää yhä enemmän erilaisia datalähteitä osana sijoituskohteiden tunnistamista. Samalla sekä tutkijat että pääomasijoittajat ovat alkaneet kehittää tehokkuutta lisääviä ja arviointia objektiivistavia menetelmiä, kuten koneoppimispohjaisia seulontatyökaluja.

3 Koneoppimisen hyödyntäminen sijoituskohteiden esiseulonnassa

Tässä luvussa käsitellään, miten koneoppimista voidaan hyödyntää sijoituskohteiden esiseulontavaiheessa. Luvussa tarkastellaan koneoppimisen teknistä perustaa esiseulonnan näkökulmasta sekä keskeisiä menetelmiä, joita kirjallisuudessa on sovellettu sijoituskohteiden tunnistamiseen. Tarkastelun tavoitteena on jäsentää, millä tavoin suuria yritysmassoja voidaan analysoida ja suodattaa arviointiin soveltuviksi kohteiksi.

3.1 Datatyypit ja esikäsittely

Koneoppiminen perustuu eri lähteistä kerätyn datan monipuoliseen hyödyntämiseen. Tässä tutkielmassa datalla viitataan sijoituskohteiden tunnistamisessa hyödynnettävään yritysdataan, joka voi olla sekä numeerista että tekstimuotoista.

Varhaisen vaiheen yrityksissä perinteisen talousdatan puutteen vuoksi esiseulontavaiheen suodatus ja arviointi eivät voi perustua yksinomaan tällaiseen aineistoon, vaan ne painottuvat merkittävissä määrin muihin ulkopuolisiin aineistoihin. Rahoituskirjallisuudessa perinteisen talousdatan ulkopuolista aineistoa kutsutaan vaihtoehtoiseksi dataksi (engl. *alternative data*), jota on teknologisen kehityksen myötä yhä helpommin saatavilla internetin sekä erilaisten teknisten rajapintojen kautta [14, 15]. Vaihtoehtoinen data koostuu tyypillisesti tekstistä, kuvista ja äänestä, minkä

vuoksi se eroaa perinteisestä numeerisesta talousdatasta. Riskipääomasijoittamisessa vaihtoehtoinen data muodostuu usein yritysten tuottamasta tekstimuotoisesta sekä digitaalisesta aineistosta, kuten yritysesittelydokumenteista (engl. *pitch deck*), verkkosivusisällöistä, sosiaalisen median profiileista ja julkaisuista [14].

Nämä aineistot sisältävät usein erilaisia datatyyppejä, jotka voidaan luokitella niiden rakenteen ja formaatin perusteella. Formaatit (engl. *formats*) määrittävät, miten data tallennetaan ja käsitellään, kun taas rakenteet (engl. *structures*) kuvaavat, miten data on järjestetty ja kuinka helposti koneoppimismallit voivat hyödyntää sitä. Näiden erojen ymmärtäminen auttaa arvioimaan, millainen aineisto soveltuu parhaiten koneoppimismallien kouluttamiseen.

Hyödynnettävä data voi olla rakenteeltaan strukturoitua, strukturoimatonta tai puolistrukturoitua [16]. Strukturoitu data on jäsenneltyä ja organisoitua siten, että se esitetään tyypillisesti taulukkomuodossa. Koneoppimisessa taulukkomuotoisesta aineistosta käytetään nimitystä tabulaarinen data (engl. *tabular data*), jossa havainnot esitetään riveinä ja muuttujat sarakkeina. Riskipääomasijoittamisessa strukturoitu data voi sisältää tietoa esimerkiksi perustamisvuodesta, rahoituskierroksista, henkilöstömääristä ja perustajatiimin koulutustasosta [17, 18]. Strukturoimaton ja puolistrukturoitu data koostuu usein yrityksen digitaalisista materiaaleista ja muista tekstipohjaisista aineistoista.

Raakadata ei ole suoraan käyttökelpoista sellaisenaan, vaan se täytyy ensin esikäsitellä (engl. *data pre-processing*), jotta sitä voidaan hyödyntää koneoppimalleissa [19]. Riskipääomasijoitustoiminnassa data on kerätty aiemmin kuvatuista lähteistä, minkä vuoksi merkittävä osa datasta on luonteeltaan heterogeenistä (engl. *heterogeneous data*). Heterogeeninen raakadata on usein myös epätäydellistä, kohinaista ja epäjohdonmukaista, mikä voi johtaa virheisiin mallintamisessa [20].

Datan esikäsittelyyn kuuluu tyypillisesti datan kerääminen, puhdistaminen, integroiminen, muuntaminen, vähentäminen sekä jakaminen opetus- ja testijoukkoihin

[16, 19]. Näiden vaiheiden tavoitteena on muokata eri lähteistä kerätty aineisto sellaiseen muotoon, että sitä voidaan hyödyntää luotettavasti koneoppimismallien kouluttamisessa. Riskipääomasijoittamisessa aineisto on usein rakenteeltaan vaihtelevaa ja laadultaan epätasaista, minkä vuoksi huolellinen esikäsittely on keskeistä.

3.2 Esiseulonnan mallintaminen koneoppimistehtävänä

Koneoppimismenetelmien, mallien arvioinnin ja selitettävyyden tarkastelu edellyttää ensin esiseulontaprosessin mallintamista koneoppimistehtäväksi. Tämän vuoksi on olennaista määritellä, millaisena ongelmana sijoituskohteiden esiseulonta nähdään koneoppimisen näkökulmasta.

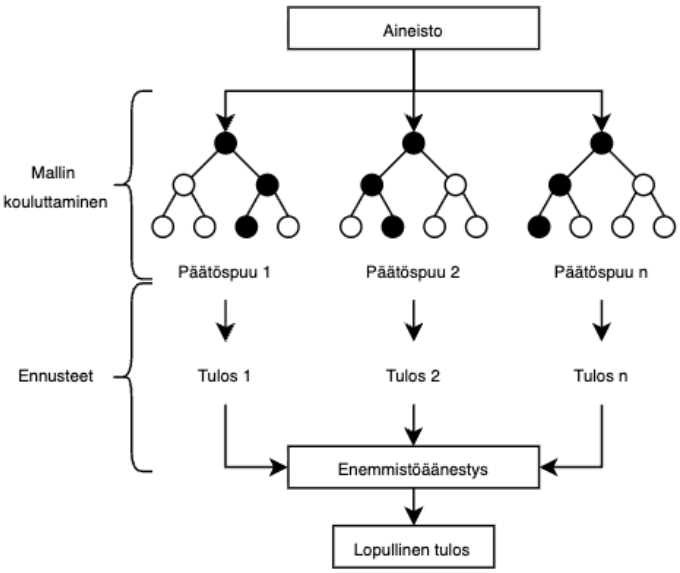
Esiseulonta voidaan mallintaa binäärisenä luokitteluongelmana (engl. *classification*), jossa koneoppimismallin tarkoituksena on ennustaa, eteneekö yritys jatkoarviointiin vai hylätäänkö se esiseulonnassa [10]. Tällöin mallin kohdemuuttuja (engl. *label*) kuvaa esiseulonnan päätöstä ja mallin tehtävänä on oppia aineistosta, millaiset yritykset tyypillisesti etenevät hankevirrassa. Mallin syötemuuttujat (engl. *features*) koostuvat aiemmin kuvatuista valintakriteereistä, joita käytetään mallin koulutuksessa. Luokittelun avulla voidaan suodattaa merkittävä osa epärelevanteista yrityksistä ja tunnistaa yritykset, jotka ovat potentiaalisia sijoituskohteita.

Esiseulontaa voidaan tarkastella myös järjestämis- ja priorisointiongelmana (engl. *ranking*), jossa yritykset asetetaan paremmuusjärjestykseen niiden ennustetun sijoituspotentiaalin perusteella [18, 21]. Tällöin mallin tavoitteena ei ole pelkästään luokitella yrityksiä, vaan tuottaa priorisoitu joukko potentiaalisia sijoituskohteita. Näin analyysi voidaan kohdistaa ensisijaisesti korkeimman potentiaalin yrityksiin.

3.3 Koneoppimismenetelmät esiseulonnassa

Luokittelu- ja järjestämisongelmien ratkaisuun voidaan hyödyntää useita koneoppimismenetelmiä. Kirjallisuudessa yleisemmin käytetyt menetelmät jakautuvat puupohjaisiin ensemble-menetelmiin (engl. *tree-based ensemble methods*) sekä klassisiin luokittelumenetelmiin (engl. *classical classification methods*) [10, 18, 22].

Tyypillisimmin puupohjaisina ensemble-menetelminä on käytetty satunnaismetsää (engl. *Random Forest*, RF), gradienttivahvistusta (engl. *Gradient Boosting*, GB) sekä XGBoostia (engl. *Extreme Gradient Boosting*). Puupohjaiset ensemble-menetelmät ovat osoittautuneet tehokkaiksi ennustettaessa varhaisen vaiheen yrityksen menestystekijöitä [10, 21]. Näiden menetelmien toimintaperiaate on esitetty kuvassa 3.1. Aineisto syötetään useille päätöspuille, joiden yksittäiset ennusteet yhdistetään enemmistöäänestyksen avulla lopullisen ennusteen muodostamiseksi. Kuvassa mustat ja valkoiset pallot havainnollistavat päätöspuiden tekemiä luokittelupäätöksiä, joissa musta kuvaa hyväksyttyä ja valkoinen hylättyä päätöstä.



Kuva 3.1: Puupohjaisen ensemble-menetelmän toimintaperiaate. Lähde: Mukailtu [16].

Klassisia luokittelumenetelmiä hyödynnetään usein vertailumalleina (engl. *baseline models*) yksinkertaisuutensa ja tulkittavuutensa vuoksi. Näiden menetelmien suorituskykyä verrataan monimutkaisempien puupohjaisten ensemble-menetelmien suorituskykyyn eri arviointimetriikoiden avulla, sillä on olennaista arvioida, voiko yksinkertaisemmilla menetelmillä saavuttaa vastaavan suorituskyvyn [22]. Vertailumalleina on käytetty muun muassa tukivektorikonetta (engl. *Support Vector Machine*, SVM), k-lähimmän naapurin menetelmää (engl. *K-Nearest Neighbors*, KNN) sekä logistista regressiota (engl. *Logistic Regression*, LR) [17, 21, 22].

Satunnaismetsä on koneoppimiseen perustuva algoritmi, jota käytetään laajasti luokitteluongelmien ratkaisemiseen. Se koostuu useista toisistaan riippumattomista päätöspuista, joiden ennusteet yhdistetään enemmistöäänestyksen avulla lopullisen luokittelun määrittämiseksi. Menetelmä soveltuu hyvin tabulaarisen ja heterogeenisen datan käsittelyyn. Tämä tekee siitä käyttökelpoisen riskipääomasijoittamisen kontekstissa, jossa sijoituskohteita kuvaava data koostuu yleisimmin useista erilaisista muuttujista. Tämän lisäksi satunnaismetsä kykenee käsittelemään epälineaarisia riippuvuuksia ja on suhteellisen vankka käsittelemään kohinaa sekä puuttuvia arvoja [23]. Menetelmä on osoittanut hyvää suorituskykyä varhaisen vaiheen yritysten valinnassa sekä niiden menestyksen ennustamisessa [17, 18].

Gradienttivahvistus on ensemble-menetelmä, jossa uusia päätöspuita rakennetaan iteratiivisesti siten, että jokainen uusi malli pyrkii korjaamaan aiempien mallien virheitä sovittamalla mallin häviöfunktion negatiiviseen gradienttiin [24]. Toisin kuin satunnaismetsässä, gradienttivahvistuksessa mallit muodostavat peräkkäisen ja keskenään riippuvan kokonaisuuden, mikä mahdollistaa korkean ennustetarkkuuden.

XGBoost on gradienttivahvistukseen perustuva optimoitu koneoppimisalgoritmi, jota käytetään luokittelu- sekä järjestämisongelmien ratkaisuun [24]. Se sisältää optimointeja regularisaatioon ja rinnakkaislaskentaan, jotka nopeuttavat mallin koulutusta sekä auttavat estämään ylioppimista. XGBoost on saavuttanut laajaa suo-

siota korkean suorituskykynsä ja laskennallisen tehokkuutensa vuoksi [24]. Nämä menetelmät soveltuvat erityisen hyvin tabulaarisen datan käsittelyyn ja kykenevät mallintamaan monimutkaisia sekä epälineaarisia riippuvuuksia muuttujien välillä.

Tukivektorikone on koneoppimisalgoritmi, jota käytetään luokittelutehtäviin. Se pyrkii löytämään optimaalisen erotuspinnan eri luokkien välille, jotta uudet havainnot voidaan luokitella mahdollisimman tarkasti [25]. Tukivektoreiksi kutsutaan niitä havaintoja, jotka sijaitsevat lähimpänä luokkien välistä rajaa ja siten määräävät erotuspinnan sijainnin. Menetelmä on saavuttaa usein hyvän suorituskyvyn myös ilman puupohjaista rakennetta.

K-lähimmän naapurin menetelmä on klassinen luokittelualgoritmi, jossa uusi havainto luokitellaan sitä lähimpänä olevien naapureiden perusteella [26]. Siinä luokittelupäätös tehdään vasta ennustevaiheessa hyödyntämällä lähimpien havaintojen tietoa. Menetelmä on yksinkertainen ja helposti tulkittava.

Logistinen regressio on vakiintunut luokittelumenetelmä, jota käytetään erityisesti binäärisiin luokitteluongelmiin [27]. Menetelmä arvioi todennäköisyyttä, että havainto kuuluu tiettyyn luokkaan. Logistisen regression vahvuutena on hyvä tulkittavuus, sillä sen avulla voidaan tarkastella eri muuttujien yhteyttä luokittelupäätökseen.

Edellä esitellyt menetelmät muodostavat keskeisen vertailukohdan koneoppimisen hyödyntämiselle sijoituskohteiden tunnistamisessa esiseulontavaiheessa. Menetelmiä vertaillaan taulukon 4.1 avulla, jossa tarkastelu kohdistuu niiden käyttötapaan, suorituskykyyn, tulkittavuuteen, datavaatimuksiin ja aiemmassa tutkimuskirjallisuudessa esitettyihin havaintoihin.

3.4 Mallien arviointi ja selitettävyys

Esiseulontavaiheessa mallien arviointiin liittyy keskeinen kompromissi: kuinka paljon ollaan valmiita hyväksymään virheellisiä positiivisia ennusteita, jotta potentiaalisia menestyjiä ei jäisi tunnistamatta. Tämä kompromissi heijastuu suoraan siihen, millaisia arviointimetriikoita käytetään ja miten mallien suorituskykyä arvioidaan.

Mallien suorituskyvyn arvioinnissa hyödynnetään täsmällisyyttä (engl. *precision*) ja herkkyyttä (engl. *recall*), joita tarkastellaan usein yhtenä kokonaisuutena. Täsmällisyys kertoo, kuinka suuri osa mallin positiivisista ennusteista on oikein, kun taas herkkyys kuvaa, kuinka suuri osa todellisista positiivisista tapauksista tunnistetaan [28]. Nämä mittarit ovat usein ristiriidassa keskenään, sillä herkkyyden parantaminen voi lisätä virheellisiä positiivisia ja siten heikentää täsmällisyyttä, kun taas täsmällisyyden parantaminen voi johtaa positiivisten havaintojen määrän vähenemiseen. Suuret arvot kummassakin mittarissa viittaavat menetelmän korkeaan suorituskykyyn. F1-pisteytys (engl. *F1-score*) yhdistää täsmällisyyden ja herkkyyden yhdeksi arvoksi, mikä helpottaa menetelmien arviointia erityisesti tilanteissa, joissa luokkajakauma on epätasapainoinen. Lisäksi voidaan hyödyntää tarkkuutta (engl. *accuracy*), joka kertoo, kuinka suuren osan luokittelupäätöksistä menetelmä on tehnyt oikein.

Mallien arvioinnissa on myös tärkeää pystyä erottamaan positiiviset ja negatiiviset havainnot toisistaan, jotta voidaan arvioida, kuinka hyvin malli tunnistaa potentiaaliset sijoituskohteet vähemmän lupaavista kohteista. Tähän on käytetty ROC-käyrää (engl. *Receiver Operating Characteristic*) sekä AUC-arvoja (engl. *Area Under the Curve Values*). ROC-käyrä kuvaa herkkyyden ja väärien positiivisten osuuden välistä suhdetta eri kynnysarvoilla, kun taas AUC-arvo tiivistää mallin erottelykyvyn yhdeksi luvuksi [28]. Täsmällisyyden ja herkkyyden välinen tasapaino riippuu usein käytetystä kynnysarvosta, jonka avulla havainto luokitellaan positiiviseksi tai negatiiviseksi.

Vaikka arviointimetriikat mahdollistavat mallien suorituskyvyn arvioinnin, ne eivät kuitenkaan tarjoa tietoa mallien päätöksenteon taustalla olevista tekijöistä. Tämän vuoksi mallien selitettävyys on tärkeää, koska mallien tuottamat ennusteet vaikuttavat suoraan esiseulontatuloksiin. Koneoppimismalleja on usein kritisoitu siitä, että ne toimivat niin sanottuina mustina laatikkoina (engl. *black boxes*), joissa syötteet ja tulokset ovat havaittavissa, mutta niiden sisäinen toimintalogiikka ei ole läpinäkyvä tai helposti tulkittavissa. Tämän haasteen ratkaisemiseksi on kehitetty selitettävän tekoälyn (engl. *Explainable Artificial Intelligence*, XAI) menetelmiä, joiden avulla mallien päätöksentekoa voidaan analysoida ja tulkita [29].

Aiemmat tutkimukset ovat tyypillisesti hyödyntäneet Lundbergin ja Leen esittämää SHAP-menetelmää (engl. *Shapley Additive exPlanations*), jota käytetään koneoppimismallien ennusteiden selittämiseen [30]. Menetelmä auttaa ymmärtämään, miten syötemuuttujat vaikuttavat mallin tekemään yksittäiseen ennusteeseen. Siinä jokaiselle syötemuuttujalle lasketaan Shapley-arvot (engl. *Shapley values*), jotka kuvaavat, kuinka paljon kyseinen syötemuuttuja muuttaa ennustetta suhteessa mallin keskimääräiseen ennusteeseen. On kuitenkin huomioitava, että Shapley-arvot riippuvat käytetystä data-aineistosta ja mallista, joten niiden suora vertailu eri tutkimusten välillä on ongelmallista.

4 Koneoppimisen hyödyt ja rajoitteet

Aiemmissa luvuissa tarkasteltiin sijoituskohteiden tunnistamisen haasteita sekä koneoppimisen hyödyntämistä niiden ratkaisemisessa. Tämän luvun tarkoituksena on tarkastella koneoppimisen hyötyjä ja rajoitteita, jotta voidaan muodostaa kokonaiskuva menetelmien nykyisestä soveltuvuudesta sijoituskohteiden tunnistamiseen. Tarkasteltu kirjallisuus jakautuu empiirisiin tutkimuksiin, jotka käsittelevät sijoituskohteiden tunnistamista ja tekoälyn hyödyntämistä riskipääomasijoittamisessa, sekä menetelmäkeskeisiin tutkimuksiin, joissa vertaillaan koneoppimismenetelmiä eri aineistoilla ja tutkimusasetelmilla. Tutkielmassa jäsennetty kirjallisuus kootaan taulukkoon 4.2.

4.1 Menetelmien vertailu

Taulukossa 4.1 kootaan yhteen tutkielmassa tarkastellut tutkimukset koneoppimisen käyttökohteiden, hyötyjen ja rajoitteiden näkökulmasta. Taulukko havainnollistaa, miten koneoppimista sovelletaan riskipääomasijoittamisen kontekstissa, millaisia mahdollisuuksia menetelmiin liitetään sekä mitkä rajoitteet toistuvat kirjallisuudessa. Vertailu toimii lähtökohtana seuraavalle analyysille, jossa tarkastellaan koneoppimisen keskeisiä käyttökohteita, operatiivisia hyötyjä ja menetelmien hyödyntämistä rajoittavia tekijöitä.

Taulukko 4.1: Menetelmien vertailu

Menetelmä	**Käyttötapa**	**Suorituskyky**	**Tulkittavuus**	**Datavaatimukset (Määrä, laatu, rakenne)**	**Lähteet**
GB	Luokittelu ja pisteytys	Korkea	Matala	Suuri, korkea, strukturoitu	[21, 22, 31, 32, 33]
LR	Luokittelu ja pisteytys	Keskitaso	Hyvä	Vähäinen, keskitaso, strukturoitu	[10, 31, 32, 33, 34, 35]
KNN	Luokittelu	Vaihteleva	Keskitaso	Keskitaso, keskitaso, strukturoitu	[17, 18]
RF	Luokittelu ja pisteytys	Korkea	Vaihteleva	Keskitaso, korkea, osittain strukturoitu	[10, 17, 18, 21, 31, 32, 34]
SVM	Luokittelu	Hyvä	Matala	Keskitaso, korkea, strukturoitu	[21, 31, 32, 33, 34, 35]
XGBoost	Luokittelu ja pisteytys	Korkea	Matala	Suuri, korkea, strukturoitu	[10, 18, 35]

Taulukon 4.1 perusteella tarkastelluissa tutkimuksissa korostuvat erityisesti puupohjaiset ensemble-menetelmät, kuten RF, GB ja XGBoost. Näitä menetelmiä hyödynnetään laajasti sekä luokittelu- että pisteytystehtävissä, mikä kertoo niiden vakiintuneesta asemasta sijoituskohteiden tunnistamisessa. Menetelmät näyttävät saavuttavan monissa tapauksissa parhaat F1-pisteet, mikä tukee käsitystä niiden korkeasta suorituskyvystä.

Samanaikaisesti voidaan havaita kompromissi suorituskyvyn ja tulkittavuuden välillä. Tämä korostuu erityisesti puupohjaisissa ensemble-menetelmissä, joissa ennuste muodostuu useiden yksittäisten päätöspuiden yhdistelmänä. Menetelmät tarjoavat usein korkean ennustetarkkuuden, mutta mallin sisäinen päätöksentekologiikka on vaikeammin tulkittavissa kuin yksinkertaisemmissa klassisissa luokittelumenetelmissä. Zbikowski ja Antosiuk [33] korostavat, että tällaisissa menetelmissä muuttujien tarkkaa vaikutusta yksittäisiin luokittelupäätöksiin on vaikea määrit-

tää. Rajoitetta voidaan osittain lieventää SHAP-menetelmällä, joka mahdollistaa piirteiden vaikutusten tarkastelun yksittäisten ennusteiden tasolla. Ross et al. [18] kuitenkin huomauttavat, että tällaiset menetelmät perustuvat approksimaatioihin, eivätkä siten täysin poista mallien tulkittavuushaasteita.

Vertailumalleina käytetyt klassiset luokittelumenetelmät, kuten LR ja KNN, ovat rakenteeltaan yksinkertaisempia kuin puupohjaiset ensemble-menetelmät, mikä voi parantaa niiden tulkittavuutta. Erityisesti LR:n avulla voidaan arvioida yksittäisten muuttujien yhteyttä mallin ennusteisiin. Näiden menetelmien suorituskyky jää kuitenkin usein ensemble-menetelmiä heikommaksi, mikä voi johtua niiden rajallisemmasta kyvystä mallintaa aineiston monimutkaisia ja epälineaarisia riippuvuuksia. Esimerkiksi LR voi saavuttaa hyvän suorituskyvyn koulutusaineistossa, mutta suoriutua selvästi heikommin testiaineistossa [31, 33]. Tämä voi tarkoittaa, ettei malli opi riittävän yleistettäviä menestystekijöitä, vaan mukautuu liiallisesti koulutusaineiston erityispiirteisiin.

Taulukosta 4.1 käy lisäksi ilmi, että menetelmien suorituskyky ei riipu pelkästään algoritmisista ominaisuuksista, vaan myös käytetyn datan laadusta ja rakenteesta. Puupohjaiset ensemble-menetelmät hyötyvät tyypillisesti monipuolisesta aineistosta, kun taas klassiset luokittelumenetelmät voivat toimia kohtuullisesti myös pienemmillä aineistoilla. On kuitenkin huomattava, että tarkastellut tutkimukset eroavat toisistaan aineistojen, arviointimenetelmien ja kokeellisten asetelmien osalta, mikä vaikeuttaa menetelmien suoraa vertailua tutkimusten välillä.

Edellä esitetyn perusteella menetelmien väliset erot vaikuttavat suoraan niiden soveltuvuuteen riskipääomasijoittamisessa. Korkean suorituskyvyn puupohjaiset ensemble-menetelmät soveltuvat erityisesti sijoituskohteiden suodattamiseen ja ennustamiseen, kun taas paremmin tulkittavat klassiset luokittelumenetelmät tukevat päätöksenteon läpinäkyvyyttä. Näin menetelmän valinta riippuu siitä, painotetaanko ennustetarkkuutta vai päätösten perusteltavuutta sijoitusprosessissa.

4.2 Hyödyt ja rajoitteet

Taulukossa 4.2 jäsennetään tutkielmassa käytettyä kirjallisuutta sen mukaan, millaisesta näkökulmasta tutkimukset käsittelevät koneoppimisen hyödyntämistä sijoituskohteiden tunnistamisessa. Taulukossa erotellaan tutkimusten tyyppi, käyttökohde sekä niissä esiin nousevat hyödyt ja rajoitteet. Tarkoituksena on havainnollistaa, miten aiempi kirjallisuus jakautuu yhtäältä menetelmien tekniseen arviointiin ja toisaalta riskipääomasijoittamisen käytännön kontekstiin. Taulukko auttaa tunnistamaan, mitkä hyödyt ja rajoitteet toistuvat tutkimuksissa useimmin.

Taulukko 4.2: Yhteenveto koneoppimisen käyttökohteista, hyödyistä ja rajoitteista.

	Tutkimustyyppi		**Käyttökohde**			**Hyödyt**			**Rajoitteet**		
Artikkeli	Empiirinen	Menetelmäkeskeinen	Menestyksen ennustaminen	Päätöksenteon tukeminen	Sijoituskohteiden suodattaminen	Tehokkuus	Skaalautuvuus	Signaalien tunnistaminen	Ihmisen merkitys prosessissa	Tulkittavuuden haasteet	Data-aineiston haasteet
Arroyo et al. (2019) [21]		x	x	x	x	x		x		x	x
Di Giannantonio et al. (2022) [12]	x			x	x	x	x	x	x		x
Bai ja Zhao (2021) [34]		x	x	x		x				x	
Kim et al. (2023) [31]		x	x	x		x			x		x
Krishna et al. (2016) [17]		x	x			x					x
Park et al. (2024) [35]		x	x	x	x	x					x
Razaghzadeh Bidgoli et al. (2024) [32]		x	x	x		x		x			
Retterath (2020) [10]	x	x	x	x	x	x	x	x	x	x	
Ross et al. (2021) [18]		x	x	x	x	x	x	x		x	
Röhm et al. (2022) [3]	x			x	x	x	x		x		
Te et al. (2023) [22]		x	x	x	x	x		x		x	
Żbikowski ja Antosiuk (2021) [33]		x	x	x		x			x		x

Taulukon 4.2 perusteella koneoppimisen käyttökohteet painottuvat sijoitusprosessin alkuvaiheeseen, erityisesti menestyksen ennustamiseen, päätöksenteon tukemiseen sekä sijoituskohteiden suodattamiseen. Kirjallisuudessa menestystä mitataan useilla sijoittajan kannalta olennaisilla tavoilla. Yhdessä lähestymistavassa ennustetaan yritysten etenemistä seuraavalle rahoituskierrokselle, yritysostoon tai pörssilistautumiseen [18, 21, 22, 31, 32, 33]. Toisessa lähestymistavassa yrityksiä luokitellaan ennusteiden perusteella onnistumisen ja epäonnistumisen kategorioihin esimerkiksi perustajatiimiä, rahoituskierroksia, toimialaa, sijaintia ja teknologisia ominaisuuksia kuvaavan tiedon perusteella [10, 17, 35]. Lisäksi Bai ja Zhao [34] tarkastelevat suoremmin sijoituspäätöksen ennustamista ja siihen vaikuttavien muuttujien tunnistamista. Näitä lähestymistapoja yhdistää se, että ennusteiden tarkoituksena on tuottaa dataan perustuvaa tietoa yrityksen potentiaalista ja tukea jatkotarkasteluun soveltuvien kohteiden tunnistamista. Tämän vuoksi koneoppimista ehdotetaan tutkimuksissa sekä päätöksenteon tueksi että sijoituskohteiden esiseulonnan välineeksi.

Hyötyjen osalta tutkimuksissa korostuvat erityisesti tehokkuus, skaalautuvuus ja potentiaalisten signaalien tunnistaminen. Nämä hyödyt liittyvät ennen kaikkea hankevirran alustavaan käsittelyyn. Taulukon 4.2 perusteella seitsemän kahdestatoista tutkimuksesta ehdottaa, että koneoppimista voidaan hyödyntää sijoituskohteiden suodattamisessa. Tämä viittaa siihen, että koneoppimisen operatiivinen hyöty syntyy esiseulontavaiheen tehostamisesta, sillä mallien avulla pääomasijoittajien tarkempi analyysi voidaan suunnata lupaavimpiin kohteisiin sen sijaan, että koko hankevirtaa suodatettaisiin manuaalisesti. Menetelmien vertailu 4.1 osoittaa, että puupohjaiset ensemble-menetelmät voivat tarjota tähän ongelmaan käyttökelpoisen ja suorituskykyisen ratkaisun.

Operatiivisen näkökulman lisäksi koneoppiminen voi tarjota myös analyyttistä lisäarvoa. Monet tutkimukset nostavat esiin, että koneoppiminen voi auttaa löytämään aineistosta jatkoarvioinnin kannalta merkityksellisiä signaaleja. Arroyo et al.

[21] esittävät, että malleja voidaan kouluttaa erikseen tiettyjen maiden tai toimialojen aineistoilla, jolloin ne voivat tuottaa kontekstisidonnaisempia havaintoja kiinnostavien sijoituskohteiden tunnistamiseen. Ross et al. [18] puolestaan korostavat, että ennusteiden lisäksi mallit voivat tuottaa piirreanalyysejä, joiden avulla voidaan tunnistaa sekä lupaavia sijoitustekijöitä että mahdollisia riskisignaaleja. Tätä täydentävät Bai ja Zhao [34], joiden mukaan koneoppimista voidaan hyödyntää sijoituspäätöksen kannalta olennaisimpien arviointikriteerien tunnistamisessa. Yhdessä nämä havainnot osoittavat, että menetelmät voivat ohjata sijoittajan huomiota niihin yritystä, tiimiä tai markkinaa koskeviin tekijöihin, jotka ansaitsevat tarkemman selvityksen.

Rajoitteiden osalta taulukko 4.2 osoittaa, että merkittävimmät haasteet liittyvät aineiston laatuun, luokkien epätasapainoon, mallien tulkittavuuteen ja ihmisen tekemän laadullisen arvioinnin merkitykseen. Nämä rajoitteet määrittävät pitkälti sen, missä määrin koneoppimista voidaan hyödyntää käytännön sijoitusprosessissa.

Datan laatuun liittyvät rajoitteet näkyvät erityisesti varhaisen vaiheen yritysten arvioinnissa. Tutkimukset [17, 21, 31, 33, 35] korostavat, että saatavilla oleva aineisto voi olla puutteellista, kohinaista ja vinoutunutta, minkä vuoksi mallit voivat painottua enemmistöluokan havaintoihin. Kim et al. [31] nostavat esiin aineiston epätasapainon, sillä heidän tutkimuksessaan IPO-yrityksiä oli huomattavasti vähemmän kuin ei-IPO-yrityksiä, mikä rajoitti mallin kykyä selittää onnistuneita tapauksia. Samankaltaisesti Park et al. [35] korostavat, että epätasapainoinen koulutusaineisto voi heikentää mallien kykyä tunnistaa vähemmistöluokkaan kuuluvia onnistuneita yrityksiä.

Tulkittavuuden haaste korostaa sitä, ettei mallin tuottama ennuste tai pisteytys yksin riitä päätöksenteon perustaksi. Retterath [10] tuo esiin, että pääomasijoittajat suhtautuvat mallien tuottamiin ennusteisiin varauksellisesti erityisesti silloin, kun niiden perusteita on vaikea ymmärtää. Samankaltaisesti Ross et al. [18] ja Arroyo et

al. [21] nostavat esiin mallien mustan laatikon ongelman, jossa niiden sisäinen päätöksentekologiikka voi jäädä käyttäjälle epäselväksi. Tätä täydentävät Żbikowski ja Antosiuk [33], jotka käsittelevät luokittelupäätösten tulkintaan liittyviä vaikeuksia. Näin selitettävyyteen liittyvät ongelmat näyttäytyvät tarkastelluissa tutkimuksissa yhtenä keskeisenä koneoppimisen hyödyntämistä rajoittavana tekijänä.

Ihmisen tekemän laadullisen arvioinnin merkitykseen liittyy se, että kaikki sijoituspäätöksen kannalta olennaiset tekijät eivät ole helposti mitattavissa tai muutettavissa mallien syötemuuttujiksi. Tämän havainnon ovat tehneet sekä empiiriset tutkimukset [3, 10, 12] että menetelmäkeskeiset tutkimukset [31, 34]. Kim et al. [31] huomauttavat, että perustajatiimin motivaatioon, intohimoon ja idean innovatiivisuuteen liittyviä tekijöitä on vaikea kvantifioida. Digianni et al. [12] vahvistavat tätä näkemystä esittämällä, että perustajatiimien resilienssi nousi haastatteluissa yhdeksi pääomasijoittajien keskeisimmistä arviointikriteereistä. Samalla tutkimus korostaa, että tällaisia yrittäjän todellisiin johtamis- ja sopeutumiskykyihin liittyviä ominaisuuksia on vaikea mitata koneellisten muuttujien avulla [12]. Myös Bai ja Zhao [34] osoittavat, että sijoituspäätösten kannalta ratkaisevat tekijät voivat olla laadullisia, kuten tiimin johtamiseen ja strategiseen suunnitteluun liittyviä arvioita. Tämä on merkittävä rajoite riskipääomasijoittamisessa, koska sijoituspäätökset perustuvat määrällisen datan lisäksi laadulliseen näkemykseen esimerkiksi perustajatiimistä, markkinasta ja liiketoimintamallin uskottavuudesta.

5 Pohdinta

Tässä tutkielmassa tarkasteltu kirjallisuus jakautuu julkaisutyypin perusteella kahteen ryhmään. Menetelmäkeskeinen kirjallisuus painottuu vertaisarvioituihin lehtiartikkeleihin, mikä osoittaa, että koneoppimismenetelmien tekninen arviointi yritysten menestyksen ennustamisessa on tutkimuskirjallisuudessa suhteellisen vakiintunut tutkimussuunta. Sen sijaan empiirinen kirjallisuus painottuu konferenssipapereihin ja esipainettuihin julkaisuihin, mikä viittaa siihen, että ratkaisujen käytännön hyödyntämistä koskeva tutkimus on vielä muotoutumassa. Näin julkaisujakaumasta voidaan päätellä, että tutkimuskenttä on kehittynyt pidemmälle menetelmien teknisessä arvioinnissa kuin niiden käytännön hyödyntämisen tarkastelussa ja päätöksenteollisessa merkityksessä.

Käytännön hyödyntämistä koskevan tutkimuksen keskeneräisyys näkyy myös tutkimusten arviointiasetelmissa. Retterath [10] huomauttaa, että koneoppimispohjaisten seulontatyökalujen käyttöönottoa voi rajoittaa vertailukelpoisen suorituskykynäytön puute suhteessa manuaaliseen sijoittajalähtöiseen arviointiin. Myös tämän tutkielman havainnot tukevat tätä näkemystä. Tarkastelluista menetelmäkeskeisistä tutkimuksista yhdeksän kymmenestä arvioi menetelmien suorituskykyä teknisin mittarein ilman vertailua manuaaliseen sijoittaja-arviointiin. Vaikka tarkastellut tutkimukset esittävät koneoppimisen tehokkaana ratkaisuna sijoituskohteiden tunnistamiseen, tulosten käytännön merkitystä on syytä arvioida kriittisesti. Ei ole täysin selvää, yleistyykö tutkimuksissa havaittu tehokkuus käytännön sijoitusprosesseihin.

Tämä havainto voi osittain selittää pääomasijoittajien varauksellisen suhtautumisen algoritmipohjaisiin menetelmiin.

Menetelmien hyödyntämiseen liittyy myös riskipääomasijoittamisen vakiintuneet toimintatavat. Tutkimusten [3, 10, 12] mukaan pääomasijoittajat ovat jo pitkään tunnistaneet potentiaalisia sijoituskohteita ensisijaisesti manuaalisiin menetelmiin nojautuen, ja nämä käytännöt ovat edelleen vahvasti läsnä toimialalla. Tämä osoittaa, ettei koneoppimispohjaisten ratkaisujen hidas yleistyminen selity yksin teknologian rajoitteilla, vaan myös toimialan vakiintuneilla toimintamalleilla ja päätöksentekokulttuurilla. Koska riskipääomasijoittamisessa korostuvat kokemukseen perustuva arviointi ja vastuu päätösten seurauksista, kynnys siirtää päätösvaltaa algoritmisille järjestelmille on ymmärrettävästi korkea. Näin ollen koneoppimisen rooli näyttäisi ainakin toistaiseksi olevan enemmän päätöksenteon tukiväline kuin ihmisen korvaaja. Tämä viittaa siihen, että koneoppimiseen pohjautuvien työkalujen laajempi käyttöönotto edellyttää teknologisen kehityksen lisäksi myös organisaatioiden toimintatapojen muutosta, jotta organisaatioissa voidaan rakentaa luottamusta uusiin työkaluihin ja sovittaa ne osaksi olemassa olevia sijoitusprosesseja.

Tässä yhteydessä mallien tulkittavuus nousee keskeiseksi kysymykseksi. Tarkasteltu kirjallisuus käsittelee selitettävyyttä kuitenkin melko vähän, mikä viittaa siihen, että selitettävyys jää usein ennustetarkkuuden varjoon. On huomionarvoista, että tutkimukset [10, 18, 21, 22, 33] tunnistavat mallien tulkittavuuden edelleen ongelmaksi sen sijaan, että ne osoittaisivat ongelman ratkenneen. Riskipääomasijoittamisen kontekstissa tämä on ongelmallista, sillä mallien käyttökelpoisuus ei riipu ainoastaan niiden suorituskyvystä, vaan myös siitä, pystyykö sijoittaja ymmärtämään, mihin mallin tuottama arvio perustuu. Algoritmisten työkalujen rooli sijoitusprosesseissa jää väistämättä rajalliseksi, jos niiden tuottamia arvioita ei pystytä avaamaan päätöksentekijälle ymmärrettävällä tavalla.

6 Yhteenveto

Tässä tutkielmassa tarkasteltiin, miten koneoppimista voidaan hyödyntää riskipääomasijoittamisessa sijoituskohteiden tunnistamisessa sekä mitä hyötyjä ja rajoitteita sen käyttöön liittyy. Kirjallisuuden perusteella aihetta koskeva tutkimus keskittyy tällä hetkellä pääosin erilaisten koneoppimismenetelmien testaamiseen vaihtelevilla aineistoilla ja tutkimusasetelmilla. Menetelmien käytännön soveltaminen pääomasijoittajien arjessa näyttää kuitenkin olevan vielä rajallista, mikä voi johtua mallien suorituskyvyn yleistettävyyteen ja ennusteiden perusteiden tulkittavuuteen liittyvistä haasteista.

Tarkasteltujen tutkimusten perusteella koneoppimisen suurin potentiaali liittyy sijoituskohteiden esiseulonnan tehostamiseen ja päätöksenteon tukemiseen. Useissa tutkimuksissa suositellaankin hybridimallia, jossa koneoppimispohjaiset seulontatyökalut suodattavat potentiaalisia sijoituskohteita, minkä jälkeen pääomasijoittajat arvioivat valitut kohteet tarkemmin. Tämä korostaa sitä, että koneoppimista ei tule tarkastella sijoitusammattilaisen korvaajana, vaan työkaluna, joka voi jäsentää hankevirtaa ja kohdistaa sijoittajan huomion jatkoarvioinnin kannalta olennaisimpiin kohteisiin ja arviointitekijöihin. Tutkimuskysymyksiin voidaan vastata seuraavasti:

TK 1: Koneoppimista voidaan hyödyntää erityisesti sijoituskohteiden esiseulonnassa ja päätöksenteon tukena. Mallien avulla voidaan käsitellä laajoja ja monimuotoisia aineistoja, luokitella tai pisteyttää yrityksiä sekä rajata lupaavimmat kohteet jatkoarviointiin. Erityisesti puupohjaiset ensemble-menetelmät soveltuvat tähän tarkoitukseen vahvan ennustekykynsä vuoksi.

TK 2: Koneoppimisen keskeisiä hyötyjä ovat tehokkuus, skaalautuvuus ja päätöksenteon kannalta relevanttien signaalien tunnistaminen. Keskeisiä rajoitteita ovat puolestaan aineistojen puutteellisuus ja epätasapaino, mallien heikko yleistettävyys, tulkittavuuden haasteet sekä se, että sijoituspäätösten kannalta olennaisia laadullisia tekijöitä on vaikea mallintaa kattavasti.

Jatkotutkimuksen kannalta olisi perusteltua keskittyä menetelmien käytännön arviointiin pääomasijoitusyhtiöiden todellisissa sijoitusprosesseissa. Erityisen tärkeää olisi vertailla koneoppimismallien tuottamia arvioita pääomasijoittajien manuaaliseen arviointiin, jotta voitaisiin tarkemmin arvioida, tuottavatko mallit lisäarvoa nykyisiin käytäntöihin verrattuna. Lisäksi tulisi tarkastella, miten hyvin pääomasijoittajat pystyvät tulkitsemaan mallien tuottamia ennusteita ja niiden perusteita. Tämä auttaisi arvioimaan, millaisia selitettävyysratkaisuja ja käyttöönoton edellytyksiä koneoppimispohjaiset työkalut vaativat riskipääomasijoittamisen kontekstissa. Samalla jatkotutkimuksessa olisi tärkeää selvittää, millaisia organisatorisia muutoksia menetelmien laajempi käyttöönotto edellyttää. Tällaisia kysymyksiä ovat esimerkiksi se, miten koneoppimispohjaiset työkalut sovitetaan osaksi olemassa olevia sijoitusprosesseja ja miten sijoittajien luottamusta uusiin työkaluihin voidaan rakentaa.

Lähdeluettelo

[1] V. H. Fried ja R. D. Hisrich, "Toward a Model of Venture Capital Investment Decision Making.", *FM: The Journal of the Financial Management Association*, vol. 23, nro 3, s. 28–37, 1994. DOI: 10.2307/3665619.

[2] P. A. Gompers, W. Gornall, S. N. Kaplan ja I. A. Strebulaev, "How Do Venture Capitalists Make Decisions?", *Journal of Financial Economics*, vol. 135, nro 1, s. 169–190, 2020. DOI: 10.1016/j.jfineco.2019.06.011.

[3] S. Röhm, M. Bick ja M. Boeckle, "The Impact of Artificial Intelligence on the Investment Decision Process in Venture Capital Firms", teoksessa *Artificial Intelligence in HCI*, Springer International Publishing, 2022, s. 420–435. DOI: 10.1007/978-3-031-05643-7_27.

[4] J. A. Timmons ja W. D. Bygrave, "Venture Capital's Role in Financing Innovation for Economic Growth", *Journal of Business Venturing*, vol. 1, nro 2, s. 161–176, 1986. DOI: 10.1016/0883-9026(86)90012-1.

[5] F. Bertoni, M. A. Ferrer ja J. Martí, "The Different Roles Played by Venture Capital and Private Equity Investors on the Investment Activity of Their Portfolio Firms", *Small Business Economics*, vol. 40, nro 3, s. 607–633, 2013. DOI: 10.1007/s11187-011-9384-x.

[6] D. Klonowski, "The Venture Capital Investment Process in Emerging Markets: Evidence from Central and Eastern Europe", *International Journal*

of Emerging Markets, vol. 2, nro 4, s. 361–382, 2007. DOI: 10 . 1108 / 17468800710824518.

[7] W. A. Sahlman, "The Structure and Governance of Venture-Capital Organizations", *Journal of Financial Economics*, vol. 27, nro 2, s. 473–521, 1990. DOI: 10.1016/0304-405X(90)90065-8.

[8] I. Sanz-Prieto, L. de-la-fuente-Valentín ja S. Ríos-Aguilar, "Technical Due Diligence as a Methodology for Assessing Risks in Start-up Ecosystems: An Advanced Approach", *Information Processing & Management*, vol. 58, nro 5, s. 102 617, 2021. DOI: 10.1016/j.ipm.2021.102617.

[9] A. Retterath ja R. Braun, *Benchmarking Venture Capital Databases*, SSRN Scholarly Paper, Rochester, NY, 2020. DOI: 10.2139/ssrn.3706108.

[10] A. Retterath, *Human Versus Computer: Benchmarking Venture Capitalists and Machine Learning Algorithms for Investment Screening*, SSRN Scholarly Paper, Rochester, NY, 2020. DOI: 10.2139/ssrn.3706119.

[11] *Bernstein*, *Korteweg* ja *Laws*, "Attracting Early-Stage Investors: Evidence from a Randomized Field Experiment.", *Journal of Finance (John Wiley & Sons, Inc.)*, vol. 72, nro 2, s. 509–538, 2017. DOI: 10.1111/jofi.12470.

[12] R. Di Giannantonio, M. Murawski ja M. Bick, "The Impact of Machine Learning-Based Techniques on the Scouting and Screening Processes of Early-Stage Venture Capital Firms", teoksessa *The Role of Digital Technologies in Shaping the Post-Pandemic World*, Springer International Publishing, 2022, s. 136–147. DOI: 10.1007/978-3-031-15342-6_11.

[13] G. A. Akerlof, "The Market for "Lemons": Quality Uncertainty and the Market Mechanism", *The Quarterly Journal of Economics*, vol. 84, nro 3, s. 488–500, 1970. DOI: 10.2307/1879431.

[14] L. W. Cong, B. Li ja T. Zhang, *Alternative Data for FinTech and Business Intelligence*, SSRN Scholarly Paper, Rochester, NY, 2019. DOI: 10.2139/ssrn.3521349.

[15] Y. Sun et al., "Alternative Data in Finance and Business: Emerging Applications and Theory Analysis (Review)", *Financial Innovation*, vol. 10, nro 1, s. 127, 2024. DOI: 10.1186/s40854-024-00652-0.

[16] I. H. Sarker, "Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective", *SN Computer Science*, vol. 2, nro 5, s. 377, 2021. DOI: 10.1007/s42979-021-00765-8.

[17] A. Krishna, A. Agrawal ja A. Choudhary, "Predicting the Outcome of Startups: Less Failure, More Success", teoksessa *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, Barcelona, Spain: IEEE, 2016, s. 798–805. DOI: 10.1109/ICDMW.2016.0118.

[18] G. Ross, S. Das, D. Sciro ja H. Raza, "CapitalVX: A Machine Learning Model for Startup Selection and Exit Prediction", *The Journal of Finance and Data Science*, vol. 7, s. 94–114, 2021. DOI: 10.1016/j.jfds.2021.04.001.

[19] K. Maharana, S. Mondal ja B. Nemade, "A Review: Data Pre-Processing and Data Augmentation Techniques", *Global Transitions Proceedings*, International Conference on Intelligent Engineering Approach(ICIEA-2022), vol. 3, nro 1, s. 91–99, 2022. DOI: 10.1016/j.gltp.2022.04.020.

[20] S. Kamm, S. S. Veekati, T. Müller, N. Jazdi ja M. Weyrich, "A Survey on Machine Learning Based Analysis of Heterogeneous Data in Industrial Automation", *Computers in Industry*, vol. 149, s. 103 930, 2023. DOI: 10.1016/j.compind.2023.103930.

[21] J. Arroyo, F. Corea, G. Jimenez-Diaz ja J. A. Recio-Garcia, "Assessment of Machine Learning Performance for Decision Support in Venture Capital In-

vestments", *IEEE Access*, vol. 7, s. 124 233–124 243, 2019. DOI: 10.1109/ACCESS.2019.2938659.

[22] Y.-F. Te, M. Wieland, M. Frey, A. Pyatigorskaya, P. Schiffer ja H. Grabner, "Making It into a Successful Series A Funding: An Analysis of Crunchbase and LinkedIn Data", *The Journal of Finance and Data Science*, vol. 9, s. 100 099, 2023. DOI: 10.1016/j.jfds.2023.100099.

[23] L. Breiman, "Random Forests", *Machine Learning*, vol. 45, nro 1, s. 5–32, 2001. DOI: 10.1023/A:1010933404324.

[24] T. Chen ja C. Guestrin, "XGBoost: A Scalable Tree Boosting System", teoksessa *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, sarja KDD '16, New York, NY, USA: Association for Computing Machinery, 2016, s. 785–794. DOI: 10.1145/2939672.2939785.

[25] C. Cortes ja V. Vapnik, "Support-Vector Networks", *Machine Learning*, vol. 20, nro 3, s. 273–297, 1995. DOI: 10.1007/BF00994018.

[26] G. Guo, H. Wang, D. Bell, Y. Bi ja K. Greer, "KNN Model-Based Approach in Classification", teoksessa *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, Berlin, Heidelberg: Springer, 2003, s. 986–996. DOI: 10.1007/978-3-540-39964-3_62.

[27] S. Dreiseitl ja L. Ohno-Machado, "Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review", *Journal of Biomedical Informatics*, vol. 35, nro 5, s. 352–359, 2002. DOI: 10.1016/S1532-0464(03)00034-0.

[28] O. Rainio, J. Teuho ja R. Klén, "Evaluation Metrics and Statistical Tests for Machine Learning", *Scientific Reports*, vol. 14, nro 1, s. 6086, 2024. DOI: 10.1038/s41598-024-56706-x.

[29] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI", *Information Fusion*, vol. 58, s. 82–115, 2020. DOI: 10.1016/j.inffus.2019.12.012.

[30] S. Lundberg ja S.-I. Lee, *A Unified Approach to Interpreting Model Predictions*, 2017. DOI: 10.48550/arXiv.1705.07874.

[31] J. Kim, H. Kim ja Y. Geum, "How to Succeed in the Market? Predicting Startup Success Using a Machine Learning Approach", *Technological Forecasting and Social Change*, vol. 193, s. 122 614, 2023. DOI: 10.1016/j.techfore.2023.122614.

[32] M. Razaghzadeh Bidgoli, I. Raeesi Vanani ja M. Goodarzi, "Predicting the Success of Startups Using a Machine Learning Approach", *Journal of Innovation and Entrepreneurship*, vol. 13, 2024. DOI: 10.1186/s13731-024-00436-x.

[33] K. Żbikowski ja P. Antosiuk, "A Machine Learning, Bias-Free Approach for Predicting Business Success Using Crunchbase Data", *Information Processing & Management*, vol. 58, nro 4, s. 102 555, 2021. DOI: 10.1016/j.ipm.2021.102555.

[34] S. Bai ja Y. Zhao, "Startup Investment Decision Support: Application of Venture Capital Scorecards Using Machine Learning Approaches", *Systems*, vol. 9, nro 3, s. 55, 2021. DOI: 10.3390/systems9030055.

[35] J. Park, S. Choi ja Y. Feng, "Predicting Startup Success Using Two Bias-Free Machine Learning: Resolving Data Imbalance Using Generative Adversarial Networks", *Journal of Big Data*, vol. 11, nro 1, s. 122, 2024. DOI: 10.1186/s40537-024-00993-8.