TURUN YLIOPISTO
Kauppakorkeakoulu

Data-analytiikan käyttö sisäpiiritiedon väärinkäytön tunnistamisessa

Laskentatoimen ja rahoituksen
kandidaatintutkielma

Laatija:
Konsta Uusimäki

Ohjaaja:
KTT Vesa Partanen

17.4.2025
Turku

Kandidaatintutkielma

Oppiaine: Laskentatoimi ja rahoitus
Tekijä: Konsta Uusimäki
Otsikko: Data-analytiikan käyttö sisäpiiritiedon väärinkäytön tunnistamisessa
Ohjaaja: KTT Vesa Partanen
Sivumäärä: 45 sivua
Päivämäärä: 17.4.2025

Laittomat sisäpiirikaupat ovat olleet sääntelyn kohteena jo 1930-luvun lamasta lähtien, eikä säännöstahti ole viime vuosina ainakaan hidastunut. Sisäpiiririkokset ovat rikosmuotona haasteellisia havaita ja todistaa, koska ne vaativat tarkkaa näyttöä sekä aikajanan kuin myös tuottamuksellisuuden osalta. Viime vuosikymmeninä tietotekniikka ja etenkin data-analytiikka on kehittynyt edistyneeseen nykymuotoonsa, ja tässä kandidaatin tutkielmassa perehdytäänkin siihen, miten perinteisiä data-analytiikan menetelmiä ja koneoppimista hyödynnetään tai voidaan hyödyntää laittomien sisäpiirikauppojen tunnistamisen apuna.

Laittomien sisäpiirikauppojen tunnistaminen edellyttää yhä useammin suuren rahoitusdatamassan analysointia, johon perinteiset tilastolliset menetelmät eivät aina riitä. Koneoppimismenetelmät ja muut edistyneet algoritmit tuovat tähän helpotusta, ja etenkin ohjaamaton ja ohjattu oppiminen tarjoavat keinoja paljastaa pinnanalaisia poikkeamia eli anomalioita kaupankäynnissä. Tutkielmassa tarkastellaan miten näitä keinoja yhdistetään esimerkiksi perinteiseen verkostoanalyysiin, ja tavoitteena on muodostaa kokonaiskuva data-analytiikan mahdollisuuksista ja haasteista modernin markkinavalvonnan osana käsittelemällä esimerkkinä sisäpiiritiedon väärinkäyttöä.

Tutkielman tausta-ajatuksena on ollut pyrkiä monitieteellisyyteen käsittelemällä rahoitustieteellisten teorioiden ohella aiheen yritysjuridista, tilastotieteellistä sekä tietojenkäsittelytieteellistä taustaa. Analyyttiset menetelmät kattavat paljon tilastollisia analyysitekniikkoja kuten tunnuslukuanalyysin, regressioanalyysit sekä koneoppimisen. Tämän ohella erityistä huomiota kohdennetaan dataohjatun päätöksenteon sekä big datan suureen merkitykseen nykyisessä valvontaympäristössä.

Yhtenä tutkielman keskeisenä tavoitteena on pohtia miten valvontaviranomaiset hyödyntävät data-analytiikkaa aikaista enemmän sisäpiirikaupparistkien seulonnassa. Perinteisesti valvonta on pohjautunut ilmoituksiin ja tapauskohtaisiin tarkasteluihin, mutta teknologian kehityttyä on avautunut mahdollisuuksia ennaltaehkäisevämpää valvontaa kohti. Tutkielma perustuu kirjallisuuskatsaukseen ja siinä käsitellään tämän kaiken lisäksi sisäpiirikauppojen oikeudellista kontekstia, taloudellisia markkinatehokkuuteen liittyviä vaikutuksia sekä markkinaluottamusta. Aiheen ajankohtaisuus korostuu nykyisen teknologisen kehityksen nopeuden sekä sääntelyyn liittyvien vaatimusten myötä. Tutkielman tulosten perusteella data-analytiikka ja koneoppimismenetelmät tarjoavat paljon lupaavia keinoja sisäpiiririkosten seulontaan etenkin suurissa datakokonaisuuksissa, joskin menetelmien onnistunut hyödyntäminen edellyttää muun muassa laadukasta dataa ja menetelmien syvällistä ymmärtämistä.

Avainsanat: data-analytiikka, sisäpiirikaupat, tilastolliset menetelmät, dataohjattu päätöksenteko, tekoäly, koneoppiminen, anomaliatunnistus, big data

SISÄLLYS

KUVIOT

TAULUKOT

1 Johdanto

1.1 Johdatus tutkielman aiheeseen

Viimeisten vuosikymmenten aikana tiedon merkitys päätöksenteon kannalta on kasvanut merkittävästi, ja erityisesti data-analytiikka on noussut keskeiseksi työkaluksi monella eri toimialalla ilmiöiden ymmärtämisessä, ennustamisessa ja ohjaamisessa. Finanssisektorilla sen rooli on erityisen vahvasti korostunut, ja käyttö lisääntyy jatkuvasti (Köseoğlu 2022). Data-analytiikka on yleisluonteinen käsite erilaisille analyysitekniikoille ja -menetelmille, joilla tietoaineistoista saadaan esille käyttökelpoisia oivalluksia. Olipa kyseessä sitten yrityksen taloudellinen data tai uuden kehiteltävän lääkkeen vaikutusten arviointi, data-analytiikka auttaa tekemään järkeviä tietoon perustuvia päätöksiä sekä paljastamaan pinnanalaisia trendejä.

Sisäpiirikaupoilla viitataan tässä tutkielmassa sen laittomaan versioon, eli sisäpiiritiedon väärinkäyttöön. Kaikki sisäpiirikaupat eivät ole laittomia, mutta kaikki sisäpiiritiedon väärinkäyttö on laitonta (RL 51:1). Laittomat sisäpiirikaupat ovat liiketapahtumia, joissa pörssiyhtiön sisäiseen käyttöön tarkoitettua tietoa käytetään yksilön oman edun tavoitteluun. Tämä johtaa epätasaiseen informaation jakautumiseen eli asymmetriseen informaatioon eri markkinatoimijoiden välillä, jota pidetään selkeänä esimerkkinä markkinoiden tehottomuudesta. Markkinatehottomuus voi puolestaan johtaa epäoikeudenmukaisiin kilpailuasetelmiin, sijoituspäätösten vääristymiseen ja varallisuuden virheelliseen kohdistumiseen, mikä heikentää luottamusta markkinoihin laajemmin ja siten johtaa laajempiin ongelmiin.

Vaikkakin yhdysvaltalainen lainsäädäntö sisäpiirikaupoista – johon myös muun maailman sääntely perustuu – ulottuu 1900-luvun alkuun saakka, alettiin ilmiötä juridisesti säännellä vasta 1930-luvun laman myötä, ja varsinainen nykyaikainen lainsäädäntö alkoi kehittyä vasta 1960-luvulla. Suomessa sisäpiiritiedon väärinkäytöstä tuli laitonta vuonna 1989, kun arvopaperimarkkinalain ensimmäinen versio astui voimaan (Kurenmaa 2003). Vuonna 2016 voimaantullut Euroopan unionin laajuinen markkinoiden väärinkäyttöasetus eli MAR on puolestaan entisestään laajentanut voimassa olevaa aihealueen lainsäädäntöä.

Tässä kontekstissa data-analytiikalla ja etenkin sen moderneilla tekoälyyn ja koneoppimiseen pohjautuvilla menetelmillä on erityisen tärkeä merkitys, sillä niiden avulla on mahdollista tunnistaa epäilyttäviä sisäpiiritiedon väärinkäyttöön tai muihin arvopaperimarkkinarikoksiin viittaavia transaktioita (Cheng ym. 2022, 3). Perinteisesti laittomia sisäpiirikauppoja on pidetty vaikeana rikostyyppinä

tunnistaa, ja tähän data-analytiikka tuo lisätehokkuutta. Tavoitteena ei ole enää vain rikollisen toiminnan jälkijättöinen havaitseminen, vaan myös valvonnan tehokkaamman kohdistamisen mahdollistaminen.

Tehokkaiden markkinoiden hypoteesin (EMH) mukaan markkinoilla hinnat heijastavat kaikkea saatavilla olevaa informaatiota, eikä yksittäisillä toimijoilla pitäisi olla mahdollisuutta hyötyä epäsymmetrisestä informaatiosta. Tämä ajatus perustuu siihen, että markkinoiden tehokkuus edellyttää, että kaikki tiedot ovat julkisesti saatavilla ja sisältyvät hintaan (Fama 1970). Sisäpiirikauppojen sääntelyä pidetään siten yhtenä pääasiallisena keinona ehkäistä markkinoiden vääristymiä sekä ylläpitää luottamusta markkinoihin, joskin myös esimerkiksi kurssimanipulaatiolainsäädäntö pyrkii ennaltaehkäisemään vääristymiä ja ylläpitämään luottamusta. Sanktiot, kuten sakot ja kaupankäyntikiellot, pyrkivät tekemään sisäpiiritiedon väärinkäytöstä taloudellisesti kannattamatonta sekä vahvistamaan markkinoiden läpinäkyvyyttä.

Aihe on ajankohtainen, koska teknologian kehittyminen mahdollistaa entistä paremmin laajojen rahoitusdatamassojen analysoinnin. Ennen teknologian kehittymistä sisäpiirikaupankäynnin valvomista on siis pidetty vaikeasti toteutettavana käytännössä (Kurenmaa 2003, 30). Tämän ohella markkinoiden toimivuus ja reiluus ovat edelleen tärkeitä kysymyksiä muuttuvassa maailmassamme myös koko kansantalouden ja yhteiskunnan toiminnan kannalta. Tässä tutkielmassa keskitytään erityisesti siihen, miten data-analytiikkaa voidaan soveltaa laittomien sisäpiirikauppojen havaitsemisessa ja millä tavoin analytiikka auttaa markkinoiden läpinäkyvyyden lisäämisessä.

1.2 Tutkielman tavoite ja rajaukset

Tutkielman päämääränä on tutkia miten data-analytiikkaa hyödynnetään ja voidaan hyödyntää laittomien sisäpiirikauppojen tunnistamisessa. Tarkastelun kohteena on erityisen vahvasti se, millaisia menetelmiä käytetään suurien rahoitusdatamassojen analysointiin ja poikkeamien eli anomalioiden tunnistamiseen. Tämän lisäksi tarkentavia alatutkimuskysymyksiä ovat seuraavat:

1. Miten tehokkaita nykyiset data-analytiikan menetelmät ja koneoppimisalgoritmit ovat sisäpiirikauppojen havaitsemisessa?

2. Mitkä data-analytiikan ja koneoppimisen menetelmät soveltuvat parhaiten sisäpiirikauppojen tunnistamiseen?

3. Millaisia haasteita data-analytiikan sisäpiirikauppojen valvontakäytössä on?

Tutkimusasetelma on valittu siten, että se mahdollistaa syvällisen ja laajan tarkastelun data-analytiikan käytöstä sisäpiirikauppojen tunnistamisessa. Kysymykset tehokkuudesta, soveltuvuudesta ja haasteista luovat kokonaisvaltaisen pohjan ilmiön tarkastelulle. Tällaisessa kontekstissa kirjallisuuskatsaus toimii tutkimusmenetelmänä erinomaisesti, koska se mahdollistaa aiempien tutkimusten ja menetelmien tarkastelun. Tutkielman toissijaisena tavoitteena on pyrkiä monitieteellisyyteen, ja tutkielma käsitteleekin laskentatoimen ja rahoituksen lisäksi aihealuetta jossain määrin myös tietojärjestelmätieteen, tietojenkäsittelytieteen, tilastotieteen sekä yritysjuridiikan perspektiiveistä.

Tutkielman aihepiiri on rajattu koskemaan lähinnä data-analytiikan käyttöä sisäpiirikaupoissa ja anomaliatunnistamisessa, mutta myös muita esimerkkejä saatetaan havainnollistamisen vuoksi esitellä. Data-analytiikan käsittelyä syvennetään tarkastelemalla sen roolia ja vaikutusta sisäpiirikauppojen analysoinnissa, erityisesti massadatan ja koneoppimisen näkökulmasta. Sisäpiirikauppoja käsitellään oikeudellisen kontekstin ja rahoitusteoreettisen vaikutuksen kautta. Sisäpiirikauppojen sääntelyn tausta-argumenteista keskitytään markkinoiden tehokkuusargumenttiin sekä oikeudenmukaisuusargumenttiin. Tutkielman tarkoituksena ei ole ottaa kantaa sisäpiirikauppojen laillisuuden tai laittomuuden hyväksyttävyyteen, vaan keskittyä kuvailemaan kuinka ne vaikuttavat markkinoiden läpinäkyvyyteen sekä toimintaan. Tekoälyä on hyödynnetty tutkielmassa aiheen keksimisen ja rakenteen suunnittelun apuna.

2 Datatiede ja data-analytiikka

2.1 Datan, datatieteen ja data-analytiikan määritelmät

Sanan data etymologia juontaa juurensa latinan kieleen ja se on monikkomuoto sanasta datum. Datum on yksittäinen tiedon määräyksikkö, ja data puolestaan viittaa useisiin tiedon määräyksiköihin, vaikkakin sitä käytetään nykykielessä viittaamaan myös yksittäisiin tietoihin (Cambridgen sanakirja: Datum 2025). Taloudellisen yhteistyön ja kehityksen järjestö OECD määrittelee sanan seuraavasti: ”Data on havainnoinnin avulla kerättyjä piirteitä, jotka esitetään yleensä numeerisessa muodossa.” (OECD: Glossary of Statistical Terms 2008).

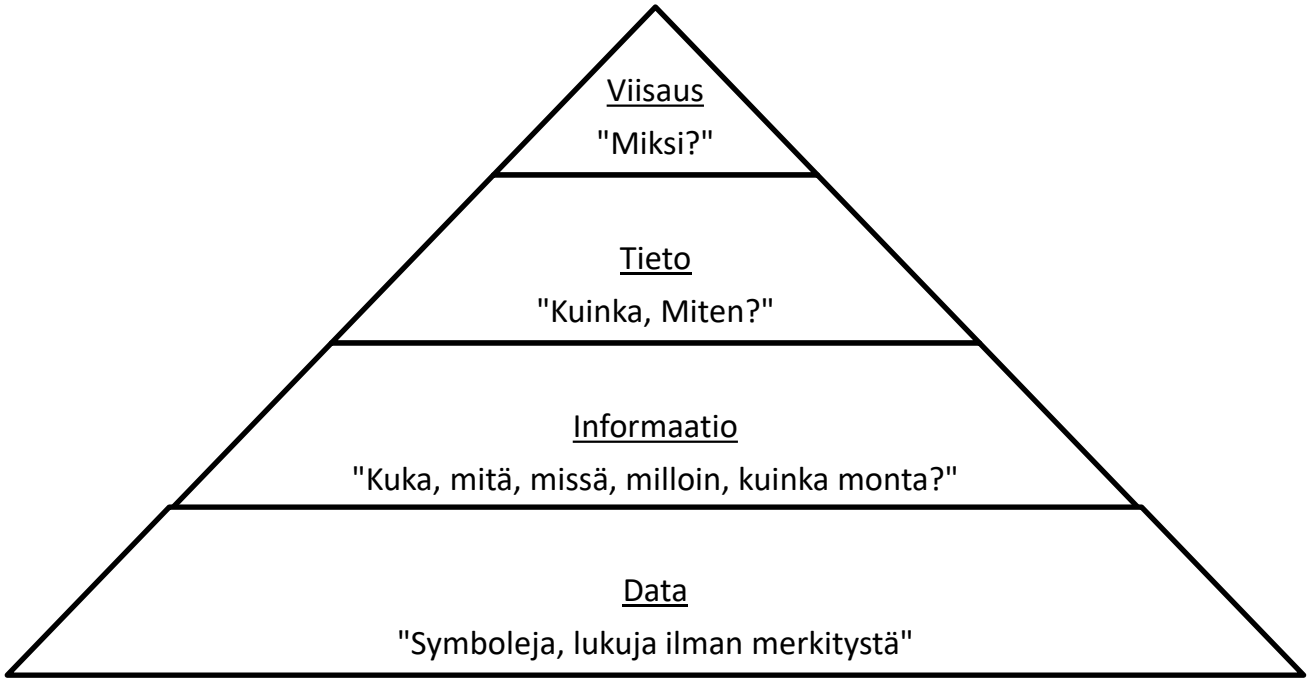
Datan suomenkielinen määritelmä on ongelmallinen, sillä kielessämme sanalla tieto voidaan viitata kontekstista riippuen englanninkielisiin sanoihin data, information tai knowledge. Sana data käännetään suomen kielessä vakiintuneesti aineistoksi tai tietoaineistoksi. Data tarkoittaa yksittäisiä lukuja tai symboleja, joihin ei vielä ole liitetty mitään varsinaista merkitystä. Data on siis yksinään epäinformatiivista (Nurmi & Pyykkönen 2022).

Informaatio (information) tarkoittaa merkityksellistä asiaa, joka voi antaa vastauksia yksinkertaisiin kysymyksiin. Informaatio on yleensä datasta jalostettua ja se on jo itsessään käytettävää. Informaatiolla voidaan vastata muun muassa kysymyksiin: ”Mitä?”, ”Kuka?”, ”Missä?” ja ”Milloin?” (Nurmi & Pyykkönen 2022).

Tieto tai tietämys (knowledge) tarkoittaa puolestaan informaation soveltamista käytäntöön. Tiedolla ymmärretään siis, miten informaatioon reagoidaan. Tieto vastaa näin ollen abstraktimpiin kysymyksiin kuten: ”Kuinka?” ja ”Miten?” (Nurmi & Pyykkönen 2022). Tieto on informaatiota, jota ihminen ymmärtää niin hyvin, että sen perusteella pystytään toimimaan vaadittaessa (Kelleher & Tierney 2018, 56).

Viisaus (wisdom) voidaan määritellä kokonaisvaltaisena ymmärryksenä ”toiminnan perusteista ja kontekstisidonnaisuudesta” eli se vastaa kysymykseen ”Miksi?” (Nurmi & Pyykkönen 2022). Viisauden perusteella tiedetään, kuinka tiedon perusteella kannattaa toimia parhaiten (Kelleher & Tierney 2018, 56).

Teemaa mallinnetaankin akateemisessa kontekstissa usein DIKW-pyramidin kautta, joka kuvaa eri tietotasojen suhdetta toisiinsa, sekä esittää miten ne jalostuvat tasolta toiselle. Nimensä pyramidi saa sanojen Data, Information, Knowledge ja Wisdom lyhenteestä. Suomeksi samasta mallista puhutaan usein viisauden hierarkiana.



Kuvio 1: Viisauden hierarkia pyramidimallina (Nurmi & Pyykkönen 2022)

Datatiede on etenkin tilastotieteen ja tietojenkäsittelytieteen yhteenliittymänä syntynyt monitieteinen tieteenala, joka käsittelee kuinka kirjavasta ja suuresta data-aineistosta poimitaan ongelmanratkaisuun soveltuvaa tietoa (Hayashi 1998). Datatieteessä on huomioitavaa moneen muuhun tieteenalaan verrattuna se, että se yhdistää data-analyysin ja informatiikan menetelmät lähes aina jonkin toisen tieteenalan, kuten esimerkiksi rahoituksen taikka lääketieteen, ongelmien ratkaisemiseksi. Rahoituksen ja datatieteen yhdistelmää voidaan kutsua esimerkiksi rahoitusanalytiikaksi, ja sillä on käyttökohteita finanssialalla esimerkiksi algoritmisen kaupankäynnin suorittamisessa tai tämän tutkielman teeman parissa, eli sisäpiirikauppojen anomaliatunnistamisessa (Zheng ym. 2024, 57-59).

Datatiede on sukua koneoppimiselle, mutta käsite ei ole täysin sama, sillä soveltamismahdollisuudet ovat datatieteessä laajemmat (Kelleher & Tierney 2018, 1). Datatieteilijät käyttävätkin perinteisten data-analytiikan työkalujen ohella koneoppimista ja tekoälysovelluksia ongelmanratkaisuun, ja painotus on erityisesti tulevaisuuden ennustamisessa, eli prediktiivisessä ja preskriptiivisessä analytiikassa (Kelleher & Tierney 2018). Prediktiivinen tarkoittaa suoraan suomennettuna sanaa ennustava ja preskriptiivinen voidaan mieltää esimerkiksi sanaksi ohjaileva. Molemmat ovat siis vahvasti yhteydessä tulevaisuuteen.

Data-analytiikalla, tai pelkällä analytiikalla, ei ole myöskään olemassa yhtä ainoaa tarkkaa määritelmää. Yleisesti voidaan sanoa, että data-analytiikka kattaa erilaisia tilastollisia menetelmiä, tekniikoita sekä lähestymistapoja datan keräämiseen, käsittelyyn, analysointiin ja hyödyntämiseen päätöksenteossa, eli sitä voidaan mahdollisesti pitää suppeampana ja käytännönläheisempänä osana datatiedettä (Aasheim ym. 2015, 104). Joissain konteksteissa data-analyysin synonyyminä käytetään tilastotieteellisiin menetelmiin liittyvää tilastollista analyysiä. Näillä käsitteillä on kuitenkin pieni määritelmällinen ero: tilastotiede on perinteisesti käsitellyt pienempiä tietoaineistoja, kun taas data-analytiikka keskittyy suurien datamäärien eli big datan analysoimisen. Tämän taustalla on se, että aikaisemmin suurien ja sopivien datamassojen kerääminen oli haastavaa ja kallista (Aasheim ym. 2015, 104). Tämä jako ei kuitenkaan ole ideaali, sillä myös tilastotieteen hyödyntämä aineisto voi olla erittäin laaja, ja siten myös koko jaon olemassaoloa voi perustellusti kyseenalaistaa.

Toinen tärkeä taksonomia datatieteen ja data-analytiikan välillä liittyy tekniikoihin ja suuntautuneisuuteen. Datatieteessä katse on pitkälti tulevaisuuden ennustamisessa, kun taas data-analytiikassa pyritään historiallista dataa käyttämällä ymmärtämään trendejä päätöksenteon tukena (Aasheim ym. 2015, 104). Data-analytiikassa työkaluina käytetään muun muassa Microsoft Exceliä, Python-, R- ja JavaScript-ohjelmointikieliä sekä tietokantakieliä kuten SQL:ää. Data-analytiikkaan liitetään usein termi Business Intelligence eli BI, joka on liiketoimintatietoon liittyvä data-analytiikan alalaji. BI:ta käytetään yleensä tiedon raportointiin ja visualisointiin, ja sen tarkoituksena on tehostaa päätöksentekoa.

Taulukko 1: Datatieteen, Data-analytiikan ja Business Intelligencen taksonomia

	Suuntautuminen	**Menetelmiä**
Datatiede	Tulevaisuus	Koneoppiminen, prediktiivinen analytiikka, Python, Pythonin kirjastot kuten Pandas ja PyTorch
Data-analytiikka	Nykyhetki, menneisyys (ja tulevaisuus)	Excel, ohjelmointikielet, BI-työkalut, SQL
Business Intelligence	Nykyhetki ja menneisyys	Excel, BI-työkalut, SQL

Datatieteestä ja data-analytiikasta on huomioitavaa se, että kaikissa tapauksissa analyyttisten metodien käyttäminen ei ole järkevää. Jos trendit huomataan esimerkiksi datan visualisoinneista helposti, ei vaikeiden ja aikaa vievien menetelmien hyödyntäminen ole tehokasta (Kelleher & Tierney 2018, 19). Näitä epäformaaleja ongelmanratkaisumenetelmiä kutsutaan heuristiikoiksi, ja joskus niiden

tulokset voivat riittää tarpeeksi hyvään ymmärrykseen, jolloin täsmällisemmälle ja tutkimuksellisemmalle analytiikalle ei ole tarvetta. Heuristiikat eivät kuitenkaan ole aina toimivia. Esimerkiksi Tversky & Kahneman (1973) huomasivat, että ihmiset arvioivat tapahtumien todennäköisyyttä virheellisesti, koska he nojaavat analyyseissään intuitiivisiin mutta systemaattisesti väärin tuloksiin johtaviin heuristiikkoihin kuten edustavuusheuristiikkaan. Edustavuusheuristiikka on yleensä harhaanjohtava nyrkkisääntö, jonka mukaan ihmiset arvioivat tapahtuman todennäköisyyttä sen perusteella, miten se vastaa jo olemassa olevaa mielikuvaa tarkasteltavana olevasta asiasta. Ihmiset saattavat esimerkiksi tehdä vääriä johtopäätöksiä ihmisen työpaikoista pukeutumisen perusteella: eriskummallisesti ja värikkäästi pukeutuva henkilö tuskin mielletään ensimmäisenä ammatiltaan rahoitusalan ammattilaiseksi.

2.2 Data-analytiikan jaottelu

Data-analytiikan jaotteluun on olemassa useita eri teoreettisia viitekehyksiä. Todennäköisesti tunnetuin näistä on DDPP-malli, jossa analytiikka jaotellaan kuvailevaan, diagnostiseen, ennakoivaan ja ohjaavaan analytiikkaan. Malli on kehittynyt analytiikan ja sen menetelmien kehittymisen myötä, mutta erityisesti sitä käytetään BI-ympäristössä analyysin tasojen jaotteluun.

Kuvaileva eli deskriptiivinen analytiikka on data-analyysin perusmuoto, jossa datamassasta tarkastellaan sen tilastollisia ominaisuuksia, kuten keskiarvoa, mediaania, moodia, hajontaa tai varianssia tilastollisia ohjelmistoja käyttämällä. Deskriptiivisen analytiikan juuret ovat vakaasti tilastotieteelliset, ja tilastotieteessä lähes saman käsitteen kuvaamiseen käytetäänkin käsitettä deskriptiivinen tilastotiede. Deskriptiivinen analytiikka sisältää menetelminään yleiset datan visualisointimenetelmät, ryhmittelyt sekä segmentoinnit, ja vastaa pitkälti kysymykseen: ”Mitä tapahtuu?”. Yksinkertaistetusti voidaan sanoa, että deskriptiivisessä analytiikassa dataa analysoidaan raporttien ja visualisointien tuottamista varten (Köseoğlu 2022, 30).

Diagnostisen analytiikan tarkoituksena on kertoa miksi datasta löytyy trendejä, korrelaatiota ja eritoten syy-seuraussuhteita eri muuttujien välillä, ja se vastaa kysymykseen: ”Miksi näin tapahtui?” (Köseoğlu 2022, 30). Pääasiallisena työkaluna myös diagnostisen analytiikan tekemisessä ovat tilastolliset ohjelmistot. Diagnostista analytiikkaa käytetään muun muassa korrelaatioiden ja syy-seuraussuhteiden selittämiseen hypoteesitestauksen, korrelaatioanalyysin ja kausaalianalyysin keinoin. Lepenioti ym. (2020) pitää diagnostista analytiikkaa kuvailevan analytiikan alalajina.

Prediktiivisessä eli ennustavassa analytiikassa käytetään historiallista dataa ennustamaan tulevaisuuden skenaarioita sekä trendejä, eli se vastaa kysymykseen: ”Mitä todennäköisesti tapahtuu

tulevaisuudessa?" (Köseoğlu 2022, 30). Prediktiivistä analytiikkaa tehdään nykyään pitkälti automaattisesti algoritmipohjaisesti, mutta sitä voidaan toteuttaa myös perinteisin tilastollisin menetelmin, kuten muun muassa lineaariregressiota soveltamalla. Lepenioti ym. (2020) jakaa prediktiivisen analytiikan menetelmät kolmeen eri alalajiin: todennäköisyysmalleihin, tilastolliseen analyysiin sekä koneoppimis/tiedonlouhimis -pohjaisiin menetelmiin. Edellä mainittu lineaariregressio kuuluu tilastollisen analyysin alle, kun taas esimerkiksi Bayes-verkko on todennäköisyysperusteinen ja neuroverkko on koneoppimis/tiedonlouhimis -pohjainen.

Viimeinen mallin osa on preskriptiivinen analytiikka. Preskriptiivinen eli ohjaava analytiikka vastaa pitkälti kysymykseen: "Mitä pitäisi tehdä?", ja on siten analytiikan terävintä ja toteutuksellisesti haastavinta kärkeä (Köseoğlu 2022, 30). Preskriptiivinen analytiikka ei ainoastaan ennusta mitä tapahtuu, vaan suosittelee lisäksi toimenpiteitä halutun tilan saavuttamiseksi. Preskriptiivistä analytiikkaa on vaikea soveltaa perinteisin menetelmin, jolloin mukaan tulee erityisesti koneoppiminen ja tekoäly. Preskriptiivistä analytiikkaa voidaan toki toteuttaa tekoälyn ja koneoppimisen ohella myös esimerkiksi optimointialgoritmeja ja simulaatiomalleja hyödyntämällä. Preskriptiivisen analytiikan menetelmät Lepenioti ym. (2020) jakaa prediktiivisen analytiikan todennäköisyysperusteisten ja koneoppimis/tiedonlouhimis -pohjaisten mallien lisäksi matemaattiseen ohjelmointiin, evolutionääriseen laskentaan, simulointimalleihin sekä logiikkaperusteisiin malleihin.

Taulukko 2: Deskriptiivinen, Diagnostinen, Prediktiivinen ja Preskriptiivinen analytiikka

	Tavoite	**Tekniikat**	**Työkalut**
Deskriptiivinen	Ymmärtää ja kuvailla dataa	Tilastolliset tunnusluvut ja visualisointi	Tilastolliset työkalut (Excel, R, SPSS, Python)
Diagnostinen	Tunnistaa syy-seuraussuhteet	Hypoteesitestaus, kausaalianalyysi	Tilastolliset työkalut
Prediktiivinen	Ennustaa tulevaisuutta	Koneoppiminen, lineaariregressio, muut algoritmit	Tilastolliset työkalut ja etenkin ohjelmointikielet (Python, R)
Preskriptiivinen	Suositella toimenpiteitä halutun tulevaisuuden saavuttamiseksi	Koneoppiminen ja tekoäly, optimointialgoritmit, Monte Carlo -simulaatiot	Ohjelmointikielet, LLM-kielimallit

DDPP-mallin ohella data-analytiikan jaotteluun on olemassa myös muita malleja. Esimerkiksi Bhagattjee (2014) jaottelee data-analytiikan tutkimukselliseen (exploratory), vahvistavaan (confirmatory) sekä ennustavaan (predictive) analytiikkaan. Tutkimuksellinen analytiikka keskittyy aiemmin

tuntemattomien ilmiöiden löytämiseen datasta ilman ennakko-oletuksia ja käyttää apunaan erityisen paljon visualisointeja. Vahvistava analytiikka tarkoittaa puolestaan pitkälti hypoteesitestausta ja se arvioi havaintojen tilastollista merkittävyyttä. Ennustava analytiikka tarkoittaa tässä kontekstissa pitkälti samaa ennustuksellista analytiikkaa kuin aiemmassa DDPP-mallissa.

2.3 Tilastolliset analyysimenetelmät data-analytiikassa

Koska datatieteen ja data-analytiikan pohja on vahvan tilastotieteellinen, on tarpeellista tutustua sen relevantteihin osa-alueisiin hieman syvällisemmin. Vahvaa tilastotieteellistä osaamista voidaan pitää tämän perusteella edellytyksenä data-analytiikan onnistuneelle toteuttamiselle. Tilastollisia analyysimenetelmiä on liian monia tarkkaa ja kattavaa läpikäyntiä varten, joten käsitellään tässä alaluvussa niistä olennaisimmat.

2.3.1 Keskiluvut ja tunnuslukuanalyysi

Ensimmäinen analyysin osa-alue liittyy deskriptiiviseen tilastotieteeseen (Wooldridge 2016, 628). Tähän osa-alueeseen kuuluu muun muassa tilastollinen tunnuslukuanalyysi, jakauman muodon analysointi sekä näiden tilastollinen hypoteesitestaus. Deskriptiivinen tilastotiede on data-analytiikan kannalta kriittinen osa data-analytiikkaa, ja jotkin lähteet pitävät jopa tilastotiedettä tieteellisen data-analyysin synonyyminä (Helping Engineers Learn Mathemathics 36, 1).

Tilastollinen tunnuslukuanalyysi tarkoittaa datasta saatavien keskilukujen ja hajontalukujen määrittämistä sekä niiden avulla olennaisen tiedon esittämistä. Keskilukuihin kuuluu esimerkiksi keskiarvot, mediaani sekä moodi. Keskiarvosta huomioitavaa on se, että kansankielisesti keskiarvosta puhuessa viitataan yleensä sen suosituimpaan lajiin eli aritmeettiseen keskiarvoon. Keskiarvoja on aritmeettisen keskiarvon ohella useita, kuten esimerkiksi painotettu keskiarvo, geometrinen keskiarvo sekä harmoninen keskiarvo (De 2016, 1119). Muille keskiarvomenetelmille löytyy sovelluksia etenkin rahoitustieteen ja taloustieteen parista.

Keskilukujen ohella tunnuslukuanalyysiä tehdään myös niin sanotuista hajontaluvuista, eli esimerkiksi keskihajonnasta ja varianssista, taikka vaihteluvälistä ja kvartaaliväleistä (Helping Engineers Learn Mathemathics 36, 17). Hajontaluvut kuvaavat kuinka paljon havaintojen arvot poikkeavat keskiluvun ympärillä. Varianssi on todennäköisesti merkittävin hajontaluku ja se kuvaa kuinka hajanaista data on. Keskihajonta on varianssille läheistä sukua, mutta eroaa siitä siten, että sitä käytetään

mittarina keskimääräiselle hajonnalle datan alkuperäisessä mittayksikössä, kun taas varianssi ilmaisee keskihajonnan neliöitynä eli eri mittayksikössä.

Vaihteluväli tarkoittaa datan suurimman ja pienimmän arvon välistä erotusta ja se antaa yksinkertaisen käsityksen datan hajonnasta (Upton & Cook 1996, 55). Kvartaaliväli puolestaan tarkoittaa väliä, joka kattaa keskimmäiset 50 % havainnoista ja antaa kuvan datasta ilman äärimmäisiä arvoja. Kvartaaliväli onkin erityisen hyödyllinen työkaluna silloin, kun data sisältää poikkeavia arvoja ja siten vääristää keskihajontaan perustuvia tulkintoja (Upton & Cook 1996, 55). Kvartaalivälin ohella tärkeä matemaattinen konsepti on ala- ja yläkvartiilit. Alakvartiili sisältää pienimmät 25 % havainnoista ja yläkvartiili suurimmat 25 % havainnoista

Jakauman muodon analysointi tarkoittaa menetelmänä pitkälti sen vinouden ja huipukkuuden laskemista ja tutkimista. Jakauman muoto tarkoittaa sitä, miten data on jakautunut keskiluvun ympärille, eli kuinka symmetrinen se on. Vinouden ja huipukkuuden analysointi on tärkeää, koska ne voivat vaikuttaa tilastollisten testien toimintaan haitallisesti (Wooldridge 2016, 658). Monet edistyneemmät tilastolliset menetelmät olettavat, että data on normaalijakautunutta, ja jos data on vinoutunutta tai korkeahuippuista, voi olla järkevämpää käyttää muita testejä ja menetelmiä analyysien tekoon.

Jos jakauma on vino, voidaan histogrammista tai muusta visualisointitavasta huomata datapisteiden suhteeton painottuminen jommallekummalle puolelle häntää (Bowers 1991, 26). Huipukkuus kuvaa jakauman terävyyttä normaalijakaumaan verrattuna. Korkea huipukkuus viittaa siihen, että suurin osa havaintoarvoista on lähellä keskiarvoa, kun taas matala huipukkuus kertoo tasaisemmasta jakaumasta (Bowers 1991, 26).

Tilastollinen hypoteesitestaus on tilastotieteen perusmenetelmä, jolla voidaan testata oletuksia ja yleistää otoksesta huomatut seikat populaatiotasolle (Bowers 1991, 137). Menetelmää käytetään yleisesti kvantitatiivisen tutkimuksen tekemisessä, mutta yhä etenevissä määrin myös liiketoiminnallisten päätöksien perustelussa. Hypoteesitestauksen menetelmiä on datasta ja kontekstista riippuen useita, mutta ehkäpä tärkein niistä on Studentin t-testi. Studentin t-testin avulla verrataan kahden eri ryhmän keskiarvoja ja pyritään siten selvittämään ovatko niissä olevat aineistoerot tilastollisesti merkittäviä (Encyclopedia Britannica: Student's t-test 2025). Muita tunnettuja hypoteesitestausmenetelmiä on muun muassa Khiin neliö -testi sekä ANOVA eli varianssianalyysi.

2.3.2 Regressioanalyysit ja kausaalianalyysi

Toinen tässä tutkielmassa läpikäytävä tilastollisen analyysin osa-alue liittyy regressioon ja kausaalisuuden tunnistamiseen, eli sitä kautta etenkin diagnostiseen ja prediktiiviseen analytiikkaan.

Regressioanalyysissä pyritään selvittämään ”yhden tai useamman muuttujan yhteyttä selitettävään muuttujaan” (Kaakinen & Ellonen 2025). Regressiosta on olemassa erilaisia variaatioita, mutta niistä tunnetuimmat lienevät lineaariregressio ja logistinen regressio.

Lineaarisessa regressioanalyysissä eli lineaariregressiossa oletetaan olevan lineaarinen suhde riippuvan muuttujan ja riippumattoman/riippumattomien muuttujien välillä (Wooldridge 2016, 20). Lineaarista regressiota käytetään laajasti esimerkiksi rahoitusmalleissa ja makrotaloudellisten ennusteiden kehittämisessä ja se loistaakin parhaiten yksinkertaisten ennusteiden luomisessa sekä yleisten trendien huomaamisessa. Lineaariregressio ei kuitenkaan sovellu monimutkaisten ilmiöiden tai klassifikaatiopohjaisen datan analysointiin sen lineaarisen perusolettamuksensa vuoksi. Lineaariregression luontiin on olemassa useita menetelmiä, joista tunnetuin ja käytetyin on pienimmän neliösumman menetelmä (Bowers 1991, 194).

Logistista regressiota käytetään etenkin silloin, kun riippuva muuttuja on binäärisesti kategorinen muuttuja, mutta mallia on mahdollista soveltaa myös muun kategorisen datan käsittelemiseksi (Sperandei 2014). Perinteisesti logistista regressiota käytetään tilanteissa, joissa halutaan mallintaa todennäköisyyttä sille, että tapahtuma toteutuu tai ei toteudu. Logistinen regressiomalli perustuu sigmoidifunktioon, joka muuntaa arvot välille 0–1. Tämä muunnos tekee mahdolliseksi mallin tulkinnan todennäköisyysperusteisesti, eli kertoimien avulla voidaan arvioida kuinka kukin riippumaton muuttuja vaikuttaa tapahtuman todennäköisyyteen. Tämä on mahdollistanut sen, että logistinen regressio toimii perustana kehittyneemmille klassifikaatioalgoritmeille erityisesti koneoppimisen saralla (Thabtah ym. 2019). Logistiselle regressiolle on lineaariregression tavoin lukematon määrä käyttökohteita, mutta ehkäpä merkittävimmät niistä liittyvät anomalia- ja petostunnistamiseen sekä tartuntatautien leviämisen mallintamiseen (Sperandei 2014).

Korrelaatioanalyysi on analyysimenetelmä, joka tutkii kahden tai useamman muuttujan välillä tilastollista yhteyttä tai sen puutetta (Gogtay & Thatte 2017). Yleisimmät korrelaatioanalyysin metodit liittyvät korrelaatiokertoimien laskemiseen, kuten esimerkiksi Pearsonin korrelaatiokertoimeen. Pearsonin korrelaatiokerroin mittaa lineaarista riippuvuutta kahden muuttujan välillä ja voi saada arvon välillä -1 ja 1. Jos korrelaatiokertoimen arvo on lähellä nollaa, ei lineaarista yhteyttä ole. Jos arvo on lähellä jompaakumpaa ääripäätä, on korrelaatio puolestaan todella merkittävä.

Kausaalianalyysi on kehitetty korrelaatioanalyysin jatkoksi, koska pelkkä korrelaatio ei vielä merkitse, että ilmiöiden välille syntyy syy-seuraussuhdetta. Kausaalianalyysi on tutkielman aiheen kannalta merkittävä aihealue etenkin siksi, koska eräs data-analytiikan syvimmistä tarkoituksista on

löytää syy-seuraussuhteita datasta. Korrelaatioanalyysin tavoin myös kausaalianalyysiin on olemassa useita eri menetelmiä, ja sopivin niistä määrittyy aina aineiston ja tutkimustavoitteen perusteella.

2.3.3 Aikasarja-analyysi ja ennustemallit

Aikasarja-analyysi on analyysin muoto, jota käytetään, jos data on asetettu aikajärjestykseen ja datapisteissä on arvoeroja. Aikasarja-analyysi eroaa siten olennaisesti aikaisemmin esitellyistä poikittaistutkimuksellisista menetelmistä (Wooldridge 2016, 312). Aikasarja-analyysiä voidaan hyödyntää ajassa muuttuvan datan mallintamiseen ja tutkimiseen, ja sen tarkoituksena on tunnistaa esimerkiksi trendejä, syklisyyttä tai muita rakenteita datasta. Aikasarja-analyysi on luonteeltaan prediktiivistä (Köseoğlu 2022, 30-31).

Aikasarja-analyysin tekemiseen on olemassa monia menetelmiä ja ne voidaan luokitella taajuustasoanalyyseihin sekä aikatasoanalyysiin. Taajuustasoanalyysiin perustuvissa metodeissa pyritään analysoimaan tapahtumien määrää eli frekvenssiä, kun taas aikatasoanalyyseissä huomio on ajallisesti peräkkäisissä havainnoissa ja niiden välisissä riippuvuuksissa. Menetelmiä on luonnollisesti paljon, joten käsitellään aikasarja-analyysin tunnetuimmat ennustemallit, Autoregressive Integrated Moving Average (ARIMA) sekä Generalized Autoregressive Conditional Heteroskedasticity (GARCH).

ARIMA hyödyntää aikasarjadataa tunnistaakseen ilmiöitä sekä ennustaakseen tulevaisuuden trendejä. Esimerkiksi rahoitustieteessä sitä voidaan käyttää osakkeiden tulevien hintojen ennustamiseen (Ho ym. 2021). ARIMA koostuu kolmesta pääkomponentista: autoregressiosta, integraatiosta sekä liukuvasta keskiarvosta. Autoregressio tarkoittaa yksinkertaistetusti sitä, että menneisyys vaikuttaa nykyhetkeen. Integraatio tarkoittaa mallissa sitä, että aikasarjasta pyritään tekemään stationaarinen eli ajassa muuttumaton siten, että arvot kuvaavat eroja datapisteiden välillä. Liukuva keskiarvo eli Moving Average kuvaa tietyn ajanjakson keskiarvoa ja sitä päivitetään jatkuvasti uusien datapisteiden myötä. Se auttaa mallissa tasoittamaan vaihtelua ja poistamaan heilahduksia, jolloin trendit ja datan rakenne on helpompi havaita.

GARCH on erityisesti rahoitusmarkkinoilla käytetty malli, jota käytetään ennustamaan volatiliteetin kehittymistä aikasarjadataan perustuen (Engle, 2001). GARCH koostuu kahdesta pääelementistä. Autoregressiivinen osa tarkoittaa tässä mallissa sitä, että aikaisemmat volatiliteettiarvot vaikuttavat myös nykyiseen volatiliteettiin. Ehdollinen heteroskedastisuus tarkoittaa puolestaan sitä, että mallissa otetaan huomioon aiemmat havaintovirheet. Ominaisuus volatiliteetin tunnistamiseksi tekeekin mallista erityisen hyödyllisen riskienhallinnassa, optiohinnoittelussa ja sijoituspäätösten tukena (Engle, 2001).

Yhteenvetona voidaan todeta, että tilastolliset analyysimenetelmät muodostavat merkittävän osan data-analytiikan tieteellisestä perustasta. Tilastollinen lähestymistapa mahdollistaa rationaalisten ja systemaattisten johtopäätösten tekemisen. Kaikki tässä luvussa käsitellyt menetelmät tukevat kaikkia analytiikan muotoja deskriptiivisestä prediktiiviseen analytiikkaan, joskin preskriptiivisen analytiikan ennustaminen vaatii myös vielä entisestään kehittyneempiä malleja. Tilastollinen ymmärrys luo siis välttämättömän pohjan data-analytiikan syvemmälle ymmärtämiselle.

3 Big data ja tekoäly

3.1 Dataohjattu päätöksenteko ja big data

Datan ja data-analytiikan merkitys on noussut viimeisen muutaman vuosikymmenen aikana dataohjatun päätöksenteon suosion noustessa ja datan varastoimiseen liittyvien hintojen pudotessa (Aasheim ym. 2015, 104). Dataohjattu päätöksenteko (data-driven decision-making) painottaa datan ja sen analysoimisen merkitystä liiketoimintapäätöksiä tehtäessä. Päätökset eivät siten nojaa vain intuitioon tai heuristiikkoihin (Provost & Fawcett 2013, 3). Käytännössä tämä tapahtuu esimerkiksi BI-järjestelmiä hyödyntämällä, ja visualisointisovellukset kuten Power BI ja Tableau ovatkin saavuttaneet suuren suosion yritystoiminnan ohjaamisessa. Suuret määrät tarkkaa dataa esimerkiksi myynneistä, kuluista ja mainonnasta mahdollistavat niiden hyödyntämisen laaja-alaisesti liiketoimintapäätösten tukena.

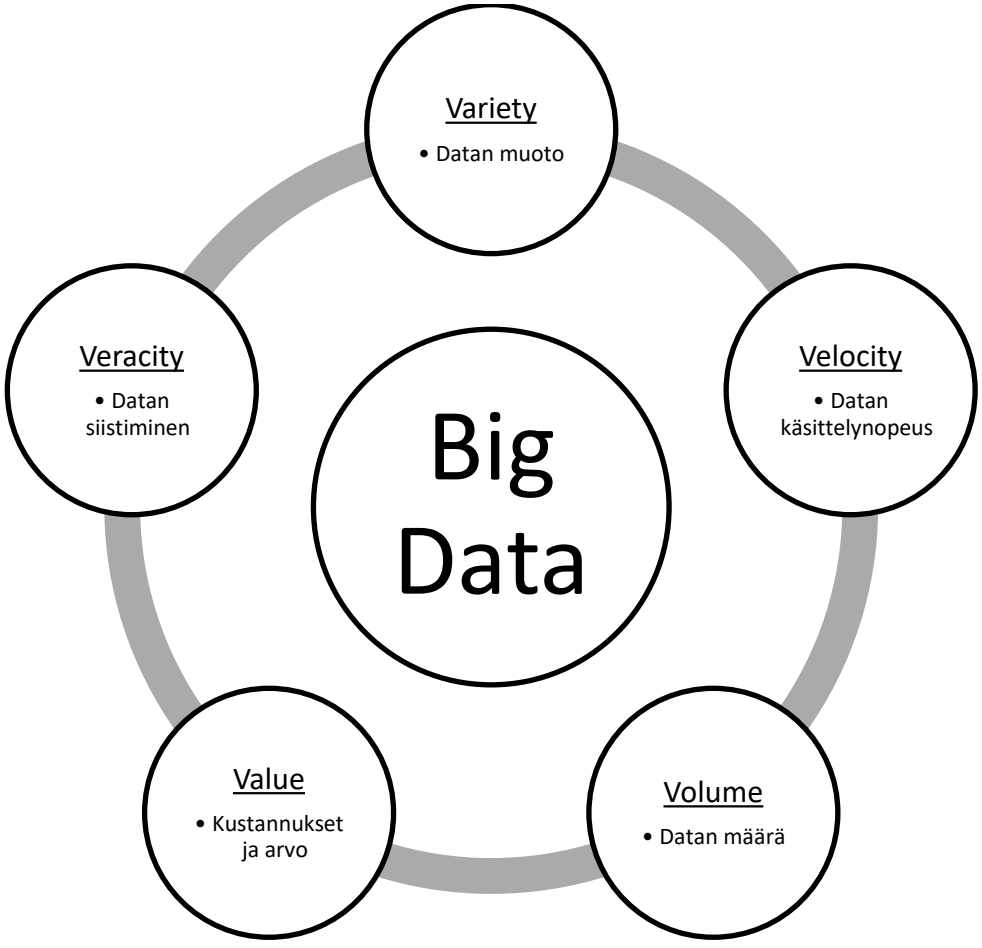
Tästä suuresta datamäärästä ja sen hyödyntämisestä käytetään termejä big data tai massadata. Ihmiset tuottavat nykyisessä tietotekniikkaperusteisessa maailmassamme päivittäin noin viisi eksatavua eli noin miljardi gigatavua erimuotoista dataa, kuten videoita, kuvia ja tekstiä. Tämä on sama määrä, mitä ihmisten arvioidaan tuottaneen kirjoitustaidon keksimisestä lähtien vuoteen 2003 asti, eli datan tuotantovauhti on kiihtynyt (Kelleher & Tierney 2018, 9). Koska monet perinteiset data-analytiikkatyökalut ovat osoittautuneet liian jäykiksi, tehottomiksi tai hitaiksi näin suurien data-aineistojen käsittelyyn, on sitä varten kehitetty uusia ja tehokkaampia teknologioita. Esimerkiksi pilvipohjaisten analytiikkapalveluiden avulla voidaan analysoida ja visualisoida suuria datamassoja nopeammin. Tämä kehitys on tehnyt dataohjatusta päätöksenteosta nopeampaa ja tehokkaampaa.

Big datalle on olemassa useampia teoreettisia malleja, mutta tunnetuimmat niistä ovat niin kutsuttuja V-malleja. Tietoteknologian tutkimus- ja konsultointiyritys Gartnerin analyytikko Doug Laney kehitti alkuperäisen 3V-mallinsa kuvailemaan big datan ulottuvuuksia vuonna 2001, joskin mallia on väärinymmärretty. Alkuperäisessä mallissa 3V tarkoitti vain yhtä osaa big datan määritelmästä (Forbes 2013). Myöhemmin mallia on laajennettu muun muassa 4V-, 5V-, 6V-, 7V- ja jopa 12V-malliin. Alkuperäinen 3V-malli sisältää kuitenkin vain kolme ulottuvuutta: Volume, Velocity ja Variety.

Volume (määrä) viittaa mallissa datan suureen määrään, jota nykyaikaiset sensorit pystyvät keräämään ja tietokannat sisältämään. Velocity (nopeus) viittaa datan prosessointinopeuteen, jonka nykyaikainen tietotekniikka on mahdollistanut (Kelleher & Tierney 2018, 9), ja Variety (monipuolisuus) puolestaan viittaa datan erilaisiin muotoihin, eli strukturoituun, strukturoimattomaan sekä semi-strukturoituun dataan (Kelleher & Tierney 2018, 9).

Strukturoitu data on selkeästi järjesteltyä ja tallennettu esimerkiksi relaatiotietokantoihin taulukkomuodossa. Strukturoimaton data ei puolestaan noudata tätä rakennetta, vaan voi koostua tekstin lisäksi esimerkiksi videoista, kuvista taikka metadatasta. Semi-strukturoitu on näiden kahden välimuoto, eli jotain rakenteellisia elementtejä on, mutta se ei ole täysin järjesteltyä tai eheää.

3V-mallia on laajennettu ajan saatossa sisältämään enemmän ulottuvuuksia. Nykyään käytetyin V-malli lienee 5V-malli. 5V-malli sisältää aikaisemmin mainittujen Volumen, Velocityn ja Varietyn lisäksi tasot Value (arvo) ja Veracity (todenperäisyys) (Lomotey & Deters 2014, 181). Value viittaa mallissa siihen, että big dataan liittyy varastointikustannuksia, mutta myös kvantifioitavaa taloudellista arvoa. Veracity-taso viittaa puolestaan siihen, että data saattaa sisältää ”saastetta”, joka pitää siistiä pois ennen sen hyödyntämistä (Lomotey & Deters 2014, 181).



Kuvio 2: 5V-malli (Lomotey & Deters 2014, 181)

3.2 Tekoäly ja koneoppimismenetelmät data-analytiikassa

Tekoälyn ja etenkin laajojen kielimallien esiinmarssi 2020-luvun alkupuoliskolla neljännen teollisen vallankumouksen yhteydessä on mahdollistanut niiden hyödyntämisen myös data-analytiikassa (Sarker, 2021). Tekoäly tai AI tarkoittaa yksinkertaisimmillaan menetelmien joukkoa, joilla pyritään matkimaan inhimillistä älykkyyttä (Hamet & Tremblay 2017).

Tekoälyn tärkeimmät osa-alueet ovat koneoppiminen, neuroverkot sekä syväoppiminen. Huomioitavaa on kuitenkin se, että tämä ei ole kattava listaus tekoälymenetelmistä, sillä jokainen näistä pitää sisällään erilaisia alamenetelmiä ja algoritmeja.

Koneoppiminen mielletään yleensä tekoälyn osa-alueeksi, joka kattaa laajan määrän erilaisia algoritmeja, joilla järjestelmät oppivat löytämään syy-seuraussuhteita ja rakenteita datasta (Kelleher & Tierney 2018, 97). Koneoppiminen jaetaan yleensä neljään eri paradigmaan, jotka ovat ohjattu oppiminen, ohjaamaton oppiminen, osittain ohjattu oppiminen ja vahvistusoppiminen. Suurin osa koneoppimisalgoritmeista kuuluu joko ohjattuun oppimiseen tai ohjaamattomaan oppimiseen (Kelleher & Tierney 2018, 99).

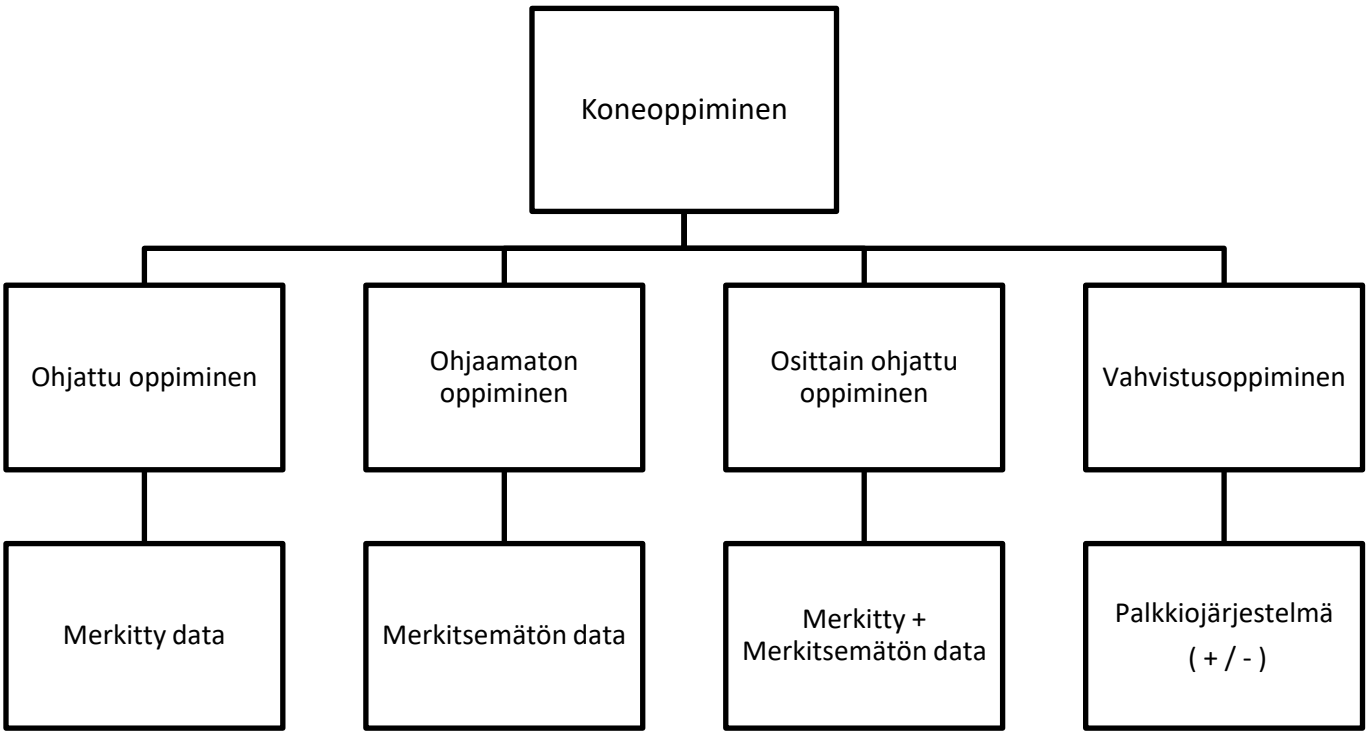
Ohjatussa oppimisessa järjestelmä oppii valmiiden vastauksien kautta. Syötedatalle määrätään ennalta oikea vastaus, jonka jälkeen järjestelmä rakentaa koulutusdatan pohjalta tuloksia ennustavan mallin. Tästä datamallista käytetään termiä merkitty data. Aiemmin läpikäydyistä tilastoista muun muassa lineaariregressio ja logistinen regressio kuuluvat koneoppimiskontekstissa ohjatun oppimisen alle.

Ohjaamaton oppiminen tarkoittaa sitä, että järjestelmä löytää itse rakenteita datasta. Ohjaamaton oppiminen eroaa ohjatusta oppimisesta siten, että syötedatalle ei ole ennalta määrättyä oikeaa vastausta, eli data on muodoltaan merkitsemätöntä. Tämä johtaa siihen, että järjestelmä huomaa rakenteet paremmin kuin ohjatussa oppimisessa, mutta tietyn ongelman ratkaiseminen vaikeutuu (Kelleher & Tierney 2018, 102).

Osittain ohjattu oppiminen on näiden kahden edeltä mainitun yhdistelmä, jossa järjestelmä oppii sekä ohjattujen ja ohjaamattomien menetelmien avulla. Järjestelmä voi siis ottaa vastaan sekä oikeaksi merkittyä dataa tai vaihtoehtoisesti strukturoimatonta dataa.

Vahvistusoppiminen puolestaan perustuu niin sanottuun ”yritykseen ja erehdykseen”. Vahvistusoppimisessa järjestelmää palkitaan oikean vastauksen saavuttamisesta ja rangaistaan väärästä

vastauksesta. Vahvistusoppimisen ideana on siis se, että järjestelmä oppii luontaisesti parhaan ”reitin” toistojen lisääntyessä.



Kuva 3: Koneoppimisparadigmat (Sarker 2021)

Nykyaikaisen massadatapohjaisen tekoälyn on mahdollistanut koneoppimismenetelmien ohella neuroverkkoteknologia. Neuroverkko on biologiasta ja ihmisen hermoston toiminnasta inspiraatiota saanut malli matematiikassa sekä tietojenkäsittelyssä, joka jäljittelee ihmisaivojen rakennetta ja toimintaa. Neuroverkko koostuu useasta kerroksesta, joissa keinotekoiset hermosolut, eli neuronit, käsittelevät dataa ja siirtävät sen eteenpäin seuraaviin kerroksiin lisäkäsittelyä varten (Kelleher & Tierney 2018). Tämä mahdollistaa syvällisten ja vaikeiden ongelmien ratkaisemisen.

Neuroverkkojen ja koneoppimisen yhdistelmäalalajia kutsutaan syväoppimiseksi eli Deep Learningiksi. Syväoppiminen on 2010-luvun alussa tapahtunut teknologinen innovaatio, joka on mullistanut tekoälyn erityisesti suurien datamäärien käsittelyssä. Syväoppiminen on menetelmänä erityisen tehokas, koska sen avulla voidaan tunnistaa datasta rakenteita ilman, että ihminen merkitsee dataa.

4 Sisäpiirikaupat

4.1 Sisäpiirikauppojen määritelmä ja sääntely

Laittomalla sisäpiirikaupalla, eli sisäpiiritiedon väärinkäyttämisellä rahoitusvälinekaupassa, on käytetystä lähteestä ja maan juridisesta kontekstista riippuen erilaisia tarkkoja määritelmiä. Suomessa sisäpiiritiedon väärinkäyttämisestä säädetään arvopaperimarkkinalaissa, jossa sitä säännellään yhdessä sisäpiiritiedon väärinilmaisun, markkinoiden manipuloinnin ja arvopaperimarkkinoiden tiedottamisrikoksen kanssa (AML 18:2). Maailmassa on olemassa 103 rahoitusmarkkinapaikkaa, joista 87:ssä on olemassa jonkin tasoista sisäpiirikauppasääntelyä (Bhattacharya & Daouk 2002).

Sisäpiirikaupan määritelmään liittyy olennaisesti sisäpiiritieto. Suomen rahoitusmarkkinoita valvovan viranomaisen, eli Finanssivalvonnan, mukaan sisäpiiritiedolla tarkoitetaan suurelle yleisölle saavuttamatonta tietoa, joka julkistettuna todennäköisesti vaikuttaa huomattavasti siihen liittyvän rahoitusvälineen hintaan (Finanssivalvonta 2018). Sisäpiiritiedosta käytetään joskus myös käsitettä yksityinen informaatio, ja konkreettisesti tämä tieto voi liittyä esimerkiksi ”yrityksen tulontuottokykyyn, tuleviin kassavirtoihin ja investointimahdollisuuksiin” (Niskanen & Niskanen 2013, 290).

Ajantasainen suomalainen lainsäädäntö sisäpiirikaupoista perustuu arvopaperimarkkinalain ohella pääasiassa EU:n vuoden 2014 markkinoiden väärinkäyttöasetukseen MAR:iin (Market Abuse Regulation) ja sen määritelmään sisäpiiritiedosta. Tästä huolimatta sisäpiirikaupat ovat olleet regulaation kohteena maailmanlaajuisesti jo ainakin 1930-luvun alkupuoliskolta lähtien (Perino 2018). MAR:n mukaan sisäpiiritiedolle on ominaista se, että se on riittävän tarkkaa johtopäätösten tekoon rahoitusvälineen markkina-arvosta sekä se, että se on tapahtumasidonnaista, eli se kertoo asioista, jotka ovat jo tapahtuneet tai tulevat tapahtumaan. Näistä kahdesta ominaisuudesta käytetään yhdessä termejä sisäpiiritiedon täsmällisyys ja olennaisuus (Finanssivalvonta 2018). Sisäpiiritiedolle on lisäksi tunnusomaista se, että sisäpiiritietoa omaava, eli sisäpiiriläinen, saa ilmaista tietoaan ulkopuolisille vain silloin kun se on tarpeellista esimerkiksi ammatin suorittamisen kannalta.

Sisäpiiriläisellä tarkoitetaan henkilöä, joka on merkitty julkisen osakeyhtiön sisäpiiriluetteloon. Sisäpiiriluettelo sisältää ne henkilöt, jotka pääsevät käsiksi sisäpiiritietoon ja/tai työskentelevät liikkeeseenlaskijalle (Finanssivalvonta 2018). Sisäpiiriluetteloita on kahdenlaisia: hankekohtaisia ja pysyviä. Pysyvä sisäpiiriluettelo sisältää kaikki henkilöt ja työntekijät, joilla on jatkuva pääsy kaikkeen sisäpiiritietoon. Tähän listaan voi kuulua esimerkiksi yhtiön hallitus, toimitusjohtaja, talousjohtaja ja muut johtajat. Hankekohtainen sisäpiiriluettelo sisältää puolestaan kaikki henkilöt, joilla on sisäpiiritietoa jonkin hankkeen tiimoilta. Esimerkiksi ulkoiset konsultit ja tilintarkastajat merkitään usein

näihin listoihin (NASDAQ: Pörssin sisäpiiriohje 2020, 5). Finanssivalvonnan tehtävänä onkin valvoa näiden sisäpiiriluettelolaisten ja heidän läheistensä kaupankäyntiä, jotta lain rajojen sisällä pysytään kauppoja tehdessä.

Suomessa Finanssivalvonta on laatinut 10 toimintaohjetta, joilla sisäpiiriläinen voi käydä kauppaa rahoitusvälineillä rikkomatta markkinalainsäädäntöä. Monet ohjeista on luonteeltaan intuitiivisia ja ymmärrettäviä, kuten esimerkiksi se, että optiokaupan kanssa tulee olla erityisen varovainen sekä se, että yhtiön sisäpiirilistoista vastaavalta kannattaa kysyä ohjausta, jos suunnitelmissa on yhtiön rahoitusvälineiden vaihdanta (Finanssivalvonta 2018).

Taulukko 3: Sisäpiiriläisen 10 kaupankäyntiohjetta (Finanssivalvonta 2018)

1. Pyri tekemään pitkäaikaisia sijoituksia
2. Voit käyttää kaupankäyntiohjelmia
3. Harkitse tarvetta rajata omaisuudenhoitosopimuksen ulkopuolelle yhtiösi rahoitusvälineet, jos sinulla tai lähipiirilläsi on sellainen
4. Tee liiketoimet muulloin suljetun ajanjakson ulkopuolella
5. Ajoita kaupankäyntisi tulosjulkistuksen jälkeiseen ajankohtaan
6. Varmista yhtiön sisäpiirivastaavalta, onko kaupankäynnillesi mahdollista sisäpiiriestettä
7. Kannustinjärjestelmään liittyvien optioiden vastaanottaminen ja niiden merkitseminen on lähtökohtaisesti mahdollista
8. Älä myy tai osta optioita, jos hallussasi on sisäpiiritietoa
9. Rahoitusvälineiden ostaminen on mahdollista, jos on objektiivisesti perusteltua olettaa, että hallussasi oleva sisäpiirikielto on sen hinnan kannalta selkeästi kielteistä
10. Rahoitusvälineiden ostaminen, myyminen ja merkitseminen on mahdollista, jos tiedät liiketoimen toisella osapuolella olevan hallussaan sama sisäpiiritieto kuin sinulla

Sisäpiiriläisenä näistä ohjeista tärkein lienee se, että kaupankäynnin tulee tapahtua suljetun jakson ulkopuolella. Suljettu ajanjakso tarkoittaa 30 päivän aikaikkunaa ennen osavuosikatsauksen, listayhtiön taloudellisen raportin tai tilinpäätöksen julkistamista, ja tuona aikana yhtiön hallituksen jäsen, toimitusjohtaja tai muu johtotehtävissä toimiva ei saa itse käydä kauppaa tai suositella muita tekemään kauppaa kyseenomaisilla rahoitusvälineillä (NASDAQ: Pörssin sisäpiiriohje 2020, 5). Tämän lisäksi kyseisen ”taloudellisen raportin valmisteluun osallistuvan henkilön ei ole suositeltavaa tehdä liiketoimia” suljetun ikkunan aikana, vaikkakaan täyttä kieltoa ei ole (NASDAQ: Pörssin sisäpiiriohje 2020, 6). Huomioitavaa on kuitenkin se, että kielto ei ole täysin absoluuttinen, vaan siitä voidaan joustaa MAR-asetuksen asettamien edellytyksien täyttyessä. Eräs esimerkki tästä on se, jos ”vakavat rahoitusvaikeudet edellyttävät osakkeiden myyntiä” (MAR 19:12).

Toinen sisäpiiriläiselle tärkeä ohje liittyy sallituista sisäpiirikaupoista ilmoittamiseen. Sisäpiiriläisen tai hänen lähipiiriinsä kuuluvan tulee ilmoittaa Finanssivalvonnalle ja yhtiölle kolmen arkipäivän kuluessa kaupankäyntitapahtuman toteutuksesta (NASDAQ: Pörssin sisäpiiriohje 2020, 6).

Yleisenä ohjenuorana sisäpiiriläiselle voidaan lisäksi sanoa, että askarruttavissa tilanteissa tulisi ottaa yhteyttä yhtiön sisäpiirivastaavaan. Sisäpiirivastaava tarkoittaa pörssilistatuissa yhtiöissä henkilöä, joka valvoo sisäpiirihallintoon kuuluvien tehtävien suorittamista (Nasdaq Helsinki, 2021). Tämän kaiken ohella Finanssivalvonnan ohjeistuksessa kerrotaan muun muassa, että sisäpiiritietoon liittyvää rahoitusvälinettä voidaan ostaa, myydä tai merkitä, jos sama sisäpiiritieto on hallussa myös vaihdannan vastapuolella.

Sisäpiiritiedon tahallinen tai törkeästä huolimattomuudesta johtuva väärinkäyttö on Suomessa rangaistavaa rikoslain 51 luvun 1 pykälän nojalla. Törkeästä tekomuodosta säädetään samaisen luvun 2 pykälässä (RL 51:1–2). Sisäpiiritiedon väärinkäyttämisestä voidaan tuomita sakkoon tai vankeuteen enintään kahdeksi vuodeksi, ja törkeän tekomuodon osalta rangaistus on vähintään neljä kuukautta ja enintään neljä vuotta vankeutta. Törkeä tekomuoto tulee kyseeseen, kun henkilö väärinkäyttää sisäpiiritietoa tahallisesti tavoitellakseen erityisen suurta hyötyä tai huomattavaa henkilökohtaista etua, taikka käyttää rikoksen tekemisessä hyväksi erityisen vastuullista asemaansa laissa määritellyissä yhteisöissä, taikka rikos tehdään erityisen suunnitelmallisesti, ja sisäpiirintiedon väärinkäyttö on myös kokonaisuutena arvostellen törkeä. Myös väärinkäytön yritys on rangaistavaa.

Sisäpiiritiedon väärinkäyttämisestä voi kiinni jäädessä seurata rikosoikeudellisten seurauksien ohella hallinnollisia seurauksia. Markkinoiden väärinkäyttöasetus velvoittaa jäsenvaltiot säätämään valtuudesta toteuttaa hallinnollisia seurauksia. MAR:iin perustuen oikeushenkilölle, kuten vaikkapa yritykselle tai yhdistykselle, voidaan määrätä kokonaisliikevaihdosta mitattuna enintään 15 % suuruinen tai 15 000 000 € kokoinen hallinnollinen seurausmaksu. Luonnollisille henkilöille samainen seuraus on enintään 5 000 000 €. Tämän lisäksi voidaan määrätä erilaisia lisäseuraamuksia, kuten toimiluvan peruuttaminen tai kielto toimia johtotehtävissä, mutta mahdolliset seuraamukset eivät rajoitu vain näihin (MAR 30 artikla). MAR:n lisäksi Suomen kansallinen lainsäädäntö myöntää lisämääreitä asetusta täydentämään. Koska MAR on suunnattu liikkeeseenlaskijoille, jättää se sääntelyn ulkopuolelle muun muassa eläkeyhtiöt ja rahastoyhtiöt. Kotimaisen lainsäädännön tarkoituksena on siis laajentaa lainsäädäntöä koskemaan myös muita rahoitusmarkkinatoimijoita.

Vaikka sisäpiirikauppa on ollut Suomessa säädeltyä vuodesta 1989 alkaen ja sakkotuomioita ja hallinnollisia seuraamusmaksuja on tullut arvopaperimarkkinalain voimaantulohetkestä lähtien, ovat varsinaiset vankeustuomiot yleistyneet vasta vuodesta 2006 alkaen (Pietiläinen 2008). Ehkäpä

Suomen tunnetuin sisäpiirikauppaan liittyvä tapaus on Talvivaaran Kaivososakeyhtiö Oyj:n johdon tekemät laittomat sisäpiirikaupat ja tiedottamisrikokset vuosina 2011–2013. Tuolloin kaivoksen johtajana toiminut Lassi Lammassaari teki osakekauppoja satojen tuhansien eurojen arvosta, ansaiten yhteensä 220 000 € rikoshyötyä. Lammassaari myi osakkeita tietäessään, että toteutunut nikkelituotanto oli alhaisempi, kuin julkiset ennusteet antoivat ymmärtää, sekä lisäksi koko kaivoksen kannattavuuteen vaikuttava nikkelipitoisuus oli matalampi kuin julkisuudessa oli kerrottu. Lopulta Lammassaari tuomittiin kuuden kuukauden ehdolliseen vankeusrangaistukseen, menettämään rikoshyötynä 50 000 € sekä maksamaan 30 päiväsakkoa, eli hänen tuloillaan noin 3 000 € (MTV Uutiset 2017). Myöhemmin samassa vyyhdissä syytettiin sisäpiirikaupoista myös yhtiön entistä toimitusjohtaja Pekka Perää, entistä varatoimitusjohtaja Saila Miettinen-Lähdettä sekä yhtiön entistä kaupallista johtajaa Pekka Erkinheimoa. Syyttäjän mukaan kolmikko syyllistyi törkeään sisäpiiritiedon väärinkäyttöön käydessään Talvivaaran osakkeiden merkintäoikeuksilla kauppaa vuonna 2013. Lopulta Helsingin käräjäoikeus kuitenkin hylkäsi syytteet, sillä se katsoi, ettei vuosituotantotavoitteen laskulla ja nikkelipitoisuuden laskevalla trendillä olisi ollut huomattavaa vaikutusta osakkeen arvoon, eikä siten kolmikolla ollut hallussaan sisäpiiritietoa (Helsingin Sanomat 2020). Syyttäjät valittivat myöhemmin asiasta korkeimpaan oikeuteen, mutta eivät saaneet valituslupaa.

Talvivaaran tapaus osoittaa selkeästi sisäpiirikauppaan liittyvän epävarmuuden ja vaikeuden. Koska kaikki sisäpiiriläisten tekemät kaupat eivät ole lainvastaisia, voi olla todella vaikeaa tehdä rajanvetoa laillisen ja laittoman kaupan välillä. Koska todistustaakka on oikeusvaltiossa yleensä syyttäjällä, eikä pelkkä kaupan ajoitus riitä todisteeksi sisäpiirikaupasta, on tutkivalle viranomaiselle yleensä vaikeaa vedenpitävästi osoittaa sisäpiirikaupan tapahtuneen.

4.2 Sisäpiirikaupat ja rahoitusteoria

Rahoituksessa ja taloustieteessä sisäpiirikauppoja tarkastellaan niin sanotun tehokkaiden markkinoiden hypoteesin kautta (Efficient Market Hypothesis, EMH). EMH on taloustieteilijä Eugene Faman vuonna 1970 julkaisema ja popularisoima aikaisempaan markkinatehokkuuden tutkimukseen pohjautuva teoria, jonka keskiössä on se, että julkisilla markkinoilla kauppaa käytävän sijoitushyödykkeen hinta heijastaa kaiken julkisen ja yksityisen informaation, jolloin kenelläkään ei ole mahdollisuutta ansaita systemaattisesti ylituottoa sijoituksesta (Fama 1970). Markkinatehokkuus, eli tarkemmin allokatiivinen tehokkuus, tarkoittaa yksinkertaisimmillaan sitä että ”taloudelliset varat ohjautuvat sinne, mistä saa parhaan tuoton” (Kurenmaa 2003).

EMH:n tausta-ajatuksena on se, että markkinat voidaan jakaa kolmeen eri kategoriaan niin sanotun tehokkuuden kautta. Markkinoita on teorian mukaan heikosti tehokkaita, puolivahvasti tehokkaita

sekä vahvasti tehokkaita (Fama 1970). Heikosti tehokkailla markkinoilla teorian mukaan kaikki historiallinen markkinatieto näkyy sijoitusinstrumentin hinnassa. Heikosti tehokkailla markkinoilla tekninen analyysi ei siis teorian mukaan luo ansaintamahdollisuuksia. Puolivahvasti tehokkailla markkinoilla puolestaan teorian mukaan ajatellaan, että kaikki julkinen tieto on heijastunut sijoitusinstrumentin hintaan. Tässä tapauksessa tekninen- tai fundamenttianalyysi ei siis luo ansaintamahdollisuuksia sijoitushyödykkeellä, koska kaikki julkinen tieto on jo heijastunut hintaan. Viimeinen ja aihealueemme kannalta merkittävin osuus koskee kuitenkin vahvasti tehokkaita markkinoita. Vahvasti tehokkailla markkinoilla ajatellaan teorian mukaan, että kaikki julkinen sekä yksityinen tieto heijastuu osakkeen hintaan. Tällaisessa tilanteessa ei siis pitäisi olla teorian mukaan mahdollista saada systemaattista ylituottoa minkäänlaisin informaatioon pohjautuvin keinoin (Vallely, 2018). Näistä kolmesta markkinatehokkuuden muodosta seuraa teorian mukaan se, että koska osakkeen hinta heijastaa aina käyvän markkina-arvonsa, on ylituottoa pitkällä ajanjaksolla mahdotonta saada ylimääräistä riskiä ottamatta (Fama 1970).

EMH:lle on kuitenkin esitetty lukuisia vasta-argumentteja. Aihetta kuvaa ehkä parhaiten seuraava taloustieteilijöiden keskuudessa ikoniseksi muodostunut vitsi:

> Kaksi ekonomistia kävelee kadulla. Toinen heistä sanoo: "Katso, tuolla on kahdenkymmenen dollarin seteli maassa!" Toinen ekonomisti vastaa: "Ei ole. Jos olisi, joku olisi jo poiminut sen (Corcoran 2024).

Vitsiin kyteytyy EMH-kritiikin kärki. On melko paradoksaalista, että samaan aikaan markkinoiden oletetaan olevan täysin tehokkaat ja heijastavan samalla kaiken olemassa olevan informaation, mutta käytännössä sijoittajat etsivät jatkuvasti anomalioita ja mahdollisuuksia ylituottoon. Jos markkinat todella olisivat vahvasti tehokkaat, ei poikkeamia tai osakkeiden aliarvostuksia pitäisi käytännössä esiintyä. Useat empiiriset havainnot ovat osoittaneet, että markkinoilla esiintyy anomalioita. Hyväksi esimerkiksi tästä voimme nostaa vaikkapa momentum-ilmiön. Momentum-ilmiö viittaa markkinoilla havaittavaan kaavaan, jossa aikaisempina kuukausina hyvin tuottaneet osakkeet nousevat myös tulevaisuudessa (Jegadeesh & Titman 2001). Tämä on ilmiselvästi ristiriidassa EMH:n kanssa, sillä teorian mukaan investointihyödykkeen arvon ei pitäisi määräytyä historiallisen datan seurauksena. EMH saattaakin siis olla melko idealistinen kuvaus markkinoiden toimintamekanismeista, ja jättää teoriana huomiotta muun muassa behavioraalisen rahoituksen näkökulmat sijoittajien rationaalisuudesta.

Sisäpiirikauppojen kannalta huomattavinta on, että jos oletetaan markkinoiden toimivan tehokkaasti, ei sisäpiiritiedolla pitäisi pystyä ansaitsemaan ylituottoa. Käytännössä kuitenkin tutkimukset ja empiria ovat osoittaneet sisäpiirikauppojen tuottavan odotettua paremmin, joka viittaa siihen, että markkinat eivät ole täysin vahvasti tehokkaat. (Doffou 2007, 5) Kysymykseksi syntyykin se, että missä

määrin yksityinen informaatio vaikuttaa hintoihin ja kuinka tehokkaasti markkinat todella heijastavat informaatiota.

Toinen sisäpiirikauppoihin merkittävästi kytkeytyvä rahoitusteoria liittyy signalointiteorioihin. Signalointi voi tarkoittaa rahoituksen ja taloustieteen kontekstissa esimerkiksi yrityksen pääomarakenteeseen liittyviä ominaisuuksia (Niskanen & Niskanen 2013, 290 - 291), mutta sisäpiirikauppojen yhteydessä signalointi tarkoittaa yrityksen johdon tai muiden sisäpiiriläisten tekemiä kauppoja, joissa on mahdollisuus välittää markkinainformaatiota esimerkiksi yrityksen tulevaisuudennäkymistä. Koska sisäpiiriläisillä on velvollisuus julkistaa yritykseen liittyvät kauppansa, voivat esimerkiksi yritykseen liittyvät osakeostot olla merkki siitä, että yrityksen uskotaan olevan aliarvostettu. Vastavuoroisesti myyntien voidaan uskoa viestivän mahdollisista ongelmista yrityksessä, ja siten viestiä osakkeen olevan esimerkiksi yliarvostettu. Yrityksen sisäpiiriläisten tekemät myynti- tai ostopäätökset voivat siis toimia signaaleina markkinoille yrityksen arvosta. Toisaalta signalointiteorian sovellettavuuden ongelmaksi syntyy se, että sisäpiiriläiset saattavat tehdä myös ei-sanktioitua sisäpiirikauppaa esimerkiksi verotuksellisista syistä, jolloin myynti- tai ostosignaalit eivät ole aina vedenpitäviä. Kaikki ostot eivät siis välttämättä tarkoita aliarvostusta, eivätkä kaikki myynnit ennusta laskua.

Kolmas ja tässä tutkielmassa viimeinen läpikäytävä sisäpiirikauppoihin liittyvä rahoitusteoria on informaatioasymmetriaan perustuva teoria. Asymmetrinen eli epätasaisesti jakautunut informaatio tarkoittaa rahoitusteorian ja sisäpiirikauppojen kontekstissa markkinatilannetta, jossa joillain markkinatoimijoilla on hallussaan enemmän informaatiota kuin toisilla. Sisäpiiri-ulkopiiri-jako onkin erinomainen malliesimerkki tästä tilanteesta. Asymmetristä informaatiota pidetään epätoivottavana, sillä se voi johtaa markkinoiden tehottomuuteen tai teorian tasolla jopa luhistumiseen haitallisen valikoitumisen ja moraalikadon kautta (Akerlof 1970).

Haitallinen valikoituminen (adverse selection), tarkoittaa sitä, että markkinatoimijat eivät erota hyvää ja huonoa informaatiota toisistaan, mikä johtaa siihen, että huonommin informoidut sijoittajat tekevät itselleen epäedullisia päätöksiä. Sisäpiirikontekstissa tämä tarkoittaa sitä, että sijoittaja ei välttämättä tiedä, mikä osa hinnasta perustuu yhtiön ominaisuuksiin eli fundamentteihin, ja mikä osa perustuu sisäpiiritietoon. Moraalikato (moral hazard) puolestaan tarkoittaa kontekstissamme sitä, että asymmetrisesta informaatiosta nauttiva ottaa suurempia riskejä, kantamatta kuitenkaan täysiä vastuita ja jonkun toisen kustannuksella. Esimerkiksi yritysjohto saattaisi tapauksessamme olla altis tekemään yritykseen liittyviä päätöksiä, jotka eivät hyödyttäisi osakkeenomistajia, vaan johtoa itseään (Padilla, 2002). Molemmat sisäpiirikauppojen informaatioasymmetriaan liittyvät ongelmat ovatkin vankasti yhteydessä päämies-agentti-ongelmaan.

Sisäpiirikauppojen kieltoa puoltavien henkilöiden sääntelyn teoreettinen ajattelu perustuu yleensä kolmeen eri ajattelumalliin. Merkittävimmän ajattelutavan mukaan sisäpiirikauppojen laillisuus voi johtaa rahoitusmarkkinoilla ”likviditeettipuutteeseen, johdon vääriin tavoitteisiin tai sijoittajien luottamuksen menetykseen pääomamarkkinoita kohtaan” (Fishman & Hagerty 1992, 106). Sijoittajat saattavat menettää siis luottamuksensa markkinoiden reiluuteen ja oikeudenmukaisuuteen, erityisesti silloin, kun he kokevat, että sisäpiiriläiset hyödyntävät etuoikeutettua tietoa. Tällöin sijoittajat vähentävät kaupankäyntiään peläten, että heidän sijoituksensa eivät ole turvassa. Tämä puolestaan johtaa kaupankäyntivolyymien vähenemiseen, mikä heikentää markkinoiden likviditeettiä. Vähäinen likviditeetti tekee osakkeiden osto- ja myyntiprosesseista vaikeampia ja altistaa ne suurille kurssivaihteluille, vaarantaen potentiaalisesti jopa koko talouden toiminnan. Toisen ajattelumallin mukaan sisäpiirikaupan salliminen puolestaan ”vahingoittaa kyseiset arvopaperit liikkeeseen laskenutta yhtiötä” (Kurenmaa 2003). Kolmas ajattelutyyli puolestaan perustelee kieltoa siten, että kyseessä oleva informaatio on yhtiön omaisuutta, ja siksi sitä ei saa päästää suunniteltua ennen markkinoille. (Kurenmaa 2003). Tämän lisäksi joidenkin sääntelyä kannattavien mielestä sisäpiirikaupan hyväksyminen lannistaisi sijoittajia informaation keräämisestä sekä vinouttaisi informaatiojakaumaa entisestään, johtaen tehottomampiin markkinoihin (Fishman & Hagerty 1992, 107).

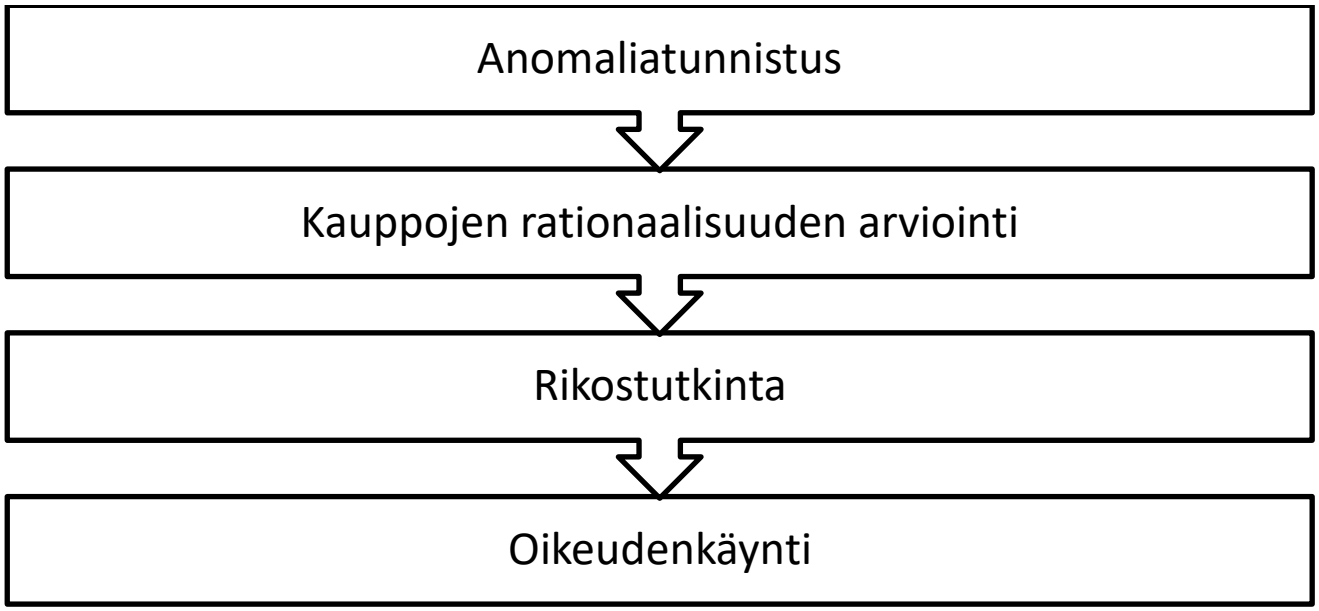
Vastakkaisen mielipiteen mukaan sisäpiirikaupan salliminen voisi parhaimmissa tapauksissa jopa edistää markkinoiden tehokkuutta. Manne (1966) tunnetusti argumentoi, että sisäpiirikaupat voisivat nopeuttaa informaation pääsyä markkinoille, joka puolestaan parantaisi markkinahintojen paikkansapitävyyttä. Jos yrityskaupat olisivat sallittuja, sijoittajille saattaisi siis tulla suurempi kannustin seurata ja analysoida yrityksen toimintaa, josta voisi seurata markkinatiedon tehokkuutta. Markkinoiden tehokkuuden parantuminen puolestaan vaurastuttaisi koko markkinataloutta, sillä resursseja ylijäämäsektorilta alijäämäsektorille ohjaava markkinamekanismi toimisi paremmin (Kurenmaa 2003). Tiivistetysti voidaankin väittää kiellon vastustajien ajatusten pohjautuvan siihen, että koska sisäpiirikaupan kieltoa on vaikea jollei jopa mahdoton estää, tulisi siitä luopua. Jakolinjat vaikuttavat syntyneen tässä suhteessa siis ”kieltoa vastustavien ekonomistien ja kieltoa puoltavien oikeustieteilijöiden” väliin (Kurenmaa 2003). Toisaalta myös taloustieteen näkökulmasta sisäpiirisääntelylle voi olla perusteita. Bhattacharyan & Daoukin (2002) mukaan sisäpiiritiedon väärinkäytön laittomuus alentaa pääoman hankintakustannuksia.

5 Data-analytiikka sisäpiirikauppojen tunnistamisessa

5.1 Perinteiset ja tilastolliset menetelmät

Tässä kappaleessa käydään läpi, miten epäilykset laittomista sisäpiirikaupoista heräävät ja miten epäilyksiä tutkitaan perinteisiin ja tilastollisiin menetelmiin perustuen. Laittomista sisäpiirikaupoista ilmoitetaan valvovalle viranomaiselle harvoin, koska kyseessä on rikos, jossa ei nähdä olevan asianomistajaa perinteisessä mielessä, sillä rikoksen uhrit jäävät yleensä piiloon (Kurenmaa 2003, 275). Sisäpiirikauppa on rikoksena perinteisesti ollut aliedustettu tuomioistuimissa todelliseen rikosmäärään nähden, koska rikoksia on ollut vaikea huomata ja vielä vaikeampaa todistaa rikoksiksi, sillä yksi epäilyttävästi ajoitettu transaktio ei vielä riitä sisäpiiritiedon väärinkäytön todisteeksi.

Laittoman sisäpiirikaupan huomaaminen ja todistaminen on monimutkainen ja monivaiheinen prosessi (Mazzarisi ym. 2024, 2). Yleensä sisäpiiririkosten väärinkäytön tutkiminen alkaa pörssitiedotteen julkistamisesta. Rahoitusvälineiden liikkeeseenlaskijan tulee säännöllisen ja jatkuvan tiedonantovelvollisuuden perusteella julkaista sisäpiiritieto välittömästi ja samanaikaisesti, ettei sisäpiiritietoa voida hyödyntää epäoikeudenmukaisesti (Kurenmaa 2003, 276-277). Tutkinnan alkuvaiheessa kerätään tietoa suurista ja epäilyttävästi ajoitetuista kaupankäyntitapahtumista halutulta aikaväliltä, esimerkiksi suljetun ajanjakson ajalta. Tässä vaiheessa ei vielä epäillä ketään, vaan kyse on tavanomaisesta tiedonkeruusta. Jos tästä tietoaineistosta huomataan merkittäviä väärinkäytöksiin viittaavia anomalioita, kohdistetaan havaittuihin transaktioihin lisätutkintaa (Mazzarisi ym. 2024, 2). Tämän jälkeen selvitetään, onko kauppoja ollut mahdollista tehdä järkeviin analyyseihin tai strategiaan perustuen. Jos tälle ei löydy tukea, aletaan selvittämään onko kaupantekijöillä kytköksiä liikkeeseenlaskijaan, eli selvitetään ovatko hankekohtaiset sisäpiiriläiset, pysyvät sisäpiiriläiset tai sisäpiiriläisten läheiset hyödyntäneet saamaansa sisäpiiritietoa laittomasti. Sisäpiiriläisten lähipiiri ja lähiyhtiöt tutkitaan erikseen siksi, koska on melko harvinaista sekä tekijältään ajattelematonta käyttää sisäpiiritietoa hyväkseen, tietäen että on mahdollisten tutkintojen kohteena. Ajatuksena on siis se, että tietoa vuodetaan lähipiirille tai -yhtiölle sijaisen kaupankäynnin toteuttamisen vuoksi (Kurenmaa 2003, 281). Viimeinen prosessin vaihe on oikeuskäsittely, jossa mahdollinen sisäpiiritiedon väärinkäyttö käsitellään (Mazzarisi ym. 2024, 2).



Kuvio 4: Sisäpiirikauppojen tunnistamis- ja todistamisprosessi (Mazzarisi ym. 2024, 2)

Anomaliatunnistuksen perinteisiä havaintomenetelmiä on monia. Kurenmaan (2003, 286–287) mukaan ainakin 2000-luvun alussa suuri osa väärinkäyttötapauksista on tullut ilmi päivittäisen rahoitusmarkkinoiden seurannan mukana. Tähän kategoriaan kuuluu pörssitiedotteiden kuten muun muassa yritysfuusioita tai julkisia ostotarjouksia edeltävien kaupankäyntipäivien seuranta. Tilastollisten menetelmien alaisuuteen kuuluu puolestaan päivittäisten kaupankäynnin tunnuslukujen, kuten vaihdantavolyymin tai hintamuutosten seuranta. Tilastollisten menetelmien tarkoituksena on nostaa esiin piikkejä tai muita huomattavia kuvioita kaupankäynnissä ja siten tutkia tarvetta mahdolliselle lisätutkinnalle.

Jos edellytykset täyttyvät, voi varsinainen sisäpiiritutkinta alkaa. Koska kaikkien esiinnousseiden sisäpiirikauppojen tutkinnalle ei välttämättä löydy riittävästi resursseja, tehdään yleensä lisäkarsintaa tutkittavista kaupoista niiden arveltujen markkinavaikutusten perusteella. Lisätutkinnassa priorisoidaan siis sellaisia kauppoja, joiden arvellaan vaikuttavan kaikista negatiivisimmin markkinoiden toimintaan ja markkinaluottamukseen (Kurenmaa 2003, 289). Varsinaisen rikostutkinnan toteuttaa Suomen poliisi Finanssivalvonnan tutkintapyynnöstä.

Tämän lisäksi Finanssivalvonta tutkii epäiltyjä väärinkäytöksiä ns. whistleblowing-järjestelmän kautta. Whistleblowing tarkoittaa tässä kontekstissa sitä, että yksityinen henkilö tekee joko nimettömästi tai omalla nimellään Finanssivalvonnalle ilmoituksen epäillyistä väärinkäytöksistä. Whistleblowing-järjestelmää voidaan käyttää myös muiden markkina- ja talousrikoksien selvittämiseen, mutta erityisen hyödyllinen se on muuten vaikeasti havaittavien sisäpiirikauppojen huomaamisessa

(Finanssivalvonta: väärinkäytösepäily 2025). Whistleblowingin kaltaisia menetelmiä hyödyntävät myös markkinatoimijat. Markkinatoimijat ovat velvoitettuja valvomaan kaupankäyntiä ja ilmoittamaan Finanssivalvonnalle kyseenalaisista transaktioista (Finanssivalvonta: sisäpiiritiedon käyttö- ja ilmaisukiellot koskevat muitakin kuin sisäpiiriläisiä 2022).

Vaihdantavolyymin ja hintamuutosten seurannan menetelmiä on olemassa useita. Esimerkiksi poikkeava ylituotto joko päivätasolla tai kumulatiivisesti voivat viitata laittomiin sisäpiirikauppoihin. Poikkeava ylituotto (abnormal return, AR) mittaa kuinka paljon tuotot poikkeavat odotetuista tuotoista, ja signaloi siten epänormaaliutta rahoitusvälinevaihdannassa. Poikkeavan ylituoton kaava on yksinkertainen:

$$AR = R_t - E[R_t] \qquad (1)$$

missä AR on poikkeava ylituotto, R_t on toteutunut tuotto ja $E[R_t]$ on rahoitusvälineen odotettu tuotto.

Vastaavasti kumulatiivinen ylituotto (cumulative abnormal return, CAR) arvioi tuottotasoa halutulla aikaikkunalla:

$$CAR = \sum AR \qquad (2)$$

Vaihdantavolyymia voidaan arvioida puolestaan volyymisuhteen (volume ratio) avulla. Volyymisuhde tarkoittaa yksinkertaisesti suhdelukua, joka mittaa päivittäistä vaihdantavolyymiä suhteutettuna keskimääräiseen vaihdantavolyymiin jollain tutkitulla aikavälillä.

Edellä mainittujen menetelmien hienous liittyy niiden toteutukselliseen helppouteen, sillä suurien poikkeamien etsiminen datasta ei ole kovin vaikeaa yksinkertaisten analyysimuotojen avulla. Ne eivät kuitenkaan ole toiminnallisesti täydellisiä, sillä merkittävä osa sisäpiirikaupoista on piilotettu jopa useamman henkilö- tai yrityskytköksen päähän. Tämä tarkoittaa kysyntää edistyneemmille menetelmille, kuten sosiaaliselle verkostoanalyysille tai aiemmin esitellylle tilastotieteelliselle GARCH:ille.

Sosiaalinen verkostoanalyysi tai pelkkä verkostoanalyysi tarkoittaa erilaisia tutkimusmenetelmiä, joilla voidaan tutkia muun muassa informaation tai muiden resurssien vaihtoa toimijoiden kesken (Haythornthwaite 1996). Verkostoanalyysi keskittyy toimijoiden välisiin yhteyksiin sekä näiden yhteyksien rakenteiden analysoimiseen ja selvittämiseen. Rahoitustieteen ja sisäpiirikauppojen tunnistamisen ohella sosiaalisella verkostoanalyysillä onkin käyttökohteita muun muassa rikostutkinnassa ja epidemiologiassa. Sisäpiirikauppojen tunnistamisessa verkostoanalyysiä saatetaan toteuttaa siten,

että epäilyttävän kaupan huomattuaan tutkintaa kohdistetaan transaktion toteuttaneen henkilön lähipiiriin tai lähiyrityksiin.

Aikasarja-dataan pohjautuva GARCH:in suosio ja hyödyllisyys selittyy sen volatiliteettikeskeisyydellä. Korkean volatiliteetin osakkeissa GARCH saattaa tunnistaa ne ajankohdat, joissa volatiliteetti on korkea ilman selkeää markkinatietoperusteista syytä. Jos esimerkiksi volatiliteetti kasvaa ennen tulosjulkistusta, voi se viitata sisäpiiritiedon väärinkäyttöön. Kuten muidenkin työkalujen ja menetelmien kanssa, myös GARCH toimii parhaiten osana monimetodista havainnointia.

5.2 Koneoppimismenetelmät ja tekoäly

Perinteisten ja tilastotieteellisten analyysimenetelmien lisäksi viimeisen muutaman vuosikymmenen aikana big data -analytiikkaan, koneoppimiseen ja tekoälyyn pohjautuvat metodit ovat tulleet käyttökelpoisiksi ja suosituiksi erityisesti tilanteissa, joissa laajasta ja monimutkaisesta datasta tulisi tunnistaa epäsäännöllisiä ilmiöitä kuten sisäpiiritiedon väärinkäyttöä. Koneoppimisen käyttökelpoisuus sisäpiirikauppojen tunnistamisessa piilee siinä, että se voi havaita monimutkaisia kuvioita ja yhteyksiä, joita perinteisillä tilastollisilla menetelmillä ei pystytä tunnistamaan.

Syväoppimisen kaltaiset tekoälypohjaiset menetelmät voivat käsitellä valtavia tietomääriä ja siten löytää epäilyttäviä kaavoja kaupankäynnissä. Rahoitusmarkkinavalvontaa tekevien viranomaisten, kuten suomalaisen Finanssivalvonnan ja Yhdysvaltalaisen SEC:n, valvontajärjestelmät saattavat esimerkiksi käyttää ohjattuun oppimiseen perustuvia malleja, jotka koulutetaan historiallisten sisäpiirikauppojen perusteella. Vaihtoehtoisesti valvontaviranomaisilla saattaa olla käytössä ohjaamattomaan oppimiseen perustuvia valvontajärjestelmiä, jotka etsivät poikkeavuuksia datasta ilman ennakkokäsitystä.

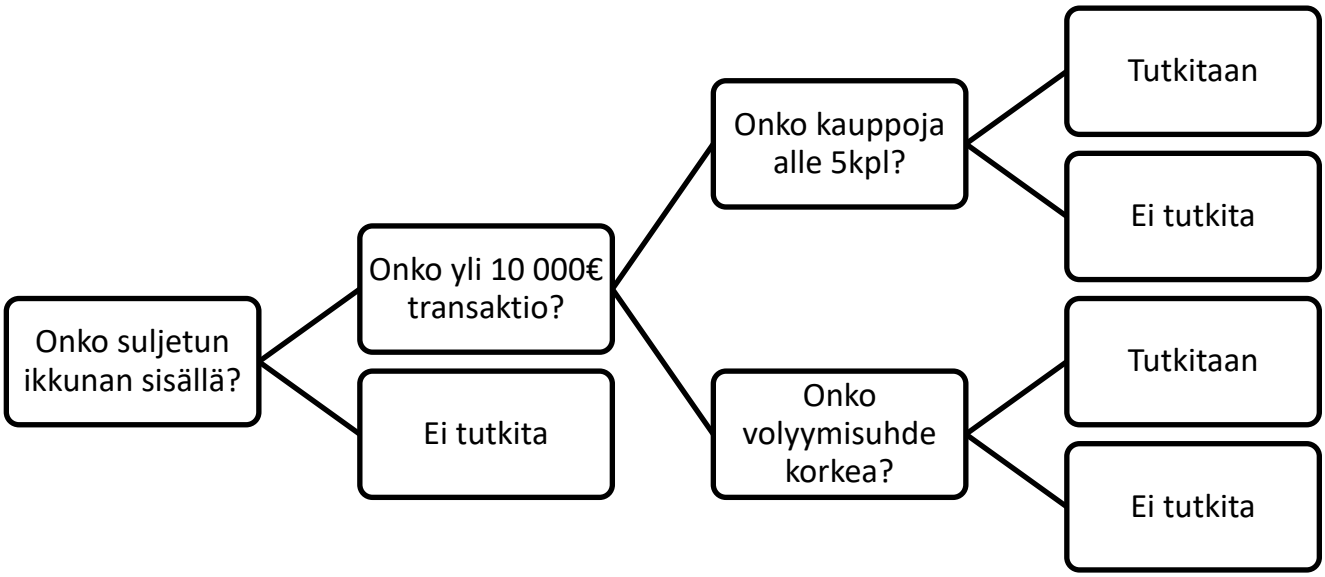
Tarkkaa tietoa valvontaviranomaisten käyttämistä järjestelmistä ei juuri ole saatavilla niiden rikostutkinnallisen arvonsa takia, mutta ainakin kahden SEC:n data-analytiikkatyökalujen toiminnallisuuden yleisperiaatteista tiedetään. ARTEMIS (Advanced Relational Trading Enforcement Metrics) on näistä ensimmäinen, ja se toimii analysoimalla SEC:n optio- ja vaihdantatietokantoja etenkin sarjarikkomusten varalta. ARTEMIS on tekoälypohjainen ja käyttää eri metriikoita asettamaan sijoittajia arvojärjestykseen markkinarikostodennäköisyyden perusteella (Hawke 2019). Luonnollisesti nämä metriikat eivät ole yleisölle julkisia. Toinen SEC:n esittelemä työkalu sisäpiirikauppojen tunnistamiseen on ATLAS (Abnormal Trading and Link Analysis System), joka keskittyy sarjarikkomusten sijaan ensikertalaisiin (Engstrom & Ho 2020, 816). Huomionarvoista työkaluissa on se, että kumpikaan ei ollut ainakaan vuonna 2019 täysin automatisoitu, joskin tilanne on saattanut muuttua.

Ohjatun oppimisen järjestelmien alle kuuluu useita eri algoritmeja. Yleistä niille kuitenkin on se, että ne ovat pitkälti luokittelualgoritmeja. Luokittelualgoritmeihin kuuluvat muun muassa päätöspuut (decision tree) ja satunnaismetsä (random forest).

Päätöspuiden toimintamekanismi perustuu siihen, että ne jakavat datan eri haaroihin päätöksiin perustuen. Algoritmi siis läpikäy erilaisia mahdollisia muuttujia ja tekee valintoja niiden perusteella, jolloin lopputuloksena syntyy lehtipuuta muistuttava rakenne. Päätöspuut ovat skaalautuvia ja niiden toimintamekanismi on helppo ymmärtää, mutta ne ovat alttiita ylisovittamiselle (Golmohammadi ym. 2014). Ylisovittaminen tarkoittaa mallien ennustavan harjoitusdatan perusteella hyvin, mutta reagoivat ja ennustavat huonosti, kun mallille syötetään uutta dataa. Ylisovittamisessa harjoitusdata on rakenteellista, mutta malli ei toimi koska rakenne ei sovellu muuta erilaista varten. Kyseessä on yleinen koneoppimiseen ja ennustealgoritmeihin liittyvä ongelma.

Päätöspuualgoritmeja on mahdollista hyödyntää tilanteissa, joissa halutaan mallintaa sisäpiirikauppaan liittyviä päätöksentekopolkuja. Muuttujat kuten kaupankäyntiajankohta ja volyymipoikkeamat voivat muodostaa mallin, joka erottaa normaalit ja asiaankuuluvat transaktiot anomalioista. Niiden etuna on etenkin tulkittavuus, sillä valvontaviranomainen voi niiden avulla jäljittää miksi jokin transaktio nähdään algoritmin silmissä riskipitoisena ja epänormaalina,

Päätöspuista on jatkokehitetty lisää algoritmeja, joista etenkin GBDT (Gradient-Boosted Decision Tree) soveltuu sisäpiirikaupan tunnistamiseen. GBDT eroaa normaalista päätöspuusta siten, että se hyödyntää gradienttitehostusta (gradient boosting) ennustetarkkuuden parantamiseksi. Gradienttitehostus lisää uusia puita iteratiivisesti siten, että jokainen pyrkii korjaamaan edellisten puiden tekemät virheet (Deng ym. 2019). Tämä johtaa siihen, että GBDT on erityisen tehokas monimutkaisten ja epälineaaristen kaavojen huomaamisessa. Dengin ym. (2019) mullistavassa tutkimuksessa huomattiin, että GBDT on tehokkain, kun suljettu aikaikkuna on 90 päivää. Tämä indikoi sitä, että suljetun aikaikkunan pituutta tulisi mahdollisesti tarkastella enemmän. Erityisen arvokasta tosin on se, että menetelmä menestyi kaikkia muita vertailtuja menetelmiä paremmin laittomien sisäpiirikauppojen tunnistamisessa, joskin huomioon on otettava se, että analysoitu data oli kiinalaista markkinadataa, joka herättää kysymyksiä tutkimuksen toistettavuudesta esimerkiksi länsimaisilla markkinoilla.



Kuvio 5: Yksinkertainen sisäpiirikauppojen päätöspuu

Satunnaismetsä on eräs tunnetuimmista klassifikaatiomenetelmistä ja se puolestaan yhdistää suuren määrän päätöspuita parantaakseen luokittelutarkkuutta ja vähentääkseen algoritmin ylisovittamista dataan. Jokainen satunnaismetsän päätöspuu koulutetaan siis erikseen, ja lopullinen algoritmin lopputulos määräytyy näiden puiden kokonaisennusteena. Satunnaismetsän tehokkuus perustuu niin sanottuun bootstrap-menetelmään ja satunnaistettuihin muuttujavalintoihin jokaisessa päätöspuussa. Tämä tarkoittaa, että puut koulutetaan hieman eri otoksella datasta, joka johtaa mallin yleistettävyyden paranemiseen. Lopullinen päätös tapahtuu mallissa äänestämällä (majority voting), jossa päätöspuut määrittävät onko arvioitu tapahtuma todennäköinen vai ei. Sisäpiirikauppojen kontekstissa tämä voisi merkitä esimerkiksi transaktion epäilyttävyyden arvioimista.

Varsinaisessa anomaliatunnistuksessa paras algoritmi on monesti useamman algoritmin yhdistelmä. Deng ym. (2019) esitti tutkimuksessaan myös toisen tehokkaan algoritmin sisäpiirikaupan tunnistamiseen, jossa hyödynnettiin GBDT:n lisäksi differentiaalievoluutioksi (Differential Evolution, DE) nimitettyä stokastista eli sattumanvaraisesti etenevää algoritmia. Tämä yhdistelmäalgoritmi onnistui havaitsemaan väärinkäytöksiä kiinalaisilla rahoitusmarkkinoilla tehokkaasti. Huomioitavaa tutkimuksessa tosin on se, että kiinalaiset markkinat luetellaan vielä kehittyviksi, ja siten tämänkin algoritmin tehokkuudessa on mahdollisesti eroja markkinoiden välillä.

Edellä mainitut menetelmät ovat pitkälti ohjattuun oppimiseen ja klassifikaatioon perustuvia menetelmiä. Tämän ohella ohjaamattomalla oppimisella on paikkansa sisäpiirikauppojen tunnistamisessa. Ohjaamattoman oppimisen menetelmät tarjoavat tehokkaita tapoja tunnistaa epäilyttäviä transaktioita

ilman ennalta luotuja luokkia, ja on siten hyödyllinen varsinkin siksi, että suuri osa saatavasta datasta ei ole merkittyä sen hinnan ja saatavuuden myötä.

Ohjatun oppimisen tavoin myös ohjaamattoman oppimisen menetelmiä on olemassa lukemattomia, joten käsitellään tässä luvussa niistä vain yleisimmät. Mahdollisesti merkityksellisin näistä menetelmistä sisäpiirikauppojen kontekstissa on k-means klusterointi (k-means clustering) (Mazzarisi ym. 2024). K-means jakaa datan k-määrään ryhmiä, ja sen jälkeen jokainen datapiste kuuluu sellaiseen klusteriin, jonka keskipiste on lähimpänä. Näiden klustereiden avulla voidaan havaita anomalioita datasta, ja tietyissä tapauksissa ne voivat viestiä sisäpiirikaupoista. Ongelmallista tässä algoritmissa on tosin se, että klustereiden määrä eli k voi olla vaikeaa valita etukäteen parhaimmalla mahdollisella tavalla. Jos k:n arvo on valittu huonosti, voi lopputulos olla epätarkka.

Toinen melko yleisesti käytetty anomaliatunnistuksen menetelmä on DBSCAN (Density-Based Spatial Clustering of Applications with Noise). DBSCAN eroaa k-meansista siten, että se perustuu datan ryhmittelyyn tiheyden mukaan sekä siten, että se ei vaadi esimäärittelynään klustereiden määrää (Deng 2020). DBSCAN siis erottaa datapisteet tiheyden mukaan tehtyihin klustereihin sekä harvaan esiintyviin anomalioihin. Tämän perusteella lienee melko selkeää, että DBSCAN:ia pystytään hyödyntämään epäilyttävien transaktioiden, kuten laittomien sisäpiirikauppojen, tunnistamisessa todella hyvin, koska se tunnistaa klustereiden ulkopuolelle jäävät anomaliat onnistuneesti (Yang ym. 2014).

Koneoppimis- ja tekoälypohjaisten menetelmien käyttämisessä on myös haasteita. Väärien positiivisten löydösten määrä voi olla mallista ja algoritmista riippuen todella suuri, ja samat lainalaisuudet todistustaakasta pätevät myös näihin malleihin. Joissain yksinkertaisemmissa malleissa myös ylisovittamista voi olla vaikea välttää, etenkin jos koulutusaineisto on pieni. Koska laittomat sisäpiirikaupat on marginaalinen osuus kaikesta kaupankäyntidatasta, voi olla vaikeaa löytää suuria määriä laadukasta koulutusdataa. Tämä kaikki johtaa siihen, että data-analyyttisten menetelmien erityinen arvo on anomaliatunnistuksessa, joka ei vielä yksinään riitä todistamaan laitonta sisäpiirikauppaa. Tämän lisäksi etenkin tekoälypohjaisissa menetelmissä ongelmaksi nousee algoritmien läpinäkyvyysongelmiin liittyvät seikat. Jossain määrin kliseinen mutta todenmukainen lausahdus tekoälymenetelmien toiminnallisuuden tuntemattomuudesta siis pätee.

6 Yhteenveto ja johtopäätökset

6.1 Keskeiset havainnot

Tämä tutkielma käsittelee data-analytiikan käyttämistä laittomien sisäpiirikauppojen tunnistamisessa, ja sen keskeiset havainnot liittyvät siihen, että data-analytiikalla ja erityisesti moderneilla koneoppimismenetelmillä on suuri potentiaali laittomien sisäpiirikauppojen tunnistamisessa, eli anomaliatunnistuksen parissa.

Tutkielmassa kävi ilmi, että perinteiset tilastotieteeseen ja matematiikkaan pohjautuvat menetelmät tunnuslukuanalyysistä regressio- ja aikasarja-analyyseihin tarjoavat erinomaisia työkaluja anomalioiden ja muiden epänormaalien tuottojen analysointiin, joskin suuret ja monimutkaiset datamäärät saattavat vaikeuttaa niiden tehokasta toteuttamista (Hilal ym. 2022). Koneoppimisen ja tekoälyn vahvuus piilee monimutkaisempien petoksien huomaamisessa. Näistä menetelmistä etenkin ohjatun ja ohjaamattoman oppimisen menetelmät, kuten päätöspuut ja klusterointimenetelmät, soveltuvat mainiosti epäilyttävän kaupankäynnin tunnistamiseen laajasta massadatasta. Tämän ohella on hyvä huomata ja tiedostaa, että varsinaisia tutkintaprosesseja varten soveltuvimpien algoritmien tulee olla yleisesti selitettäviä, sillä mustan laatikon algoritmit saattavat olla huonoja indikaattoreita oikeuden ja silmissä. Logistiseen regressioon ja päätöspuihin nojaavat menetelmät loistavat siis tässä aspektissa, mutta myös ohjaamattomilla menetelmillä on vahvuutensa, etenkin luotettavan koulutusdatan vähäisyyden vuoksi.

Tutkimuksessa havaittiin lisäksi, että markkinoiden läpinäkyvyyden lisäämisessä data-analytiikka on tärkeä työkalu. Analyysimenetelmien kehittyessä markkinavalvojat kuten Finanssivalvonta ja SEC pystyvät keskittämään tutkimuksellisia resurssejaan tehokkaammin epäilyttäviin transaktioihin (Hilal ym. 2022). Tämä kehitys tukee sekä tehokkaiden markkinoiden hypoteesia sekä yleistä markkinaluottamusta. Käytännön sovelluksista lienee kuitenkin edelleen huomioitavaa, että datan laatu ja oikeellisuus vaikuttaa olennaisesti mallien tarkkuuteen ja toimivuuteen. Esimerkiksi big data saattaa sisältää ”saastetta”, joka vaikuttaa välittömästi myös lopputuotteena saatuun tietoon ja ymmärrykseen.

Sisäpiirikauppojen tunnistamisessa analyysimenetelmät ovat pitkälti korrelaatiopohjaisia. Varsinaisten kausaalianalyysin menetelmien kuten satunnaistettujen kokeiden ja kvasikokeiden hyödyntäminen on vaikeaa. Tämä ei ole kuitenkaan ongelmallista, sillä sisäpiirikauppojen tunnistamisessa on kyse pitkälti korrelaatioiden huomaamisesta, ja kausaalisuus todistetaan muita analyysimenetelmiä, kuten esimerkiksi verkostoanalyysiä, myöhemmin käyttämällä (Haythornthwaite 1996).

Verkostoanalyysi onkin muiden analytiikkamenetelmien ohella olennainen osa toimivaa ja luotettavaa prosessia.

Tutkielmassa huomattiin myös se, että koneoppimismenetelmien käytössä haastavinta on etenkin ohjatun oppimisen malleissa löytää sopivaa koulutusdataa sisäpiiritransaktioiden harvinaisuuden vuoksi. Data-analytiikka ja sen menetelmät eivät siis korvaa valvontaa vaan pikemminkin täydentävät ja tehostavat sitä entisestään. Mallien toimintaan ei siis voi suhtautua täysin kritiikittä, vaan aina täytyy pitää mielessä kenellä vastuu lopulta on, eli käyttäjällä.

6.2 Tutkimuksen arviointi ja jatkokysymykset

Kirjallisuuskatsaukseen perustuva tutkimusasetelma toimii monitieteisen ja laajan ilmiön tarkastelussa hyvin, sillä se mahdollistaa aiheen käsittelyn sekä teknisestä, oikeudellisestä kuin myös rahoitusteoreettisesta näkökulmasta. Tämän tutkimuksen suurimpana rajoitteena on kuitenkin empiirisen aineiston puute. Aikapainesyistä suunniteltuja haastatteluosuuksia ei toteutettu, eikä varsinaista sisäpiirikauppa-analyysien toimintaa voitu konkreettisesti esitellä. Näiden asioiden huomioiminen olisi tehnyt tutkimuksesta entistä laadukkaamman, ja käytännön näkökulma melko teknisten menetelmien soveltamisessa olisi lisännyt tutkimuksen validiteettia.

Erilaiset algoritmit DBSCAN:ista logistiseen regressioon toimivat hyvin sisäpiirikauppojen tunnistamisessa. Lienee tosin selvää, että varsinaista markkinavalvontaa suorittavilla viranomaisilla on todennäköisesti olemassa edistyneemmät ja monimutkaisemmat algoritmit kuin tässä tutkimuksessa on esitetty. Valvontakäytössä on edelleen haasteita, mutta tilanne on jo kokonaisuudessaan uuden teknologian myötä todella kehittyneempi, kuin esimerkiksi kaksikymmentä vuotta sitten.

Tulevaisuudessa data-analytiikan, ja eritoten koneoppimisen, rooli markkinavalvonnassa tulee kasvamaan entisestään. 2020-luvun tekoälyn vallankumous ulottaa lonkeronsa joka suuntaan, ja rahoitusalalla markkinarikosten selvittämisessä merkitys voi olla valtava. On mielenkiintoista nähdä miten markkinalainsäädäntö kehittyy niin sanotun ”shadow trading”:in suhteen. Shadow trading tarkoittaa tilannetta, jossa yritys tai yksilö käy sisäpiiritietonsa avulla kauppaa toisen yhtiön osakkeilla, hyötyen samalla tiedosta epäreilusti. Vuonna 2021 SEC onnistui ensimmäistä kertaa soveltamaan uutta oikeusteoreettista tulkintaa ja nostamaan syytteen shadow trading -tapauksessa, jossa se haastoi markkinaväärinkäytöksestä epäillyn oikeuteen SEC v. Panuwat -tapauksessa. On mahdollista, että tapaus muodostaa merkittävän ennakkotapauksen ja vaikuttaa markkinalainsäätelyyn muuallakin, nostaen oikeusteoreettisen pohdinnan tarvetta myös jatkotutkimuksien muodossa (Kershen 2022, 151).

Tutkielma osoittaa, että data-analytiikka ja koneoppiminen tarjoavat merkittäviä mahdollisuuksia sisäpiirikauppojen valvonnassa, mutta hyödyntäminen vaatii paljon teknistä osaamista, riittävää datamäärää sekä sääntelyyn perehtymistä. Teknologia ei poista ihmisen roolia, mutta se toimii tehokkaana apuvälineenä tapauksien seulonnassa.

Mahdollisissa jatkotutkimuksissa olisi järkevää tehdä empiiristä analyysiä historiallisesta kaupankäyntidatasta. Koneoppimismallien tehokkuuden vertailu todistettuihin sisäpiirikauppoihin pohjautuvassa datassa voisi tuoda lisää hyödyllistä tietoa valvottujen ja valvomattomien mallien suorituskyvystä asian tiimoilta. Jatkokysymyksiä herää tehokkuuden ohella myös datan anonymisoinnin ja yksityisyydensuojan vaikutuksesta mallien tarkkuuteen, sillä entistä tehokkaampia ja toimivampia malleja voi olla vaikeaa rakentaa ilman tietosuojalakien ja tekoälysäätelyn rikkomista. Tämän lisäksi tutkimuksen aihealuetta voidaan syventää käsittelemällä laajemmin matemaattisia metodeja esimerkiksi Monte Carlo -menetelmistä, sillä näiden menetelmien käsittely jäi tässä tutkimuksessa todella pinnalliseksi. Yritysjuridiikan ja oikeustieteen näkökulmasta edellä mainittu shadow trading on aiheena hyödyllisiä jatkokysymyksiä potentiaalisesti sisältävä, ja tekoälyn sääntelyn kysymykset ovat keskeisiä myös tämän aiheen kannalta.

7 Lähteet

Aasheim, C. L., Williams, S., Rutner, P., & Gardiner, A. (2015). Data analytics vs. Data science: A study of similarities and differences in undergraduate programs based on course descriptions. *Journal of Information Systems Education*, *26*(2), 103–116.

Akerlof, G. A. (1970). The Market for ”Lemons”: Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, *84*(3), 488–500. JSTOR. https://doi.org/10.2307/1879431

Bhagattjee, B. (2014). *Emergence and taxonomy of big data as a service*. Massachusetts Institute of Technology.

Bhattacharya, U., & Daouk, H. (2002). The World Price of Insider Trading. *The Journal of Finance*, *57*(1), 75–108. JSTOR.

Bowers, D. (1991). *Statistics for Economics and Business*. ELBS with Macmillan. https://books.google.fi/books?id=Z72cAAAACAAJ

Cheng, G., Lundblad, C. T., Yang, Z., & Zhang, Q. (2022). *Detecting Insider Trading in the Era of Big Data and Machine Learning* (SSRN Scholarly Paper No. 4240205). Social Science Research Network. https://doi.org/10.2139/ssrn.4240205

Corcoran, K. (2024, huhtikuuta 2). *Bills on the Sidewalk*. Econlib. https://www.econlib.org/bills-on-the-sidewalk/

Deng. (2020). DBSCAN Clustering Algorithm Based on Density. *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, 949–953. https://doi.org/10.1109/IFEEA51475.2020.00199

Datum. (2025, maaliskuuta 5). https://dictionary.cambridge.org/dictionary/english/datum

De, P. (2016). The arithmetic mean—Geometric mean—Harmonic mean: Inequalities and a spectrum of applications. *Resonance*, *21*(12), 1119–1133. https://doi.org/10.1007/s12045-016-0423-4

Deng, Wang, C., Wang, M., & Sun, Z. (2019). A gradient boosting decision tree approach for insider trading identification: An empirical model evaluation of China stock market. *Applied Soft Computing*, *83*, 105652. https://doi.org/10.1016/j.asoc.2019.105652

Derindere Köseoğlu, S., Ead, W. M., & Abbassy, M. M. (2022). Basics of Financial Data Analytics. Teoksessa *Financial Data Analytics: Theory and Application* (ss. 23–57). Springer.

Doffou, A. (2007). *Insider Trading: A Review of Theory and Empirical Work*.

Engle, R. (2001). GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics. *The Journal of Economic Perspectives*, *15*(4), 157–168. JSTOR.

Engstrom, D. F., & Ho, D. E. (2020). Algorithmic Accountability in the Administrative State Special Issue: Regulating the Technological Frontier. *Yale Journal on Regulation*, *37*(3), 800–854.

FAMA, E. F. (1970). EFFICIENT CAPITAL MARKETS: A REVIEW OF THEORY AND EMPIRICAL WORK. *Journal of Finance (Wiley-Blackwell)*, *25*(2), 383–417. Business Source Ultimate. https://doi.org/10.2307/2325486

Fishman, M. J., & Hagerty, K. M. (1992). Insider Trading and the Efficiency of Stock Prices. *The RAND Journal of Economics*, *23*(1), 106–122. JSTOR. https://doi.org/10.2307/2555435

Gogtay, N. J., & Thatte, U. M. (2017). Principles of correlation analysis. *Journal of the Association of Physicians of India*, *65*(3), 78–81.

Golmohammadi, K., Zaiane, O. R., & Díaz, D. (2014). Detecting stock market manipulation using supervised learning algorithms. *2014 International Conference on Data Science and Advanced Analytics (DSAA)*, 435–441. https://doi.org/10.1109/DSAA.2014.7058109

Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Insights Into the Future of Medicine: Technologies, Concepts, and Integration*, *69*, S36–S40. https://doi.org/10.1016/j.metabol.2017.01.011

Hawke, D. (2019, elokuuta 21). *SEC Data Analysis in Insider Trading Investigations | CLS Blue Sky Blog*. https://clsbluesky.law.columbia.edu/2019/08/21/sec-data-analysis-in-insider-trading-investigations/

Hayashi, C. (1998). *What is Data Science? Fundamental Concepts and a Heuristic Example* (C. Hayashi, K. Yajima, H.-H. Bock, N. Ohsumi, Y. Tanaka, & Y. Baba, Käänt.). 40–51.

Haythornthwaite, C. (1996). Social network analysis: An approach and technique for the study of information exchange. *Library & Information Science Research*, *18*(4), 323–342.

HELM 36. (ei pvm.). https://www.lboro.ac.uk/media/media/schoolanddepartments/mlsc/downloads/HELM%20Workbook%2036%20Descriptive%20Statistics.pdf

Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. *Expert Systems with Applications*, *193*, 116429. https://doi.org/10.1016/j.eswa.2021.116429

Ho, M. K., Darman, H., & Musa, S. (2021). Stock Price Prediction Using ARIMA, Neural Network and LSTM Models. *Journal of Physics: Conference Series*, *1988*(1), 012041. https://doi.org/10.1088/1742-6596/1988/1/012041

Inc, G. (ei pvm.). *Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s*. Forbes. Noudettu 15. maaliskuuta 2025, osoitteesta https://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/

Jegadeesh, N., & Titman, S. (2001). Profitability of Momentum Strategies: An Evaluation of Alternative Explanations. *The Journal of Finance*, *56*(2), 699–720. JSTOR.

Kaakinen, M., & Ellonen, N. (ei pvm.). *Regressioanalyysi—Tietoarkisto*. Noudettu 11. maaliskuuta 2025, osoitteesta https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/kvanti/regressio/analyysi/

Kelleher, J. D., & Tierney, B. (2018). *Data Science*. MIT Press. https://books.google.fi/books?i§d=UlpVDwAAQBAJ

Kershen, K. (2022). SEC v. Panuwat: The Federal Pursuit of Shadow Trading. *Brook. J. Corp. Fin. & Com. L.*, *17*, 151.

Koseoglu, S. D., & Derindere Köseoğlu, S. (2022). *Financial Data Analytics: Theory and Application* (1. p.). Springer International Publishing AG. https://doi.org/10.1007/978-3-030-83799-0

Kurenmaa, T. (2003a). Sisäpiirintiedon väärinkäyttö. Teoksessa *Suomalainen Lakimiesyhdistys*. https://edition.fi/lakimiesyhdistys/catalog/book/479

Kurenmaa, T. (2003b). Sisäpiirintiedon väärinkäyttö. *Suomalainen Lakimiesyhdistys*. https://edition.fi/lakimiesyhdistys/catalog/view/479/394/991-1

Käräjäoikeuden tuomio: Talvivaaran Perälle ja Lammassaarelle ehdollista ja yhtiölle yhteisösakkoa. (2017, kesäkuuta 2). mtvuutiset.fi. https://www.mtvuutiset.fi/artikkeli/karajaoikeuden-tuomio-talvivaaran-peralle-ja-lammassaarelle-ehdollista-ja-yhtiolle-yhteisosakkoa/6456416

Lepenioti, K., Bousdekis, A., Apostolou, D., & Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, *50*, 57–70. https://doi.org/10.1016/j.ijinfomgt.2019.04.003

MAR 19:12, 173 OJ L (2014). http://data.europa.eu/eli/reg/2014/596/oj/fin

Mazzarisi, P., Ravagnani, A., Deriu, P., Lillo, F., Medda, F., & Russo, A. (2024). A machine learning approach to support decision in insider trading detection. *EPJ Data Science*, *13*(1), Article 1. https://doi.org/10.1140/epjds/s13688-024-00500-2

Niskanen, J., Niskanen, M., Edita, kustantaja., & Edita Oppiminen, kustantaja. (2013). *Yritysrahoitus* (7. uud. p.). Edita.

Nurmi, M., & Pyykkönen, J. (2022, maaliskuuta 3). Viisauden hierarkia. *Viisauden hierarkia*. https://blogs.helsinki.fi/yhdenvertainen-liikunnallinen-lahio/2022/03/03/viisauden-hierarkia/

OECD. (2008). *OECD Glossary of Statistical Terms*. OECD. https://doi.org/10.1787/9789264055087-en

Oikeudenkäynnit | Syytteet Talvivaaran sisäpiiririkoksista hylättiin Helsingin käräjäoikeudessa. (2020, heinäkuuta 1). Helsingin Sanomat. https://www.hs.fi/suomi/art-2000006558403.html

Padilla, A. (2002). Can agency theory justify the regulation of insider trading? *The Quarterly Journal of Austrian Economics*, *5*(1), 3–38. https://doi.org/10.1007/s12113-002-1015-6

Perino, M. (2018). The Lost History of Insider Trading. *SSRN Electronic Journal*, 54. https://doi.org/10.2139/ssrn.3099682

Pietiläinen, T. (2008, tammikuuta 2). *Timo Jouhki sai vankeutta sisäpiirikaupoista*. Helsingin Sanomat. https://www.hs.fi/talous/art-2000004537466.html

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

Pörssin sisäpiiriohje. (2020, joulukuuta 2). NASDAQ Helsinki Oy.

R. K. Lomotey & R. Deters. (2014). Towards Knowledge Discovery in Big Data. *2014 IEEE 8th International Symposium on Service Oriented System Engineering*, 181–191. https://doi.org/10.1109/SOSE.2014.25

Rikoslaki, Pub. L. No. 39- 001/1889.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia medica*, *24*(1), 12–18.

Student's t-test | Definition, Formula, & Example | Britannica. (2025, tammikuuta 16). https://www.britannica.com/science/Students-t-test

Thabtah, F., Abdelhamid, N., & Peebles, D. (2019). A machine learning autism classification based on logistic regression analysis. *Health Information Science and Systems*, *7*(1), 12. https://doi.org/10.1007/s13755-019-0073-5

Time Domain Analysis vs Frequency Domain Analysis: A Guide and Comparison. (2024, heinäkuuta 17). https://resources.pcb.cadence.com/blog/2020-time-domain-analysis-vs-frequency-domain-analysis-a-guide-and-comparison

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Upton, G., & Cook, I. (1996). *Understanding Statistics*. OUP Oxford. https://books.google.fi/books?id=vXzWG09_SzAC

Vallely, B. (2018). *The Efficient Market Hypothesis, Insider Trading and their relationship with today's stock markets*. https://www.cpaireland.ie/CPAIreland/media/Education-Training/Study%20Support%20Resources/P2%20Strategic%20Corporate%20Finance/Relevant%20Articles/the-efficient-market-hypothesis-insider-trading-and-their-relationships-with-today-s-stock-markets.pdf

Wooldridge, J. M. (2016). Introductory econometrics: A modern approach. Teoksessa *Introductory econometrics: A modern approach* (Sixth edition.). Cengage Learning.

www.finanssivalvonta.fi. (2022, toukokuuta 19). *Sisäpiiritiedon käyttö- ja ilmaisukiellot koskevat muitakin kuin sisäpiiriläisiä*. www.finanssivalvonta.fi. https://www.finanssivalvonta.fi/tiedotteet-ja-julkaisut/markkinat-tiedotteet/markkinat-tiedote-12022/sisapiiritiedon-kaytto--ja-ilmaisukiellot-koskevat-muitakin-kuin-sisapiirilaisia/

www.finanssivalvonta.fi. (2025, maaliskuuta 6). *Väärinkäytösepäily*. www.finanssivalvonta.fi. https://www.finanssivalvonta.fi/finanssivalvonta/ilmoita-vaarinkaytosepailysta/

Y. Yang, B. Lian, L. Li, C. Chen, & P. Li. (2014). DBSCAN Clustering Algorithm Applied to Identify Suspicious Financial Transactions. *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, 60–65. https://doi.org/10.1109/CyberC.2014.89

Zheng, X., Gildea, E., Chai, S., Zhang, T., & Wang, S. (2024). Data Science in Finance: Challenges and Opportunities. *AI*, *5*(1), 55–71. https://doi.org/10.3390/ai5010004