Jarre Leskinen

Tukivektorikoneen luokitteluongelman valinnan merkitys osakemarkkinoiden ennustamisessa

Tietotekniikan pro gradu -tutkielma

25. elokuuta 2021

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Tekijä: Jarre Leskinen

Yhteystiedot: jarre.leskinen@protonmail.com

Ohjaaja: Raino Mäkinen

Työn nimi: Tukivektorikoneen luokitteluongelman valinnan merkitys osakemarkkinoiden ennustamisessa

Title in English: Impact of support vector machine's classification problem selection in stock market forecasting

Työ: Pro gradu -tutkielma

Opintosuunta: Ohjelmistotekniikka

Sivumäärä: 58+0

Tiivistelmä: Tutkielma käsittelee tukivektorikoneen luokitteluongelman valinnan merkitystä osakemarkkinoiden ennustamisessa. Aikaisempia tutkimuksia erilaisista luokitteluongelmista on vähän, mikä nostaa esille tarpeen tämän aiheen tutkimisen. Uutta keskihajontasuhteutettua luokitteluongelmaa verrataan aikaisemmissa tutkimuksissa suosittuun seuraavan päivän suuntaa ennustavaan luokitteluongelmaan. Tukivektorikoneiden ominaisuudet valitaan aikaisempien tutkimusten perusteella ja niiden tarkkuutta verrataan toisiinsa sekä vertailuindeksinä käytettävään DAX-osakeindeksiin. Ennustemalleista muodostetaan aktiivisia kaupankäyntistrategioita, joita analysoidaan taustatestaamalla käyttäen historiallista kurssidataa. Tulokset osoittavat uuden keskihajontasuhteutetun luokitteluongelman johtavan huomattavasti parempiin tuloksiin sekä korostavan tarvetta jatkotutkimuksille erilaisista luokitteluongelmista.

Avainsanat: Koneoppiminen, tukivektorikone, osakemarkkinat, osakemarkkinoiden ennustaminen

Abstract: This thesis examines the impact of support vector machine's classification problem selection on stock market forecasting. Previous research on different types of classification problems has been minimal which raises the need for research on this topic. A new standard

deviation adjusted classification problem is compared against a popular next day's direction forecasting classification problem. The feature engineering for the support vector machines is based on previous research and they are compared against each other as well as the benchmark stock market index DAX. The forecasting models are used to form active trading strategies that are analysed with backtesting using historical price data. The results demonstrate that the new standard deviation adjusted classification problem produces significantly better results and highlight the need for further studies on different types of classification problems.

Keywords: Machine learning, support vector machine, stock market, stock market forecasting

Kuviot

Taulukot

Sisällys

1 Johdanto

Osakemarkkinoiden ennustaminen on ollut olemassa yhtä pitkään kuin markkinat itse. Syitä ovat parempien tuottojen tavoittelu ja pyrkimys pienempään riskiin. Markkinoiden ennustettavuus on ollut akateemisesti pitkään tutkittu ja yksi edelleen keskeisessä roolissa oleva malli on tehokkaiden markkinoiden hypoteesi (engl. *efficient market hypothesis*). Hypoteesi esittää, että osakkeen hinta kuvastaa täydellisesti kaikkea sillä hetkellä saatavilla olevaa tietoa (Malkiel ja Fama 1970). Markkinoiden ollessa tehokkaat on osakekurssien liikkeet lyhyellä välillä täysin satunnaisia ja näin ollen ne muistuttavat satunnaiskulkua (engl. *random walk*). Näin ollen markkinoiden tuottoa ei riskikorjattuna pitäisi pystyä systemaattisesti ylittämään pitkällä aikavälillä.

Toinen uudempi näkemys, joka on haastanut markkinoiden tehokkuutta ja yleistynyt myöhemmin on behavioraalinen rahoitus (engl. *behavioral finance*), jossa esitetään, että sijoittajat eivät toimi täysin rationaalisesti vaan heidän käyttäytymiseensä liittyy tiettyjä harjoja. Osa harhoista muodostuu sijoittajia ohjaavista tunteista, kuten pelko ja ahneus, jotka eivät ole historian aikana muuttuneet, vaan niiden nähdään olevan osa ihmisten käyttäytymistä (Benartzi ja Thaler 1995). Esimerkkejä tällaisesta käytöksestä ovat olleet erilaiset hinnoittelukuplat, kuten jo vuosina 1634–1637 ollut tulppaanimania (Garber 1989) ja myöhemmin 2007–2008 Yhdysvalloissa muodostunut asuntomarkkinakupla (Brueckner, Calem ja Nakamura 2012). Muita hiljattain olleita tapauksia, joita tehokkaiden markkinoiden hypoteesi ei pysty selittämään on 2021 tapahtunut GameStopin osakkeen äkillinen nousu ja lasku.

Tehokkaiden markkinoiden hypoteesi on kerännyt pitkään osakseen kritiikkiä ja eriäviä näkemyksiä tai vastaesimerkkejä. Malkiel 2003 vastasi tuolloin kertyneeseen kritiikkiin säilyttäen näkemyksen, että markkinat ovat tehokkaat ja seuraavat satunnaiskulkua. Oleellinen pohjalla oleva tekijä, joka tekee aiheesta haastavan, on markkinoiden jatkuva muuttuminen. Mahdollinen ennustettavuus, joka poikkeaisi satunnaiskulusta tarjoten ylimääräistä tuottoa ilman riskin kasvamista, tulisi lakata toimimasta, kun useat markkinatoimijat kävisivät kauppaa tämän käytöksen pohjalta. Tämä vaikuttaisi hintoihin lopulta korjaten hinnoitteluvirheen.

Tietokoneiden ja laskentatehon kasvu on muuttanut osaltaan markkinoiden toimintaa elekt-

ronisen kaupankäynnin myötä sekä tuoden mahdollisuuden toteuttaa monimutkaisempia ja vaativampia malleja markkinoiden toiminnan kuvaamiseen, joka ei aikaisemmin ole ollut mahdollista. Tämän myötä erilaiset koneoppimisen algoritmit ovat nostaneet suosiotaan. Näihin liittyen on toteutettu useita akateemisia tutkimuksia testaten eri algoritmeja ja eri kohteita markkinoilla, kuten kurssiliikkeiden tulevaa volatiliteettia eli hintamuutosten keskihajontaan (Yang, Chan ja King 2002) ja tulevaa kurssiliikettä (Choudhry ja Garg 2008).

Tutkimuksissa on tarkasteltu koneoppimismallien tuloksia suhteessa hypoteesiin kurssiliikkeen satunnaiskulusta tai mallintaen aktiivisen kaupankäynnin strategiaa. Kaupankäynnistä mitataan strategian saavuttamia tuloksia, kun strategia hyödyntää koneoppimismallia osto- ja myyntipäätösten tekemiseen.

Kaupankäyntikulujen huomioiminen on useasti jätetty tutkimuksissa huomioimatta, joka ei ole oleellista tehokkaiden markkinoiden hypoteesin tutkimisen kannalta, sillä se ei ota kantaa kaupankäynnin kuluihin (Malkiel ja Fama 1970). Mallin soveltuvuus aktiiviseen kaupankäyntiin vaatii kuitenkin kaupankäyntikulujen huomioimisen, joka tässä tutkielmassa huomioidaan yksinkertaistetulla mallintamisella, jossa kaupankäyntikulut oletetaan vakioksi ja verrataan näiden vaikutusta kuluttomiin tuloksiin.

Tässä tutkielmassa rakennetaan tukivektorikoneeseen pohjautuva koneoppimismalli, jolla pyritään ennustamaan osakekurssien liikettä käyttäen tukivektorikonetta (engl. *support vector machine*). Tukivektorikoneelle valitaan uudenlainen keskihajontasuhteutettu luokitteluongelma, jota verrataan aikaisemmissa tutkimuksissa suosittuun luokitteluongelmaan. Toteutetun mallin pohjalta rakennetaan yksinkertainen aktiivisen kaupankäynnin strategia, jonka suoriutumista taustatestataan historiallisella kurssidatalla analysoiden sen tulokset verrattuna markkinoiden normaaliin liikkeeseen tuolla ajalla sekä toista luokitteluongelmaa käyttävään tukivektorikoneeseen. Malkiel ja Fama 1970 mukaisesti aikaisempien hintaliikkeiden perusteella ei pitäisi pystyä mallintamaan tulevaa kurssiliikettä satunnaiskulkua paremmin. Tästä syystä tutkimusaiheen tärkeyttä ja sen tuloksia voidaan tarkastella niin sijoittamisen ja rahoitusteorian kannalta. Tulokset voivat antaa vastauksia markkinoiden tehokkuudelle, joka on oleellista rahoitusteorian kannalta. Toinen näkökulma tulosten merkityksellisyyteen on koneoppimisen kannalta, sillä osakekurssien sisältäessä satunnaisuutta ja muistuttaen stokastista prosessia, on niistä luotettavien signaalien löytäminen haastavaa. Toimivan mallin

löytyminen voi tarjota pohjaa myös muille toteutuksilla osakemarkkinoiden ulkopuolella, jossa käytössä oleva aineisto sisältää satunnaisuutta.

Luku 2 aloittaa käsittelemällä tutkielman kannalta oleellisen käsitteistön. Tätä seurataan luvussa 3 esittelemällä tutkielman päätutkimuskysymys ja siihen liittyvät alakysymykset. Seuraavana luku 4 sisältää aikaisempien tutkimusten käsittelyä, jota hyödynnetään tutkielmassa toteutettavien mallien rakentamiseen. Luvussa 5 esitellään tämän tutkielman tukivektorikoneiden aineisto, rakenne ja toteutus. Toteutettujen mallien tulokset käsitellään luvussa 6. Lopuksi tutkimuksen yhteenveto ja tarve mahdolliselle jatkotutkimukselle käydään luvussa 7.

2 Käsitteistö

Tässä luvussa määritellään tutkielmaan liittyvää oleellista rahoitusteoriaan ja erityisesti osakemarkkinoihin liittyvää käsitteistöä. Käsitteistön määrittelyn jälkeen voidaan tutkimuskysymys määritellä luvussa 3.

2.1 Anomalia

Anomalialla (engl. *anomaly*) tarkoitetaan tässä tutkielmassa osakekurssissa tapahtuvaa tilannetta, jolloin tulevaa voidaan ennustaa satunnaista arvausta paremmin eikä kurssikehitys näin ollen sillä hetkellä noudata satunnaiskulkua. Malkiel ja Fama 1970 esittämän tehokkaiden markkinoiden hypoteesin mukaisesti aikaisemman kurssihistorian ei pitäisi mahdollistaa tulevan ennustamista satunnaista arvausta paremmin ja kurssikehitys noudattaa näin ollen satunnaiskulkua. Tämän hypoteesin näkökulmasta poikkeamia satunnaiskulusta voidaan pitää anomalioina.

2.2 Osakemarkkinat

Burton, Nesiba ja Brown 2015 mukaisesti osakemarkkinat voidaan esittää kokonaisuutena, joka yhdistää ostajat ja myyjät. Tämmöisenä markkinapaikkana toimii tyypillisesti pörssi, joka tarjoaa alustan myynti- ja ostotarjousten vastaanottamiselle ja yhdistämiselle. Osakkeiden omistajat voivat hyötyä omistamistaan osakkeista, joko osakkeen arvonnousulla ja sen myymisestä alkuperäistä ostohintaa kalliimmalla, tai osingoista, jotka ovat yritysten tuloksestaan maksamia osuuksia omistajille.

Matalimman myyntitarjouksen ja korkeimman ostotarjouksen välissä on kurssiero (engl. *spread*), josta muodostuu kulu, jos kohteena olevaa osaketta halutaan ostaa tai myydä välittömästi. Myynti- tai ostotoimeksiantojen suorittaminen tapahtuu välittäjän kautta, josta yleensä maksetaan tietty kaupankäyntikulu välittäjälle (engl. *comission*), joka lisää sijoittajan kuluja kaupankäynnissä.

Tämä tutkielma keskittyy tarkastelemaan osakemarkkinoilla tapahtuvia arvonmuutoksia ja

niiden ennustettavuutta aikaisemman kurssihistorian perusteella.

2.3 Osakekurssien tehokkuus

Tehokkaiden markkinoiden hypoteesi esittää osakekurssien kuvastavan täydellisesti kaikkea saatavilla olevaa tietoa ja näin ollen kurssiliikkeet mallintavat satunnaiskulkua (Malkiel ja Fama 1970). Kyseisen hypoteesin paikkansa pitävyys on kuitenkin hyvin kiistelty ja yhtenä vastaesimerkkinä voidaan pitää Medallion fund -hedgerahastoa, joka on saavuttanut 66 % keskimääräisen vuosituoton vuosien 1988-2018 aikana (Cornell 2020). Rahaston tiedetään käyttävän kvantitatiivisia sijoitusstrategioita, jotka hyödyntävät muiden lähteiden ohella kurssidatasta löytyneitä anomalioita, joita tehokkaiden markkinoiden hypoteesin mukaan ei tulisi olla. Näiden anomalioiden olemassaolo puoltaisi käyttäytymistieteellisen rahoituksen näkemystä, siitä etteivät markkinatoimijat ole aina täysin rationaalisia (Benartzi ja Thaler 1995).

Tämän tutkielman näkökulmasta voidaan osakekurssien liikkeen ajatella muistuttavan stokastista prosessia, josta voi löytyä kohinan seasta ennustettavia liikkeitä eli anomalioita.

2.4 Tekninen analyysi

Nazário ym. 2017 esittää erääksi teknisen analyysin määritelmäksi kokoelman välineitä, joilla ennustetaan tulevia kohde-etuuden tuottoja hyödyntäen historiallista markkinadataa, erityisesti kohde-etuuden hintaa ja kaupankäynnin volyymia. Analyysimenetelmän toimivuutta perustellaan pohjautumisella käyttäytymistieteelliseen rahoitukseen ja siihen etteivät sijoittajat toimi aina rationaalisesti. Menetelmän toimivuus on kiistelty ja yksiselitteistä näyttöä sen toimivuudesta ei ole todettu (Nazário ym. 2017).

Menetelmä jakautuu yleisesti erilaisiin indikaattoreihin, jotka ovat matemaattisia funktioita kurssidatalle ja kurssiliikkeiden visuaaliseen tarkasteluun, josta voidaan pyrkiä havaitsemaan toistuvia malleja (engl. *patterns*).

2.5 Kynttilätikut

Japanilaiset kynttilätikut (engl. *japanese candlesticks*) ovat eräs tapa esittää osakekursseja, jossa aikayksikön sisällä tapahtuneista kurssiliikkeistä ilmaistaan avaus- ja sulkemishinta sekä aikayksikön sisällä olleet matalin ja korkein hinta (Lu 2014). Tästä menetelmästä voidaan myös käyttää nimitystä OHLC tai OHLCV (engl. *open-high-low-close-volume)*, jossa O on aikayksikön avaushinta, H on aikayksikön korkein hinta, L on aikayksikön matalin hinta, C on aikayksikön sulkemishinta ja V aikayksikön sisällä ollut volyymi eli vaihdettujen osakkeiden määrä.

Menetelmä on suosittu osakekurssien visualisointiin ja tallentamiseen, mutta on myös oleellinen osa teknisen analyysin visuaalisessa tarkastelussa. OHLC aikasarjamuotoisen kurssidatan rakenne voidaan esittää seuraavanlaisesti:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} & x_{1,5} \\ x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} & x_{2,5} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i,1} & x_{i,2} & x_{i,3} & x_{i,4} & x_{i,5} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,3} & x_{n,4} & x_{n,5} \end{bmatrix}, \tag{2.1}$$

siten että:

$x_{i,1} = O_i$

$x_{i,2} = H_i$

$x_{i,3} = L_i$

$x_{i,4} = C_i$

$x_{i,5} = V_i$,

missä jokainen matriisin rivi $i \in \{1, 2, \cdots, n\}$, kuvastaa yhtä aikasarjan aikayksikköä ensimmäisestä jäsenestä 1 viimeiseen jäseneen n. Matriisin sarakkeet on määritelty seuraavasti:

O_i=Aikayksikön avaushinta

H_i=Aikayksikön korkein hinta

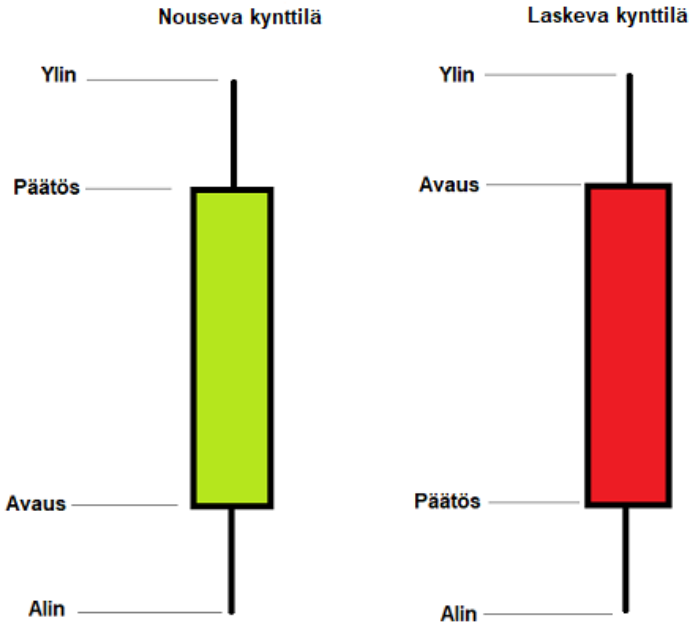
L_i=Aikayksikön matalin hinta

C_i=Aikayksikön sulkemishinta

V_i=Aikayksikön volyymi

Kuvassa 1 esitetään visuaalinen esitystapa kynttilätikuille.

Kuvio 1. Nouseva ja laskeva kynttilä (Leskinen 2019)



2.6 Kernelit

Kernelit (engl. *kernels*) ovat funktioita, joilla kaksiulotteisen aineiston xy-koordinaatistossa olevien pisteiden väliset suhdeluvut voidaan esittää korkeammassa dimensiossa. Aineistolle, jossa kahteen kategoriaan luokiteltavia pisteitä ei voida erottaa toisistaan lineaarisesti, voi löytyä lineaarinen ratkaisu korkeammassa dimensiossa. Kerneleiden avulla voidaan siis etsiä aineistolle lineaarista ratkaisua, jota ei alkuperäisessä xy-koordinaatistossa löydy. Tästä menetelmästä käytetään nimitystä "Kernel-temppu"(engl. *kernel trick*) tai Mercerin teoreema (Suykens ym. 2003). Yleisiä kernelifunktioita ovat lineaariset, polynomiset, sekä radiaaliset kernelifunktiot. Kernelifunktion optimaalinen valinta koneoppimisessa riippuu aineistosta ja käytettävästä menetelmästä.

2.7 Tukivektorikone

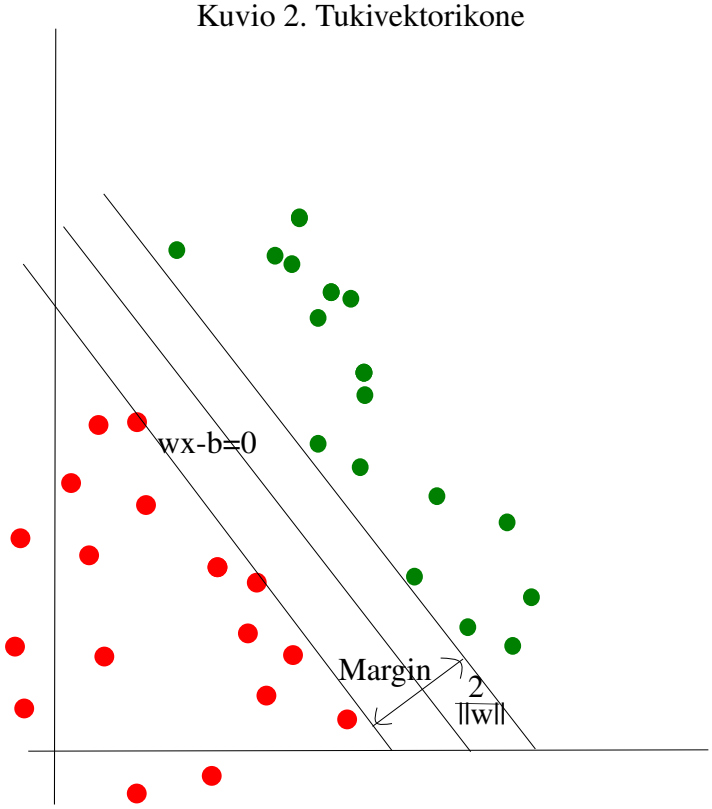
Tukivektorikone on Vladimir Vapnikin esittelemä ohjattu koneoppimismalli (Vapnik 1999). Tyypiltään kyseessä on lineaarinen luokittelumalli, jota voidaan käyttää luokitteluongelmissa. Malli toimii käyttämällä tukivektoreita, jotka luokittelevat aineiston kahteen eri luokkaan erottamalla ne lineaarisesti toisistaan.

Tukivektorikone toimii myös luokitteluongelmiin, jotka eivät ole suoraan lineaarisia. Tähän käytetäänn luvussa 2.6 esiteltyjä kerneleitä. Oikean kernelifunktion valinta vaikuttaa mallin tuloksiin ja valinta riippuu käytettävästä aineistosta. Yleinen valinta tukivektorikoneen kernelifunktiolle on radiaalinen kernelifunktio (engl. *radial basis function, RBF*) (Liu, Shen ja Wang 2014).

Tukivektorikone käyttää siis kerneleitä mallintamaan aineiston pisteiden välisiä suhteita toisessa dimensiossa, jossa tukivektorit voivat löytää lineaarisen luokittelun aineistolle, mahdollistaen luokittelun niin lineaarisille, kuin ei-lineaarisille luokitteluongelmille. Tukivektorikone sallii aineiston lineaarisessa jaottelussa virhemarginaalin, joka parantaa mallin toleranssia virheluokitteluille, tehden siitä hyvin soveltuvan aineistoon, jossa on mukana kohinaa. Virhetoleranssi vähentää myös riskiä ylisovittamisesta (engl. *overfitting*), jossa tietyn pisteen jälkeen malli ei enää paranna kykyään ratkaista kyseinen ongelma vaan alkaa poimia aineistosta muuta kohinaa (Jabbar ja Khan 2015). Ylisovittaminen saa tulokset näyttämään hyvältä mallin sovittamisvaiheessa, mutta sen kyky ratkaista kyseinen ongelma on tuloksia heikompi, joka johtaa heikompiin tuloksiin myös uudella aineistolla.

Tukivektorikoneen soveltuminen on aikaisemmissa tutkimuksissa ollut vastavirta-algoritmilla (engl. *backpropagation*) toimivaa neuroverkkoa parempi osakekurssien ennustamisessa (Tay ja Cao 2001). Kuvassa 2 esitetään havainnollistus tukivektorikoneesta, joka luokittelee aineiston kahteen eri kategoriaan.

Kuvio 2. Tukivektorikone



3 Tutkimuskysymys

Tässä tutkielmassa esitellään uusi tukivektorikoneen luokitteluongelma osakemarkkinoiden ennustamiseen. Aikaisempi tutkimus koneoppimisen käyttämisestä osakemarkkinoiden ennustamiseen on keskittynyt ominaisuuksien valintaan (engl. *feature engineering*) tai koneoppimisen algoritmin valintaan. Vähemmän tutkittu aihe on mallin ennustetyypin kuten luokitteluongelman valinnan merkitys toteutettavan mallin tulokseen.

Luvussa 5 toteutetaan kaksi tukivektorikonetta samoilla hyperparametreilla ja syötteillä, mutta eri luokitteluongelmilla. Aikaisemmissa tutkimuksissa suosittua luokitteluongelmaa verrataan tässä tutkielmassa esiteltävään keskihajontasuhteutettuun luokitteluongelmaan.

Tutkimuskysymys jakautuu seuraavaan pääkysymykseen ja sen alakysymyksiin:

1. Vaikuttaako tukivektorikoneen luokitteluongelman valinta osakekurssien ennustamiskykyyn?

    (a) Voiko tukivektorikone löytää historiallisesta kurssikehityksestä anomalioita vai seuraako osakekurssit satunnaiskulkua?
    (b) Miten iso vaikutus kaupankäyntikuluilla on toteutettuihin malleihin?

Toteutettavaa mallia verrataan vanhan luokitteluongelman lisäksi myös vertailuindeksiin, joka on tässä tutkielmassa käytettävä kohde-etuus, jonka kurssiliikkeitä pyritään ennustamaan. Vertailut toteutetaan kaupankäyntikuluja mallintaen sekä ilman, jolla havainnollistetaan kaupankäynnin määrän vaikutus tuloksiin.

Hypoteesina keskihajontasuhteutetun luokitteluongelman valinnalle on osakekurssien stokastinen luonne. Pienien kurssiliikkeiden ennustaminen saattaa altistua satunnaiselle kohinalle kurssiliikkeissä enemmän kuin keskihajontaa isompien liikkeiden ennustaminen. Näin ollen mahdolliset kurssihistoriasta löytyvät anomaliat eivät välttämättä esiinny, jos luokitteluongelma on valittu väärin.

Mahdolliset luokitteluongelman valinnan vaikutukset voivat paljastaa tarvetta jatkotutkimukselle, jos aikaisemmissa tutkimuksissa painotetut luokitteluongelmien tyypit eivät osoittaudu

johtavan tasaisiin tuloksiin muiden luokitteluongelmien kanssa. Käytettävät luokitteluongelmat esitellään luvussa 5.3.

4 Aikaisempi tutkimus

Tässä luvussa käsitellään aikaisempia tutkimuksia, joissa koneoppimismalleja sekä erityisesti tukivektorikonetta on hyödynnetty osakekurssien ennustamiseen. Tutkimusten valinta on rajattu vain niihin, joissa käytetään tukivektorikonetta tai tukivektoriregressiota ja historiallista hintadataa osakkeiden tai osakeindeksien liikkeistä aineistona. Muita aineistoja käyttävät tutkimukset kuten yritysten taloudelliset tiedot, markkinasentimentti tai muut tekijät ovat rajattu pois, sillä ne ovat tämän tutkielman kohteen ulkopuolella, joka rajoittuu pelkästään kurssihistorian hyödyntämiseen.

Aikaisemmista tutkimuksista tarkastellaan ominaisuuksien valintaa ja käytettyjä luokitteluongelmia. Läpikäydyistä tutkimuksista muodostetaan yleisnäkemys siitä, miten aineistoa on aikaisemmin esikäsitelty ja mitä malleilla on pyritty ennustamaan. Näiden pohjalta suoritetaan ominaisuuksien valinta luvussa 5.2 tämän tutkielman mallille.

Taulukossa 1 esitellään kirjallisuuskatsaukseen sisältyvät tutkimukset, niissä käytetyt algoritmit sekä ennustetyyppi. Riippumaton komponenttianalyysi (engl. *independent component analysis)* on lyhennetty taulukossa merkinnällä *ICA*, geneettiset algoritmit (engl. *genetic algorithms*) merkinnällä *GA* ja K:n lähimmän naapurin menetelmä (engl. *K-nearest neighbors*) merkinnällä *KNN*.

Taulukko 1. Aikaisemmat tutkimukset

Tutkimus	Algoritmi	Ennustetyyppi
Tay ja Cao 2001	Tukivektorikone	Suunta
Chen ja Hao 2018	Tukivektorikone	Suunta
Cao ja Tay 2003	Tukivektorikone	Suunta
Huang, Nakamori ja Wang 2005	Tukivektorikone	Suunta
Kim 2003	Tukivektorikone	Suunta
Ahmadi ym. 2018	Tukivektorikone	Suunta
Henrique, Sobreiro ja Kimura 2018	Tukivektoriregressio	Regressio
Ince ja Trafalis 2017	Tukitektorikone + ICA	Suunta
Choudhry ja Garg 2008	Tukivektorikone + GA	Suunta
Żbikowski 2015	Tukivektorikone	Suunta
Kumar, Meghwani ja Thakur 2016	Tukivektorikone	Suunta
Nayak, Mishra ja Rath 2015	Tukivektorikone + KNN	Suunta

4.1 Ominaisuuksien valinta

Historiallista kurssidataa on useissa tutkimuksissa esikäsitelty erilaisilla tavoilla. Kim 2003 hyödynsi teknisen analyysin indikaattoreita kurssihistorian esikäsittelyssä, jolla voidaan pyrkiä pehmentämään kurssiliikkeissä esiintyvää kohinaa ja antaa mallin katsoa tämän ohitse muihin tekijöihin. Vastaavasti (Tay ja Cao 2001) esikäsitteli kurssiliikkeet hyödyntäen liukuvaa keskiarvoa sekä hintamuutosprosenttia eri aikaväleiltä. Tutkimuksissa käytetyistä teknisen analyysin indikaattoreista suurin osa hyödyntää pelkästään sulkemishintoja ja näin ollen mahdolliset anomaliat, jotka liittyisivät päivän sisällä olleeseen avaushintaan, korkeimpaan tai matalimpaan hintaan jäävät tässä menetelmässä huomaamatta.

Teknisen analyysin indikaattorit ja erilaiset tilastolliset arvot ovat olleet käytössä lähes kaikissa aineistona olevissa tutkimuksissa, mutta valitut indikaattorit vaihtuvat tutkimusten välillä. Chen ja Hao 2018 esitti tutkimuksessaan, että aikaisemmissa tutkimuksissa erilaisten indikaattoreiden käyttämisen on todettu parantavan mallien tarkkuutta. Aineiston pohjalta voidaan havaita tämän menettelyn ominaisuuksien valinnassa olevan normalisoitunut käytäntö aihealueen tutkimuksissa. Rajallisen aineiston määrän takia on kurssiliikkeistä saata-

va poistettua kohinaa ja näin ollen parannettava mallin saaman syötteen laatua. Tarkkoja valintoja indikaattoreille on haastava eri tutkimusten pohjalta muodostaa, sillä poikkeavien aineistojen takia eivät tutkimusten tulokset ole suoraan verrattavissa toisiinsa.

Toinen yleinen lähestymistapa on ollut (Ahmadi ym. 2018) käyttämä esikäsittely, jossa hyödynnetään luvussa 2.5 kuvattuja kynttilätikkuja, jotka oli normalisoitu suhdeluvuiksi yhden päivän sisällä. Tämä menetelmä tuo mukanaan kurssiliikkeissä olevan kohinan selkeästi mukanaan, mutta välttää tilanteen, jossa liiallinen esikäsittely saattaisi pudottaa tietoa pois, jota tukivektorikone pystyisin hyödyntämään ennusteiden tekemiseen. Menetelmän ongelmana muodostuu, että syötettä ei voida laajentaa katsomaan kauas historiaan sillä liian laaja syötteiden määrä hankaloittaisi mallin sovittamista, koska aineiston määrä on rajallinen.

Chen ja Hao 2018 hyödynsi tutkimuksessaan molempia aikaisemmin mainittuja menetelmiä. OHLC-mallisten kynttilätikkujen lisäksi syötteeseen valittiin teknisen analyysiin indikaattoreita pohjaten valinnan aikaisempaan tutkimukseen, jossa on osoitettu kyseisten indikaattoreiden kuvastaneen oleellisia elementtejä kurssiliikkeestä. Valinta oli perusteltu poiketen edellä mainituista tutkimuksista, joissa syötteen esikäsittelyn valinnalle ei ole esitetty vahvoja perusteita. Kyseisen valinnan haasteeksi voi silti muodostua runsas syötteiden määrä, joka voi hidastaa tai hankaloittaa mallin oppimista, koska aineiston määrä on rajallinen.

Cao ja Tay 2003 käytti 5, 10, 15 ja 20 päivien hintamuutoksia prosenteissa sekä 100-päivän eksponentiaalista liukuvaa keskiarvoa mallin syötteenä, joka ennusti kohde-etuuden liikettä seuraavan viiden päivän aikana. Hintamuutosten käyttäminen on tehokas tapa normalisoida hintadataa, mutta tämä jättää huomioimatta päivän sisällä tapahtuneet hintaliikkeet sekä kaupankäynnin volyymin. Liukuva keskiarvo auttaa tuomaan mallille pidemmän aikavälin kokonaiskuvaa kompaktissa muodossa. Kaupankäynnin volyymia hyödynnettiin Żbikowski 2015 tutkimuksessa, jossa painotettiin vahvasti teknisen analyysin indikaattoreita ja erityisesti erilaisia oskillaattoreita. Kaupankäynnin volyymin huomioimiseen käytettiin OBV-indikaattoria (engl. *on-balance volume*).

Henrique, Sobreiro ja Kimura 2018 toteuttamassa tukivektoriregressioon pohjautuvassa mallissa käytettiin yleiseen tapaan indikaattoreita, kuten liukuvaa keskiarvoa sekä erilaisia oskillaattori-indikaattoreita. Poikkeavana indikaattorina oli käytössä keskimääräinen todellinen alue (engl.

average true range), joka mittaa kurssiliikkeiden volatiliteettia, huomioiden myös päivän sisäiset korkeimmat ja alimmat hinnat.

Huang, Nakamori ja Wang 2005 toteuttama malli ennusti kohde-etuuden suuntaa hyödyntämällä osakeindeksin ja valuuttakurssin normalisoitua hintamuutosta. Teknisen analyysin indikaattoreita tai päivän sisäisiä liikkeitä ei hyödynnetty mallissa, joka tekee mallin saamista syötteistä suppeat. Toteutunut malli yltää oleellisesti satunnaiskulkua parempaan ennusteeseen, mutta tutkimuksessa ei eritellä kuinka isoja toteutuneet liikkeet kohde-etuudessa ovat olleet ja malli ottaa kantaa vain liikkeen suuntaan, ei sen määrään.

Aikaisemmissa tutkimuksissa teknisen analyysin indikaattoreita on käytetty laajasti, ja niistä liukuvat keskiarvot ovat olleet suosituimpia. Chen ja Hao 2018 esitti tutkimuksessaan, että lyhyen liukuvan keskiarvon käyttäminen auttaa normalisoimaan ja poistamaan kohinaa kurssidatasta, joka olisi auttanut parantamaan tuloksia. Tämä tukee näkemystä esikäsittelymenetelmien valinnan tärkeydestä. Kurssihistorian määrä on rajallinen ja markkinoiden dynamiikan voidaan olettaa muuttuneen kaupankäyntikulujen laskiessa elektronisten kaupankäyntimahdollisuuksien myötä sekä automatisoidun kaupankäynnin kasvamisen myötä, joka on lisännyt markkinoiden tehokkuutta (Manahov, Hudson ja Gebka 2014).

Rajallinen aineiston määrä tekee ominaisuuksien valinnasta tärkeän osan koko mallia, sillä liiallinen syötteiden määrä, jotka kuvaavat samoja asioita tai muuten ovat kohinaa, voivat tehdä mallin sovittamisesta mahdotonta. Tähän ongelmaan on esitetty niin sanottuja hybriditoteutuksia, joissa malli koostuu ensimmäisessä vaiheessa käytettävästä erillisestä algoritmista, joka on vastuussa ominaisuuksien valinnasta, joita käytetään koneoppimismallissa. Choudhry ja Garg 2008 toteutti hybridimallin, jossa hyödynnettiin geneettisiä algoritmeja ominaisuuksien valintaan ja tukivektorikonetta luomaan ennusteita näiden valintojen pohjalta. Ince ja Trafalis 2017 toteutti myöhemmin vastaavalla rakenteella mallin, jossa geneettisten algoritmien sijaan käytettiin riippumatonta komponenttianalyysia ominaisuuksien valintaan. Hybriditoteutusta tutki myös (Kumar, Meghwani ja Thakur 2016), jossa käytettävissä olevista syötteistä löytyi laajasti erilaisia indikaattoreita ja useita eri menetelmiä hyödynnettiin näiden valitsemisessa. Tutkimuksissa hybriditoteutukset tuottivat parempia tuloksia, kuin pelkkä tukivektorikone ilman ominaisuuksien valinnan optimointia. Tulosten pohjalta voidaan havaita ominaisuuksien valinnan tärkeys ja hybriditoteutusten lupaavuus, joiden

haastavuutena voi kuitenkin olla riski mallin ylisovittamisesta.

Erityisesti (Kumar, Meghwani ja Thakur 2016) tutkimus sisälsi tuloksia eri indikaattoreiden hyödyllisyydestä, sen perusteella miten usein ne tulivat eri malleilla valituksi. Eniten valittujen syötteiden joukossa oli aikaisemmin esitetty keskimääräinen todellinen alue (ATR), OHLC-hinnat kyseisenä päivänä sekä liukuvat keskiarvot. Hintamuutokset ja keskiarvot muodostuvat oleellisiksi valinnoiksi, mutta volatiliteettia mittaava ATR-indikaattori on tuloksissa oleellinen havainto. Liian monimutkaiset indikaattorit eivät välttämättä tarjoa hyödyllistä esikäsittelyä aineistolle, ja perusominaisuudet vaikuttavat olevan tärkeämpiä tekijöitä.

Aikaisempien tutkimusten ominaisuuksien valintaa vertaillessa ovat OHLC-kurssiliikkeiden käyttäminen sekä erilaiset indikaattorit olleet laajasti käytössä. Erityisesti indikaattoreissa ovat erilaiset liukuvat keskiarvot sekä oskillaattorit olleet käytetyimpiä. Indikaattoreiden tuoma kurssidatan normalisointi ja kohinan poistaminen ovat osoittautuneet tutkimuksissa hyödyllisiksi, mutta eri vaihtoehtojen määrän takia täytyy osa rajata pois. Useat indikaattorit mallintavat samoja asioita, mutta eri tavalla, jolloin näiden valinta ei tuo uutta informaatiota malliin. Liiallinen indikaattoreiden määrä on ominaisuuksien valinnan ja kohinan poistamisen tarkoitusta vastaan, joka voi vaikeuttaa tukivektorikoneen sovittamista. Aikaisempien tutkimusten pohjalta vaikuttavat volatiliteetin mittaaminen, liukuvat keskiarvot, oskillaattorit sekä OHLC-kurssidata olevan kategorioina lupaavia, joista jokainen kannattaa tuoda mukaan ominaisuuksien valintaan.

4.2 Luokitteluongelmat

Tukivektorikoneen käyttäminen osakemarkkinoiden liikkeiden mallintamiseen vaatii kurssiliikkeiden ennustamisen mallintamista luokitteluongelmana (engl. *classification problem*). Taulukossa 1 havaitaan, että aikaisemmat tutkimukset ovat keskittyneet muodostamaan luokitteluongelman kohde-etuuden suunnan ennustamisena. Kyseiseen valintaan liittyy puute, sillä pelkästään suunnan ennustaminen ei välttämättä riitä tuottojen tekemiseen, sillä malli ei ota kantaa tulevan liikkeen määrästä.

Yli 50 % tarkkuus ennusteissa ei suunnan ennustamisessa riitä, jos ennusteen ollessa vää-

rässä, on kohde-etuuden liike paljon suurempi, kuin sen ollessa oikeassa. Tämä johtaa toteutuvien tappioiden olevan aina voittoja suuremmat. Huomioimatta kaupankäynnin kuluja on odotusarvo mahdollista määritellä seuraavanlaisesti, jossa w=voiton määrä, l=tappion määrä (negatiivinen), h=todennäköisyys voitolle $f(w,l,h) = wh + l(1-h)$

Voittoa tekevän mallin täytyisi täyttää ehto $f(w,l,h) > 0$, johon tutkimusten luokitteluongelmat eivät ota kantaa. Poikkeavana tapauksena on regressiomallit kuten (Henrique, Sobreiro ja Kimura 2018) toteuttama tukivektoriregressio. Suuntaa ennustavat luokitteluongelmat voivat silti antaa tuloksia markkinoiden tehokkuuteen liittyen, sillä tehokkaiden markkinoiden hypoteesin heikkojen ehtojen perusteella historiallisen kurssidatan ei pitäisi mahdollistaa satunnaiskulkua parempia ennustuksia tulevasta (Malkiel ja Fama 1970), mutta ne eivät välttämättä ole optimaalisia löytämään pidemmän aikavälin anomalioita.

Kim 2003 toteuttama malli luokittelee ongelman kohde-etuuden suunnan ennustamiseksi, määrittäen laskeeko vai nouseeko kohde seuraavana päivänä ($N+1$). Vastaavanlainen toteutus oli (Tay ja Cao 2001), jossa malli ennusti suuntaa viiden päivän päähän ($N+5$). Eri ennusteiden aikavälejä oli verrattu (Ahmadi ym. 2018), jossa vertailtiin eri malleja ja niiden ennusteita yhden ja kuuden päivän kurssiliikkeen suunnasta.

Chen ja Hao 2018 tutkimus muutti aikaisemmin suosittua luokitteluongelmaa ja toteutettava malli loi signaaleita, joiden perusteella tuli joko myydä tai ostaa kohde-etuutta. Kyseinen luokitteluongelma tekee mallista käytännönläheisemmän kaupankäynnille, mutta koska malli ei ota kantaa tavoitehintoihin tai millaista kurssiliikkeen määrä se odottaa niin saatava toteutus on lähellä mustaa laatikkoa, joka hankaloittaisi riskienhallinnan toteuttamista.

Henrique, Sobreiro ja Kimura 2018 toteuttama regressiomalli tarjoaa etunaan tarkan hinnan ennustamisen, joka ratkaisee suuntaa ennustavien luokitteluongelmien puutteen. Tarkan tulevan hinnan ennustamista voidaan pitää ideaalisimpana ennustetyyppinä malleilla, mutta haasteeksi nousee sen haastavuus. Osakekurssien sisältäessä kohinaa ja mahdollisesti mallintaen satunnaiskulkua vaihtelevissa määrin, on regressiomallien ennustetyypin toteuttaminen tarkasti haastavaa.

Mahdolliset kurssiliikkeestä löytyvät anomaliat voivat löytyä, kun pyritään erottelemaan isompia kurssiliikkeitä ja ohittamaan satunnainen kohina. Optimaalinen ennustetyyppi voi

löytyä regressiomallin ja suuntaa ennustavan luokitteluongelman väliltä, jolla ohitetaan kohinan vaikutus ennusteeseen, mutta otetaan kantaa myös liikkeen määrään. Tarkan hinnan ennustaminen ei ole välttämätöntä, jotta kurssiliikkeiden ennustamisesta voidaan hyötyä, jos ennusteen tuotto-odotuksen tiedetään olevan positiivinen.

Käytetty luokitteluongelma kurssiliikkeen suunnan ennustamisesta tietyllä aikavälillä on ollut aikaisemmissa tutkimuksissa laajasti käytössä erilaisten regressiomallien ohella. Tämä tuo esille yhden mahdollisen puutteen aikaisemmissa tutkimuksissa, sillä valittu luokitteluongelma voi olla hyvin tärkeä osa toteutettavaa mallia eikä etukäteen voida tietää varmaksi mitä niistä on mahdollista mallintaa satunnaista arvausta paremmin. Aikaisempi tutkimus on rajautunut hyvin samanlaisiin luokitteluongelmiin, jotka jättävät tarpeen uusien luokitteluongelmien tutkimiselle.

Koneoppimiseen löytyy laajasti valmiita ohjelmointikirjastoja, joissa näiden mallien toteutus on valmiina käytettävissä sekä monessa tapauksessa myös mallien hyperparametrit ovat myös automaattisesti optimoitavissa. Tämä jättää vähemmän painoa tekniselle toteutukselle, sillä kynnys valmiiden mallien käyttämiselle on madaltunut vuosien aikana. Jos kurssiliikkeistä löytyy anomalioita niin näiden löytyminen ja hyödyntäminen aktiivisessa kaupankäynnissä tulisi poistaa anomalian olemassaolo ja lisätä markkinoiden tehokkuutta. Oleellisiksi haasteiksi jää jäljelle aineiston esikäsittely eli ominaisuuksien valinta sekä miten mallin tuloksia käytetään, joka liittyy oleellisesti tässä luvussa käsiteltyyn luokitteluongelman valintaan.

Pelkästään suuntaa ennustavaan luokitteluongelmaan liittyy ongelmana, myös edellä esitetty havainto siitä, että tuotto-odotusten tai mahdollisten riskien mallintamista ei kyseisen luokitteluongelman perusteella voida tarkasti suorittaa, sillä se ei ota kantaa liikkeen määrään. Liikkeen määrää voidaan pyrkiä mallintamaan kurssiliikkeen keskihajonnalla tai luvussa 4.1 kuvatulla ja osassa aikaisempia tutkimuksia käytetyllä keskimääräisellä todellisella alueella (ATR).

Aikaisemmissa tutkimuksissa ei luokitteluongelman valinnalle ole esitetty kattavia perusteita eikä sen valintaan ole esitelty perusteltua prosessia, vaan tutkimukset ovat keskittyneet enemmän muiden osa-alueiden ympärille kuten ominaisuuksien valintaan tai käytettävään

koneoppimismalliin. Aikaisemman tutkimusnäytön puuttuessa luokitteluongelman valinnasta on aikaisemmin käytettyjen luokitteluongelmien haastaminen perusteltua ja tarpeellista.

5 Toteutus

Tässä luvussa kuvataan tämän tutkielman tukivektorikonemallien toteutus. Tukivektorikoneen toteutus muodostuu aineistosta, ominaisuuksien valinnasta, luokitteluongelmasta, tukivektorikoneen toteuttamisesta ja sovittamisesta. Näiden lisäksi toteutus sisältää taustatestausalgoritmin, jota käytetään kaupankäynnin simulointiin historiallisella kurssidatalla, mallintaen miten tukivektorikoneeseen pohjaava kaupankäyntistrategia olisi suoriutunut.

5.1 Aineisto

Toteutettava malli käyttää aineistonaan historiallista kurssidataa osakeindeksistä DAX, joka koostuu 30 eri saksalaisesta osakkeesta. Valinta osakeindeksille yksittäisen yrityksen sijaan pohjataan saatavilla olevan historiallisen datan määrään, yhtiökohtaisten uutisten tai riskien minimointiin sekä selviytymisharhan minimointiin (engl. *survivorship bias*). Kurssihistoriaa tarkastellaan päivätasolla, sillä päivän sisäisen kurssidatan saatavuus ilmaiseksi on rajattua.

Kyseisen indeksin valinta muiden vaihtoehtojen sijaan tehdään runsaan historiallisen datan saatavuuden takia sekä maantieteellisen sijainnin suhteen, joka voisi mahdollistaa tulosten skaalautuvan helpommin myös muihin euroalueen maihin. Käytetty aineisto on päivätason kurssihistoriaa ajanjaksolta [1.1.2000 - 1.4.2021] ja se on saatu yahoo finance -palvelusta. Aineisto jaetaan kahteen osaan, joista ensimmäistä käytetään tukivektorikoneen sovittamiseen ja siihen varattu kurssihistorian määrä on 70 % aineistosta. Loput 30 % varataan sovitetun mallin testaamiseen ja tulosten analysointiin.

Valinta korkealle testaamiseen varatulle aineistolle on pohjattu aineiston luonteeseen. Ylisovittamista, jossa malli oppisi kuvaamaan aineistossa olevaa kohinaa eikä mahdollisia systemaattisia anomalioita kurssiliikkeissä halutaan pyrkiä välttämään. Mahdollinen ylisovittaminen johtaisi hyviin tuloksiin historiallisella datalla testatessa, mutta riski sen toimimattomuudesta tulevaisuudessa olisi hyvin korkea. Pidempi testaamiselle varattava aineisto auttaa tämän havaitsemisessa ja välttämisessä sekä antaa tilastollisesti merkittävämpiä tuloksia.

Kuvassa 3 on visualisoitu DAX-indeksin kurssihistoria käytetyltä ajanjaksolta kynttiläkaa-

viona luvun 2.5 mukaisesti, siten että yksi kynttilä kuvastaa yhtä päivää. Mallin sovittamiseen ja testaamiseen käytettävät aineistot on visualisoitu vastaavasti kuvissa 4 ja 5. Käsiteltävän aineiston kynttiläluonnetta havainnollistetaan kuvassa 6. Visuaalisesti tarkastellen sovittamiseen käytettävässä aineistossa kurssiliike on muodostunut enemmän pitkäkestoisista trendeistä ylös ja alas, kun mallin testaamiseen valittu aineisto sisältään enemmän heiluntaa. Testausaineiston loppupuolella on havaittavissa COVID-19 pandemiasta 2020 keväällä seurannut historiallisen nopea pörssiromahdus. Eroavaisuudet aineistojen välillä sekä nopea pörssiromahdus testausaineistossa tuovat toivottavaa eroavuutta, joka auttaa havaitsemaan herkemmin mahdollisen ylisovittamisen mallissa.

Kuvio 3. DAX päivätason kurssihistoria 3.1.2000-1.4.2021

Kuvio 4. DAX harjoitusaineisto päivätason kurssihistoria 3.1.2000-11.5.2006

Kuvio 5. DAX Testausaineisto päivätason kurssihistoria 4.11.2014-1.4.2021

Kuvio 6. DAX havainnollistus päivätason kurssihistoria 20.2.2018-25.9.2018

5.2 Ominaisuuksien suunnittelu

Toteutettavan tukivektorikonemallin ominaisuuksien suunnittelulla pyritään esikäsittelemään aineisto siten, että kohinan määrä mallin saamista syötteistä on minimoitu ja vain oleelliset ominaisuudet ovat saatavilla. Onnistunut ominaisuuksien suunnittelu nopeuttaa mallin sovittamista ja auttaa saavuttamaan paremmat tulokset, jos saatavilla olevan aineiston määrä on rajallinen. Valitut ominaisuudet on pohjattu luvussa 4 käsiteltyyn aikaisempaan tutkimukseen ja niissä tehtyihin valintoihin aineiston esikäsittelyn suhteen sekä tämän tutkimuksen mallin vaatimusten suhteen. Valitut ominaisuudet pidetään yksinkertaisena eikä hybriditoteutuksissa käsiteltyjä menetelmiä ominaisuuksien valintaan hyödynnetä, sillä tutkielma keskittyy valitun luokitteluongelman vaikutuksien vertailuun.

Ensimmäinen osio on luvussa 2.5 kuvatun OLHCV-mallisen suoran kurssihistorian sisällyttäminen ja käsittely. Suoran indeksin pistelukujen sijaan halutaan kyseinen kurssihistoria normalisoida. Päivän sisällä tapahtuneet liikkeet normalisoidaan kyseisen päivän avaushinnan suhteen ja päivän avaushinta sekä sulkemishinta normalisoidaan edellisen päivän sulkemishintaan nähden. Kaupankäynnin volyymin normalisoidaan edellisen päivän volyymiin. Normalisoinnit toteutetaan seuraavanlaisesti:

$$ON_i = \frac{O_i}{C_{i-1}} - 1$$

$$HN_i = \frac{H_i}{O_i} - 1$$

$$LN_i = \frac{L_i}{O_i} - 1$$

$$CN_i = \frac{C_i}{C_{i-1}} - 1$$

$$VN_i = \frac{V_i}{V_{i-1}} - 1,$$

missä $i \in \{2,3,\cdots,n\}$, n=aikasarjan viimeisen aikayksikön indeksi ja
O_i=Aikayksikön avaushinta
H_i=Aikayksikön korkein hinta
L_i=Aikayksikön matalin hinta
C_i=Aikyksikön sulkemishinta
V_i=Kaupankäynnin volyymi

ON_i=Normalisoitu avaushinnan muutos

ON_i=Normalisoitu avaushinnan muutos

HN_i=Normalisoitu aikayksikön korkeimman hinnan muutos

LN_i=Normalisoitu aikayksikön matalimman hinnan muutos

CN_i=Normalisoitu sulkemishinnan muutos

VN_i=Normalisoitu volyymin muutos

ON_i=Normalisoitu avaushinnan muutos

Suorien hintaliikkeiden lisäksi aineistoa esikäsitellään ja mukaan tuodaan tiettyjä aikaisemmassa tutkimuksessa käytettyjä teknisen analyysin indikaattoreita. Indikaattorien arvot on laskettu suhteessa edellä esitettyyn normalisoituihin päivien sulkemishintoihin $C_i, i \in \{1,2,\cdots,n\}$

Liukuva keskiarvo (engl. *moving average*) kuvaa viimeisen n päivän kurssin keskiarvoa:

$$MA_i(C,n) = \frac{\sum_{j=0}^{n-1} C_{n-j}}{n}, i \geq n, \tag{5.1}$$

missä C sisältää edelle kuvatut kohde-etuuden sulkemishinnat, i kuvastaa tiettyä päivää ja n päivien määrää, joista lasketaan keskiarvo.

Keskiarvo normalisoidaan suhteessa tarkastelupäivän i sulkemishintaan muodostaen normalisoitu liukuva keskiarvo seuraavanlaisesti:

$$NMA_i(C,n) = \frac{MA_i(C,n)}{C_i} - 1, i \geq n, \tag{5.2}$$

missä $MA_i(n)$ on edellä kuvattu liukuva keskiarvo, C kohde-etuuden sulkemishinnat, n liukuvan keskiarvon pituus ja i aikasarjan yksittäisen päivän indeksi.

Eksponentiaalinen liukuva keskiarvo (engl. *exponential moving average*) on liukuva keskiarvo, joka antaa enemmän painoa uusimmille datapisteille ja näin ollen reagoi nopeammin muutoksiin kuin tavallinen liukuva keskiarvo $MA_i(n)$. Indikaattori määritellään seuraavanlaisesti:

$$EMA_i(C,n) = \begin{cases} C_i, & \text{jos } n = 1 \\ \frac{2}{n+1} * C_i + \frac{n-1}{n+1} * EMA_{i-1}(n-1), & \text{jos } n > 1 \end{cases}, \tag{5.3}$$

missä C on kohde-etuuden sulkemishinnat, n liukuvan keskiarvon pituus ja i aikasarjan yksittäisen päivän indeksi.

Indikaattori normalisoidaan suhteessa tarkastelupäivän i sulkemishintaan:

$$NEMA_i(C,n) = \frac{EMA_i(C,n)}{C_i} - 1, \tag{5.4}$$

missä C on kohde-etuuden sulkemishinnat, $EMA_i(C,n)$ edellä kuvattu eksponentiaalinen liukuva keskiarvo, n liukuvan keskiarvon pituus ja i aikasarjan yksittäisen päivän indeksi.

Liukuva keskihajonta (engl. *rolling standard deviation*) näyttää viimeisen n päivän keskihajonnan:

$$STDEV_i(n) = \sqrt{\frac{\sum_{j=0}^{n-1}(MA_i(CN,n) - CN_{i-j})^2}{n-1}}, \tag{5.5}$$

missä CN on kohde-etuuden normalisoidut sulkemishinnat, $MA_i(CN,n)$ edellä kuvattu liukuva keskiarvo, n liukuvan keskiarvon pituus ja i aikasarjan yksittäisen päivän indeksi.

Suhteellinen vahvuusindeksi (engl. *relative strength index*):

$$U_i = \begin{cases} C_i - C_{i-1}, & \text{jos } C_i > C_{i-1} \\ 0, & \text{jos } C_i \leq 0 \end{cases}$$

$$D_i = \begin{cases} C_i - C_{i-1}, & \text{jos } C_i < C_{i-1} \\ 0, & \text{jos } C_i \geq 0 \end{cases}$$

$$RSI_i(n) = 100 - \frac{100}{1 + \frac{EMA_i(U,n)}{EMA_i(D,n)}}, RSI_i \in [0,100], \tag{5.6}$$

missä C kuvaa kohde-etuuden sulkemishintoja, n on suhteellisen vahvuusindeksin pituus ja i aikasarjan yksittäisen päivän indeksi.

Indikaattorin keskiarvo skaalataan alaspäin lähemmäksi muiden normalisoitujen indikaattoreiden arvoja:

$$NRSI_i(n) = RSI_i(n) - 50, NRSI \in [-50,50], \tag{5.7}$$

missä $RSI_i(n)$ on edellä kuvattu suhteellinen vahvuusindeksi, n on indikaattorin pituus ja i aikasarjan yksittäisen päivän indeksi.

Edellä esitetyistä esikäsittely ja normalisointi menetelmistä muodostetaan seuraavat syötteet tukivektorikoneelle:

- Viimeisimmän päivän OHLCV kurssiliikkeet
- NMA(C, 20)
- NEMA(C, 5)
- STDEV(20)
- NRSI(14)

Syötteistä muodostetaan aikasarjan muotoinen matriisi luvun 2.5 tapaisesti, mutta vaihtaen sarakkeiden arvot esikäsittelyn mukaisiin ominaisuuksiin:

$$A = \begin{bmatrix} ON_{20} & HN_{20} & LN_{20} & CN_{20} & VN_{20} & NMA_{20} & NEMA_{20} & NRSI_{20} & STDEV_{20} \\ ON_{21} & HN_{21} & LN_{21} & CN_{21} & VN_{21} & NMA_{21} & NEMA_{21} & NRSI_{21} & STDEV_{21} \\ \vdots & \vdots & \vdots & \vdots & \vdots & & & & \\ ON_i & HN_i & LN_i & CN_i & VN_i & NMA_i & NEMA_i & NRSI_i & STDEV_i \\ \vdots & \vdots & \vdots & \vdots & \vdots & & & & \\ ON_n & HN_n & LN_n & CN_n & VN_n & NMA_n & NEMA_n & NRSI_n & STDEV_n \end{bmatrix} \tag{5.8}$$

$$\overrightarrow{a_i} = \begin{bmatrix} ON_i & HN_i & LN_i & CN_i & VN_i & NMA_i & NEMA_i & NRSI_i & STDEV_i \end{bmatrix}, \tag{5.9}$$

missä jokainen matriisin rivi kuvaa yhtä päivää, siten että 1=aikasarjan ensimmäinen päivä, n=aikasarjan viimeinen päivä ja i=aikasarjan yleinen päivä $1 \leq i \leq n$. Matriisin ensimmäinen rivi alkaa indeksistä 20, sillä kaikkia sarakkeita kuten normalisoitua liukuvaa keskiarvoa NMA_{20} ei voida määritellä, kun $i < 20$. $\overrightarrow{a_i}$ viittaa yksittäiseen matriisin riviin eli eri ominaisuuksien arvoihin yksittäisenä päivänä. Matriisin sarakkeet ovat tässä kappaleessa määritellyt ominaisuudet.

5.3 Luokitteluongelma

Luvussa 4.2 havaittiin aikaisempia tutkimuksia tarkastelemalla, että tutkimukset ovat keskittyneet oleellisesti testaamaan uusia koneoppimisen algoritmeja sekä ominaisuuksien valinnan parantamista esimerkiksi hybriditoteutuksilla, joissa ominaisuuksien valintaan voidaan käyttää menetelmiä kuten geneettisiä algoritmeja tai riippumatonta komponenttianalyysia. Malkiel 2003 mukaisesti markkinoiden tulisi korjata löytyneet hinta-anomaliat, sillä uuden anomalian löytyessä sitä aletaan käyttämään aktiivisessa kaupankäynnissä, joka vaikuttaa

hintaan, kunnes anomaliaa ei enää esiinny. Tämän perusteella voidaan ajatella, että laajan kirjastojen määrän ja koneoppimismallien implementoinnin helppouden takia, ei anomalioita pitäisi koneoppimisella löytyä helposti. Mahdolliset anomaliat voisivat löytyä parantamalla käytettyjä syötteitä ja ominaisuuksien valintaa, josta on toteutettu useita tutkimuksia. Vähemmän tutkittu on kolmas kohde, joka on mallien ennustamisen kohde, kuten tukivektorikoneen luokitteluongelma.

Aikaisempien tutkimusten malleissa yleiset koneoppimismallien ulostulot ovat joko luokitteluongelmat kohde-etuuden kurssimuutoksen suunnalle tietyllä aikavälillä, kuten seuraavalle päivälle $N+1$, tai regressiotyyliset mallit, joilla pyritään ennustamaan tarkkaa hintaa. Pelkästään suuntaa ennustavissa malleissa on ongelmana luvussa 4.2 esitetty ongelma, että malli ei ota kantaa liikkeen määrään vaan pelkästään suuntaan. Kyseinen luokittelu on riittävä haastamaan hypoteesia kurssiliikkeiden satunnaiskulusta, mutta ei ole riittävä aktiiviseen kaupankäyntiin, sillä malli ei ota kantaa tuotto-odotukseen tai riskiin.

Tässä tutkielmassa esiteltävä uusi luokitteluongelma on keskihajontasuhteutettu luokitteluongelma (engl. *standard deviation adjusted classification problem*. Luokitteluongelma jaottelee datapisteet sen perusteella saavuttaako kohde-etuus seuraavana $STDEV_i(20)$ kokoisen muutoksen ylös vai alas $[C_i - STDEV_i(20), C_i + STDEV_i(20)]$. Tämä luokitteluongelma ottaa suoraan kantaa liikkeen määrän, jolloin riski ja tuotto-odotus ennusteessa on määritelty. Kyseinen luokittelu voi myös helpottaa uusien anomalioiden tunnistamisen kurssiliikkeestä, sillä pienet liikkeet jätetään huomioimatta ja ennustetta skaalataan kuvaamaan isompaa kuvaa. Tämä helpottaa kohinan välttämistä ja voi tarjota parempia ennusteita.

Toteutettavaa luokitteluongelmaa verrataan aikaisemmin suosittuun $N+1$ suunnan ennustamiseen. Tulokset auttavat tunnistamaan luokitteluongelman valinnan merkitystä toimivan mallin toteuttamisen kannalta, mikä on vähän tutkittu aihe osakemarkkinoiden osalta.

5.4 Tukivektorikoneen toteuttaminen

Toteutettavat mallit ovat tukivektorikoneita luvussa 5.2 esitetyllä syötteellä. Mallit eroavat ainoastaan luokitteluongelmiltaan, joista ensimmäiseen valitaan luvussa 5.3 esitetty keskihajontasuhteutettu luokitteluongelma. Kyseistä luokitteluongelmaa verrataan aikaisemmissa

tutkimuksissa laajasti käytettyyn $N+1$ päivän suunnan ennustamiseen. Aineisto ladataan ja esikäsitellään luvun 5.1 mukaisesti. Tukivektoriluokittelumalliin toteutukseen käytetään sklearn-kirjaston tukivektoriluokittelijaa (engl. *support vector classification, SVC*), joka sovitetaan harjoitteluaineistoon valiten kernelifunktioksi radiaalinen kernelifunktio. Hyperparametrit pidetään oletusarvoinaan. Käytetyn ohjelmointikielen sekä käytettyjen kirjastojen versiot on esitetty taulukossa 2.

Taulukko 2. Ohjelmointikieli ja -kirjastot

Nimi	**Versio**
Python	3.8.3
numpy	1.20.2
pandas	1.2.3
sklearn	0.24.1
ta	0.7.0
pandas_datareader	0.9.0

Mallien sovittamisen jälkeen verrataan tukivektorikoneita toisiinsa sekä vakiofunktioon, joka kuvaa tilannetta, jossa kohde-etuutta ostetaan aikajakson alussa eikä myydä ollenkaan. Tilastollisesti merkittävän mallin pitäisi poiketa oleellisesti tuloksiltaan vakiofunktiosta. Sovitettuja tukivektorikoneita sekä vertailukohtana olevaa vakiofunktiota voidaan kuvata seuraavanlaisesti:

$$f(\overrightarrow{a_i}) = \begin{cases} 0 \\ 1 \end{cases} \tag{5.10}$$

$$g(\overrightarrow{a_i}) = 1, \tag{5.11}$$

missä $\overrightarrow{a_i}$ sisältää luvussa 5.2 kuvatut tukivektorikoneen saamat syötteet yksittäisenä päivänä. Ulostulo 1 kuvaa tilannetta, jossa kurssin odotetaan nousevan ja 0 tilannetta, jossa kurssin odotetaan laskevan.

Vertailu vakiofunktioon $g(\overrightarrow{a_i})$ tapahtuu vertaamalla ennustusten osumatarkkuutta (engl. *hit-rate*), jossa verrataan oikein ennustettuja tapauksia kaikkien ennustusten kokonaismäärään:

$$H(x,y) = x/(x+y), \tag{5.12}$$

missä x on oikein ennustettujen tapausten määrä ja y väärin ennustettujen tapausten määrä.

5.5 Simuloitavat strategiat ja taustatestaus

Luvussa 4.2 esitettiin, että hyödyllisen mallin tulisi tuottaa positiivinen tuotto-odotus, kun osumatarkkuus suhteutetaan mahdolliseen tappioon ja voittoon. Tätä varten kahden tukivektorikoneiden ennustusten pohjalta rakennetaan kaksi yksinkertaista kaupankäyntistrategiaa, jota simuloidaan historiallisella datalla mallintaen aktiivista kaupankäyntiä taustatestauksella. Tukivektorikoneiden pohjalta käydyn kaupan tuloksia verrataan toisiinsa sekä vakiofunktioon. Ensimmäisessä mallissa käytetään luvussa 4.2 esiteltyä suosittua luokitteluongelmaa, jossa tukivektorikone ennustaa nouseeko vai laskeeko kohde-etuus seuraavana $N+1$ päivänä. Luokitteluongelma ei ota kantaa mahdollisen kurssiliikkeen määrään. Toinen malli on luvussa 5.3 esitelty keskihajontasuhteutettu luokitteluongelma.

Tukivektorikoneiden pohjalta muodostettavat strategiat operoivat ostamalla kohde-etuutta aina, kun tukivektorikone odottaa kurssin liikkuvan seuraavaksi ylös, eli kun $f(\overrightarrow{a_i}) = 1$. Ostettua kohde-etuutta pidetään, kunnes malli odottaa kurssin laskevan seuraavaksi $f(\overrightarrow{a_i}) = 0$. Ostettuja kohde-etuuksia ei pidetä, kunnes tukivektorikoneen ennustama liikkeen määrä toteutuu, vaan kun sen ennuste tulevasta muuttuu. Tämä pohjataan tuotto-odotukseen, joka muuttuu negatiiviseksi, kun malli ennustaa kurssin laskua, jolloin kohde-etuutta ei haluta pitää.

Taustatestauksessa voidaan mallintaa kaupankäyntiin liittyviä kuluja, mikä auttaa havainnollistamaan kaupankäynnin määrän vaikutusta tuottoihin. Positiiviset tulokset kulujen huomioinnin jälkeen eivät takaa voitollista mallia, mutta negatiiviset tulokset voivat vahvistaa nollahypoteesin, joka olettaa, ettei kyseistä mallia käyttävä strategia toimisi sellaisenaan käytännössä.

Tarkka kaupankäyntiin liittyvien kulujen mallintaminen on haastavaa, sillä sen tulisi sisältää

ainakin toimeksiannosta maksettava kulu välittäjälle, joka voi vaihdella välittäjän ja toimeksiantojen koon mukaan, osto- ja myyntitasojen erotus ja mahdollinen toimeksiannon vaikutus hintatasoon tai hintatason muuttuminen ennen kuin toimeksianto keretään suorittaa (engl. *slippage*). Edellä mainitut tekijät eivät myöskään ole vakioita vaan ne voivat vaihdella. Todellisten kulujen mallintamisen haastavuuden takia, niitä voidaan pitää erillisenä tutkimusaiheena eikä siihen liittyvät tulokset ole tämän tutkielman tutkimuskysymyksen kannalta välttämättömiä. Todellisten kulujen mallintaminen ja vaikutuksen arviointi esitettyyn strategiaan esitetään mahdolliseksi jatkotutkimuksen aiheeksi.

Strategioiden taustatestauksella muodostetaan tulokset, jotka ovat kumulatiiviset tuotot pääomalle, jolla strategia käy simulaation aikana kauppaa. Tässä tutkielmassa käytetty taustatestaus määritellään seuraavanlaisesti:

$$b(s,k,a_n,c_n,i) = \begin{cases} 1, \textbf{ jos } i = 1 \\ 0, \textbf{ jos } f(a_i) = 0, f(a_{i-1}) = 0 \\ \frac{c_i}{c_{i-1}} - k - s, \textbf{ jos } f(a_i) = 0, f(a_{i-1}) = 1 \\ \frac{c_i}{c_{i-1}}, \textbf{ jos } f(a_i) = 1, f(a_{i-1}) = 1 \\ -k, \textbf{ jos } f(a_i) = 1, f(a_{i-1}) = 0 \end{cases} \tag{5.13}$$

$$(E_n) = (b(s,k,a_1,c_1), b(s,k,a_2,c_2), \cdots, b(s,k,a_n,c_n)), \tag{5.14}$$

missä (E_n) on jono, joka sisältää kumulatiiviset tuotot taustatestauksessa, siten että E_1 on simulaation ensimmäinen päivä ja n viimeinen. $b(s,k,a_n,c_n,i)$ kuvaa yksittäisen päivän tuottoa, siten että

s=Osto- ja myyntitasojen ero $\frac{m}{o}$, missä o on sen hetken paras ostotarjous ja m paras myyntitarjous

k=kaupankäynnin kulu, joka maksetaan välittäjälle suoritettavasta myynti- tai ostotoimeksiannosta

a_n=Luvussa 5.2 kuvattu tukivektorikoneen saamat syötteet eri päivinä

c_n=Luvussa 2.5 esitetty kohde-etuuden sulkemishinnat

i=Sen hetkinen päivä, missä 1 on ensimmäinen päivä ja n viimeinen

Taustatestaus alkaa lähtötilanteesta, jossa kohde-etuutta omistetaan ja sen jälkeen myydään tai ostetaan tarvittaessa strategian mukaan. Myynneistä ja ostoista muodostuu kaupankäyntikulu, joka vähentää tuottoja, sekä myydessä vähennetään lisäksi osto- ja myyntitasojen välinen erotus. Kaupankäynnin aktiivisuus vaikuttaa näin ollen tuottoihin negatiivisesti.

Taustatestaukset suoritetaan viidelle eri tapaukselle. Ensimmäisessä kaupankäyntiin liittyvät kulut oletetaan nollaksi ja taustatestaus ajetaan strategialle, jossa tukivektorikoneen luokitteluongelma ennustaa seuraavan päivän suuntaa. Seuraava strategia käyttää tukivektorikonetta, joka hyödyntää keskihajonta suhteutettua luokitteluongelmaa. Toisessa tapauksessa taustatestaukset suoritetaan samoille tukivektorikonemalleille, mutta jokaisen toteutetun toimeksiannon kuluksi oletetaan $k = 0.2\%$ sekä myynti- ja ostotasojen erotukseksi oletetaan olevan $s = 0.1\%$. Näiden neljän taustatestin lisäksi ajetaan yksi taustatestaus vertailukohteena toimivalle vakiofunktiolle $g(a_i) = 1$, joka omistaa kohde-etuutta koko simulaation ajan eikä myy sitä kertaakaan. Vakiofunktion taustatestaus tarvitsee suorittaa vain kerran sillä kaupankäyntikulut tai niiden puuttuminen eivät vaikuta sen tuloksiin.

Tämän tutkielman taustatestauksen toteutuksessa on tuoton saaminen mahdollista ainoastaan kurssiliikkeen nousemisesta. Strategian aktiivisuus pyrkii siis käytännössä minimoimaan tappioita myymällä kohde-etuuden ja pitämällä pääoman käteisenä ennen tulevaa kurssilaskua. Laskevissa kurssiliikkeissä olisi myös mahdollista tehdä tuottoa myymällä kohde-etuutta lyhyeksi eli lainaamalla esimerkiksi osakkeita, joita kaupankävijä ei itse omista ja myymällä ne markkinalle. Kohde-etuuden laskiessa voidaan kyseiset osakkeet ostaa halvemmalla takaisin ja palauttaa ne lainaajalle, jolloin lyhyeksi myyjä tekee voittoa laskeneen kurssin verran. Toinen vaihtoehto olisi hyödyntää erilaisia johdannaisia kuten optioita, joilla voi hyötyä kohde-etuuden laskemisesta. Tässä taustatestauksen toteutuksessa rajataan lyhyeksi myyminen pois vaihtoehdoissa, sillä siihen liittyvät kaupankäyntikulut voivat vaihdella runsaasti kohde-etuudesta riippuen, eikä se ole välttämätöntä, kun halutaan mallintaa yleisesti strategian kykyä ennustaa nousevia ja laskevia kursseja.

5.5.1 Taustatestauksen tulosten mittaaminen

Suoritetusta taustatestauksesta saadaan tulokseksi pääoman tuottoja kuvaava jono (E_n). Tuloksista mitataan tämän tutkielman kannalta oleellisia tekijöitä kuten kokonaistuottoa, riskiä sekä riskikorjattua tuottoa. Kokonaistuotto määritellään tuottojen kumulatiivisena tulona:

$$S = \prod_{i=1}^{n} E_i, \tag{5.15}$$

missä E_i on pääoman tuotot eri päivinä.

Kokonaistuoton lisäksi tuottoja mallinnetaan päiväkohtaisten tuottojen keskiarvolla:

$$G = \frac{\sum_{i=1}^{n} E_i}{n} \tag{5.16}$$

Strategian sisältämää riskiä mallinnetaan päiväkohtaisten tuottojen keskihajonnalla:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (G - E_i)^2}{n-1}}, \tag{5.17}$$

missä G on edellä kuvattu päiväkohtaisten tuottojen keskiarvo.

Strategian riskikorjatun tuoton mittaamiseen käytetään Sharpen lukua. Sharpen luvun käyttötarkoituksena on mallintaa strategian tuottojen ja sen sisältämän riskin eli tuottojen keskihajonnan välistä suhdetta. Rationaalisen sijoittajan oletetaan haluavan maksimoida sijoitustensa tuotto-odotus hänen ottamansa riskin määrälle tai vaihtoehtoisesti minimoimaan riskin määrä valitsemalleen tuotto-odotukselle (Markowitz 1952).

Sharpen luku määritellään tuotto-odotuksen ja tuottojen keskihajonnan välisenä suhteena (Sharpe 1966). Sharpen luku määritellään seuraavanlaisesti:

$$S_\alpha = \frac{R_p - R_f}{\sigma_p}, \tag{5.18}$$

missä R_p on sijoitusten tuotto-odotus, R_f on markkinoilta saatavilla oleva riskitön tuotto, joka tyypillisesti on valtioiden velkakirjoista saatavat korot sekä σ_p on tuottojen keskihajonta. Tässä tutkielmassa riskitön tuotto oletetaan matalan korkotason ja tulosten yksinkertaistamisen takia nollaksi $R_f = 0$. Tuotto-odotus määritellään päiväkohtaisten tuottojen keskiarvona sekä riski σ_p on päiväkohtaisten tuottojen keskihajontana.

Taustatestauksen tuloksista Sharpen luku lasketaan seuraavasti:

$$S_\alpha = \frac{G}{\sigma} \tag{5.19}$$

Sharpen luku auttaa kuvaamaan strategian riskin ja tuoton välistä suhdetta eli riskikorjattua tuottoa. Tämä mahdollistaa paremman vertailtavuuden vakiofunktion tuloksiin sekä paremman analysoitavuuden ovatko mahdolliset vakiofunktiota paremmat tuotot peräisin tukivektorikoneen hyvästä kurssiliikkeen ennustamisesta, vai mahdollisesti strategian korkeammasta riskin ottamisesta.

Edellä kuvatuilla metriikoilla pyritään mallintamaan yksinkertaistetusti taustatestatun strategian tuotto- ja riskiprofiilia sekä riskikorjattua tuottoa, jota mallinnetaan Sharpen luvulla. Tuloksissa voidaan kiinnittää enemmän huomiota itse Sharpen lukuun eikä pelkästään riskien tai tuottojen vertailuun, sillä strategian riskiä ja tuottoa voidaan aina skaalata samassa suhteessa alaspäin, jos sille allokoidaan vähemmän pääomaa käytettäväksi. Vastaavasti tuotto-odotusta ja riskiä voidaan skaalata ylöspäin ottamalla käyttöön lainarahaa ja näin ollen allokoiden strategialle enemmän pääomaa, kuin mitä sijoittajalla itsellään olisi käytettävissä. Näin ollen voidaan riskikorjattua tuottoa pitää tärkeämpänä, kuin pelkkää tuotto-odotusta. Tuotto-odotuksen skaalaamisen vaikutus Sharpen lukuun pääomaa allokoimalla voidaan esittää seuraavasti, kun velkarahan oletetaan olevan ilmaista:

$$S_\alpha = \frac{LG}{L\sigma} = \frac{G}{\sigma}, \tag{5.20}$$

missä L on pääoman allokoinnin määrä siten että, $L > 1$ tarkoittaa velkarahan käyttämistä, joka kasvattaa riskiä ja tuotto-odotusta ja $L < 1$ pääoman vähentämistä, joka laskee riskiä ja tuotto-odotusta. Tästä huomataan, että tuotto-odotuksen ja riskin skaalaaminen pääoman määrää muuttamalla ei vaikuta Sharpen lukuun, jos lainaraha on ilmaista ja näin ollen riskikorjatut tuotot ovat tulosten kannalta merkityksellisempiä, kuin pelkkien tuottojen vertailu.

5.6 Tukivektorikoneen sovittaminen

Tukivektorikoneet sovitetaan luvussa 5.2 esitettyyn aineistoon käyttäen luvussa 5.1 kuvatulla tavalla 70% aineistosta mallin sovittamiseen ja loput 30% käytetään mallin testaamiseen. Mallien tarkkuuden tulisi olla lähellä toisiaan niin harjoitusaineistossa kuin testausai-

neistossa. Isot eroavaisuudet aineistojen välillä voivat indikoida mahdollista ylisovittamista, joka johtaa huonoihin tuloksiin harjoitusaineiston ulkopuolella kuten esitettiin luvussa 2.7. Taulukossa 3 on esitetty tämän tutkielman keskihajontasuhteutetulla luokitteluongelmalla rakennetun tukivektorikoneen tarkkuus verrattuna vakiofunktioon. Vastaavasti taulukossa 4 on esitetty vanhalla $N+1$ päivän suuntaa ennustavalla luokitteluongelmalla rakennettu tukivektorikone verrattuna vakiofunktioon.

Tukivektorikoneiden tarkkuudet ovat harjoitus- ja testausaineiston välillä lähellä toisiaan. Tämän pohjalta voidaan olettaa, että selkeätä ylisovittamista ei tukivektorikoneissa ole tapahtunut ja eroavaisuudet johtuvat eroavasta kurssiliikkeestä aineistojen välillä, joka oli myös visuaalisesti todettu luvussa 5.1. Molempien mallien tarkkuus on hyvin lähellä vakiofunktiota testausaineistossa. Seuraavan päivän suuntaa ennustavan luokitteluongelman kanssa tämän voidaan todeta, ettei malli ole löytänyt systemaattisesti anomalioita kurssiliikkeestä ja malli on vaikuttanut sovittaneen itsensä lähelle vakiofunktiota ennustaen lähes joka päivälle nousua. Tämä voidaan todeta luokitteluongelman pohjalta, joka ennustaa vain kiinteästi seuraavan päivän suuntaa ottamatta kantaa tulevan liikkeen määrään. Keskihajontasuhteutetun luokitteluongelman osalta tarkkuudet ennusteissa ovat myös lähellä vakiofunktiota, mutta ilman tietoa ennusteiden tuotoista ja tappioista ei pelkkä tarkkuus riitä havaitsemaan onko malli löytänyt aineistosta anomalioita.

Keskihajontasuhteutetun luokitteluongelman tarkkuus tippui testausaineistossa yhdellä prosenttiyksiköllä, kun vakiofunktion tarkkuus nousi noin parilla prosenttiyksiköllä. Muutokset tarkkuudessa voivat liittyä testausaineiston eroavaisuuteen harjoitusaineistossa. Mallin tulosten hyödyllisyyden toteaminen vaatii tarkemman analysoinnin taustatestauksen avulla, jossa sen tuotot sekä tappiot arvioidaan, joka suoritetaan luvussa 6.

Taulukko 3. Keskihajontasuhteutetun luokitteluongelman tarkkuus

	Tukivektorikone	**Vakiofunktio**
Harjoitusaineiston tarkkuus	56.05%	53.31%
Testausaineiston tarkkuus	55.04%	55.4%

Taulukko 4. $N+1$ päivän suunnan luokitteluongelman tarkkuus

	Tukivektorikone	**Vakiofunktio**
Harjoitusaineiston tarkkuus	53.09%	52.8%
Testausaineiston tarkkuus	53.08%	52.96%

6 Tulokset

Luvussa 5.6 todettiin harjoitus- ja testausaineistojen tulosten olevan lähellä toisiaan molemmissa tukivektorikoneissa, eikä ylisovittamista pidetä ilmeisenä. Vanha luokitteluongelma saavutti harjoitus- ja testausaineistossa alle prosenttiyksikön verran paremman tarkkuuden verrattuna vakiofunktioon, mutta tätä ei voida pitää merkittävänä. Tulos liittyy oleellisesti, että yksinkertaistetulla aineistolla, joka esiteltiin luvussa 5.2 ei kyseisellä luokitteluongelmalla tukivektorikone pystynyt löytämään selkeitä anomalioita aineistosta.

Keskihajontasuhteutetun luokitteluongelman kanssa tukivektorikone saavutti vajaan kolmen prosenttiyksikön verran vakiofunktiota paremman osumatarkkuuden kuin vakiofunktio, mutta testausaineistossa olivat tarkkuudet lähellä toisiaan ja vakiofunktio hieman tarkempi. Tarkkuudet olivat yleisesti tässä luokitteluongelmassa korkeampia, kuin vanhassa luokitteluongelmassa, joka liittyy pidempään aikaväliin ennusteissa ja osakemarkkinoiden taipumukseen nousta pitkällä välillä. Tällöin nousun ennustaminen on tarkempaa, kun ennusteen aikaväliä pidennetään.

Yleisesti osumatarkkuudet olivat maltillisia ja tarkempia tuloksia varten hyödynnetään aktiivisen kaupankäynnin taustatestauksesta saatavia tuloksia. Tulokset sisältävät molemmissa luokitteluongelmissa kuluttoman tapauksen, jossa kaikki kaupankäyntiin liittyvät kulut oletetaan nollaksi ja kulut huomioiva tapaus, jossa mallinnetaan luvussa 5.5 mainittuja kaupankäyntikulua myynti- ja ostotoimeksiannoista sekä osto- ja myyntitarjousten välistä erotusta. Kulujen ei oleteta olevan tarkkoja tai kuvaavan tarkasti todellisuutta, vaan niitä hyödynnetään havainnollistamaan kulujen yleistä vaikutusta toimeksiantojen määrän kasvaessa. Tämä mahdollistaa tutkimuskysymyksen alakysymykseen vastaamisen liittyen kaupankäyntikulujen vaikutuksesta tukivektorikonemallien pohjalta suoritettavaan aktiiviseen kaupankäyntiin. Markkinoiden tehokkuuden tai hypoteesin satunnaiskulusta osakekursseissa haastamista varten ei kuluilla ole tuloksissa merkitystä, mutta ne antavat muuten hyödyllisen vertauskohdan, kun arvioidaan strategian mahdollista käytännönläheisyyttä.

Taulukossa 5 on esitetty tukivektorikoneen tulokset tämän tutkielman uudella keskihajontasuhteutetulla luokitteluongelmalla kulut huomioiden ja ilman. Strategian tuottokuvaaja, joka

kuvaa kumulatiivisia tuottoja on esitelty ilman kuluja kuvassa 7 ja kulut huomioiden kuvassa 8. Vertailukohta DAX on simulaatiossa käytettävä kohde-etuus DAX-indeksi, jonka kaupankäyntiä mallinnetaan aikaisemmin mainitulla vakiofunktiolla. Vakiofunktion mallinnetaan pitävän kohde-etuutta koko simulaation ajan eikä kaupankäyntikuluja näin ollen muodostu. Vastaavasti taulukko 6 kuvaa samat tulokset käyttäen vanhaa luokitteluongelmaa tukivektorikoneessa. Kyseisen strategian tuottokuvaajat on esitelty ilman kuluja kuvassa 9 ja kulujen kanssa kuvassa 10.

Tuloksissa keskimääräinen tuotto ja tuottojen keskihajonta kuvaavat yhden päivän aikavälillä tapahtuneita muutoksia. Kokonaistuotto kuvaa kumulatiivisia tuottoja, jotka strategia saavuttaa taustatestauksen loppuun mennessä. Sharpen luku on skaalattu kertomalla sen päivätasolla saatu arvo luvulla $\sqrt{252}$. Normaalin kalenterivuoden aikana on noin 252 päivää jolloin pörssi on auki ja näin ollen, jos osakekurssien odotetaan mallintavan normaalijakaumaa niin Sharpen luku voidaan skaalata vuositasolle kertomalla se luvulla $\sqrt{252}$. Skaalaus on tehty lukujen vertailtavuuden helpottamista varten eikä niitä ole tarkoitus verrata vuositason kurssimuutoksista laskettuihin Sharpen lukuihin, sillä oletus osakekurssien kehityksen normaalijakaumasta ei pidä täysin paikkaansa (Lo 2002).

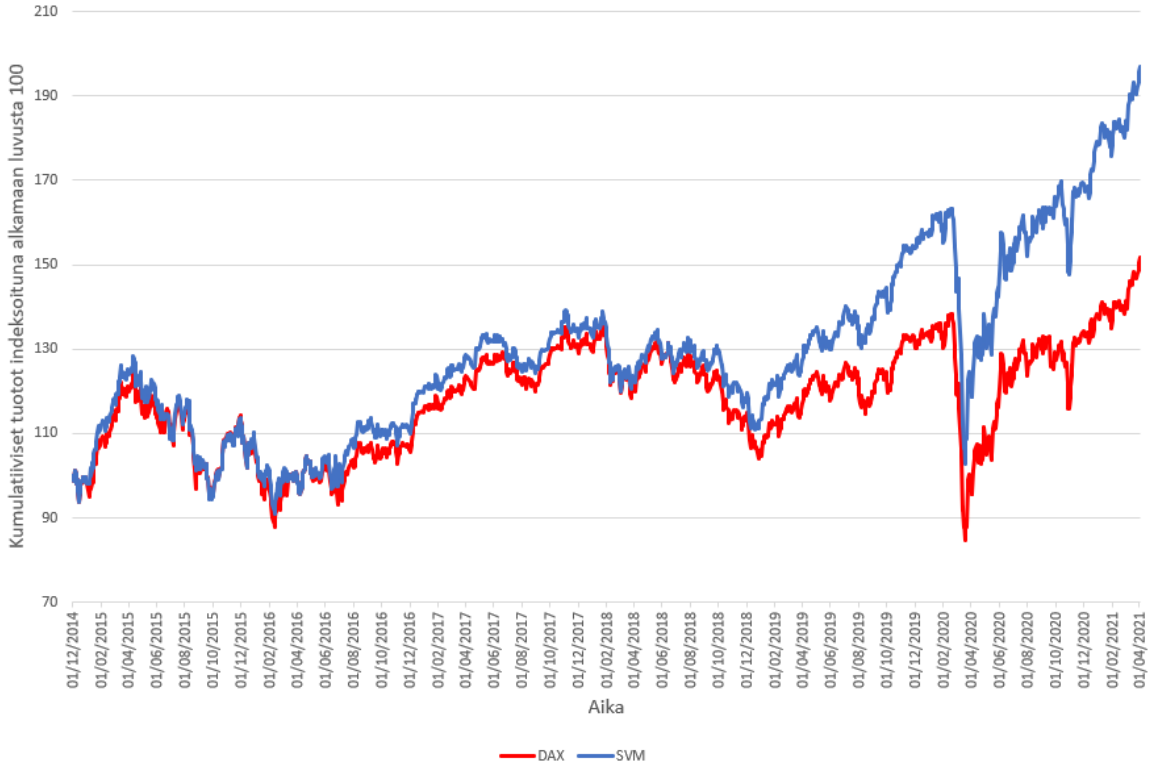
Taulukko 5. Keskihajontasuhteutetun luokitteluongelman tulokset

	Tukivektorikone kuluton	**Tukivektorikone kuluilla**	**DAX**
Keskimääräinen tuotto	0.05%	0.03%	0.03%
Tuottojen keskihajonta	1.26%	1.31%	1.31%
Sharpen luku	0.64	0.42	0.42
Kokonaistuotto	97.04%	30.73%	51.62%

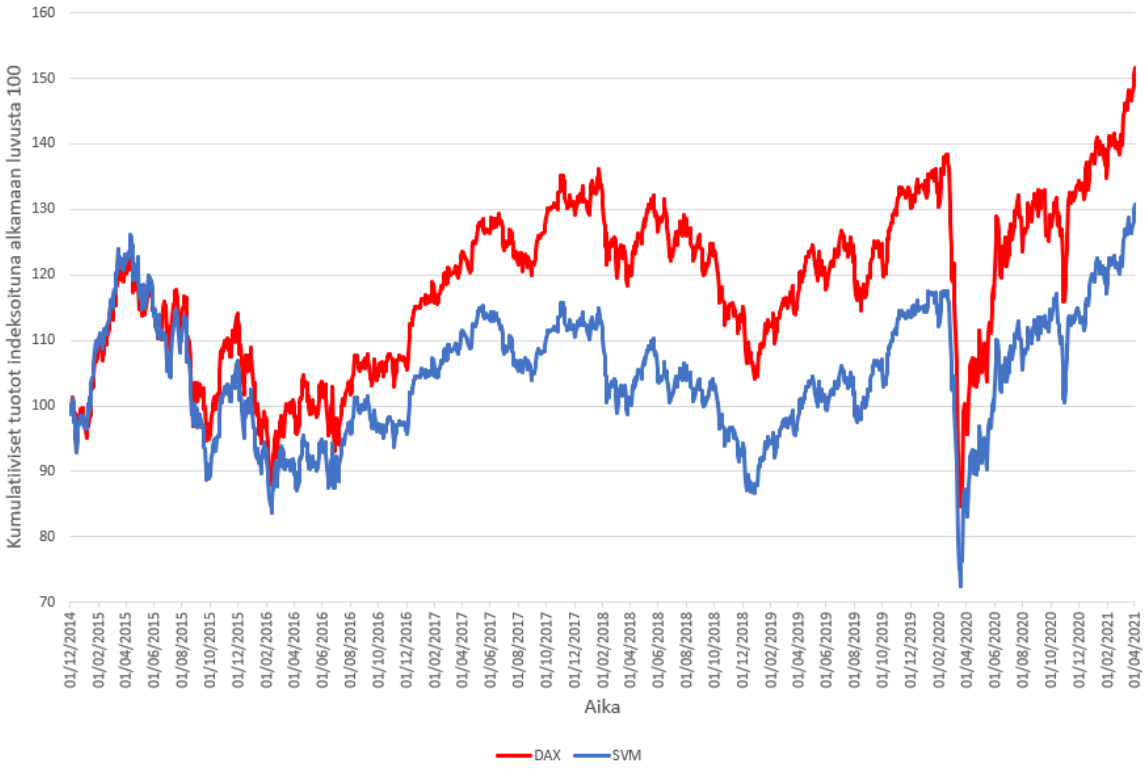
Taulukko 6. $N+1$ päivän suunnan luokitteluongelman tulokset

	Tukivektorikone kuluton	**Tukivektorikone kuluilla**	**DAX**
Keskimääräinen tuotto	0.04%	0.03%	0.03%
Tuottojen keskihajonta	1.13%	1.31%	1.31%
Sharpen luku	0.46	0.41	0.42
Kokonaistuotto	59.75%	48.6%	51.62%

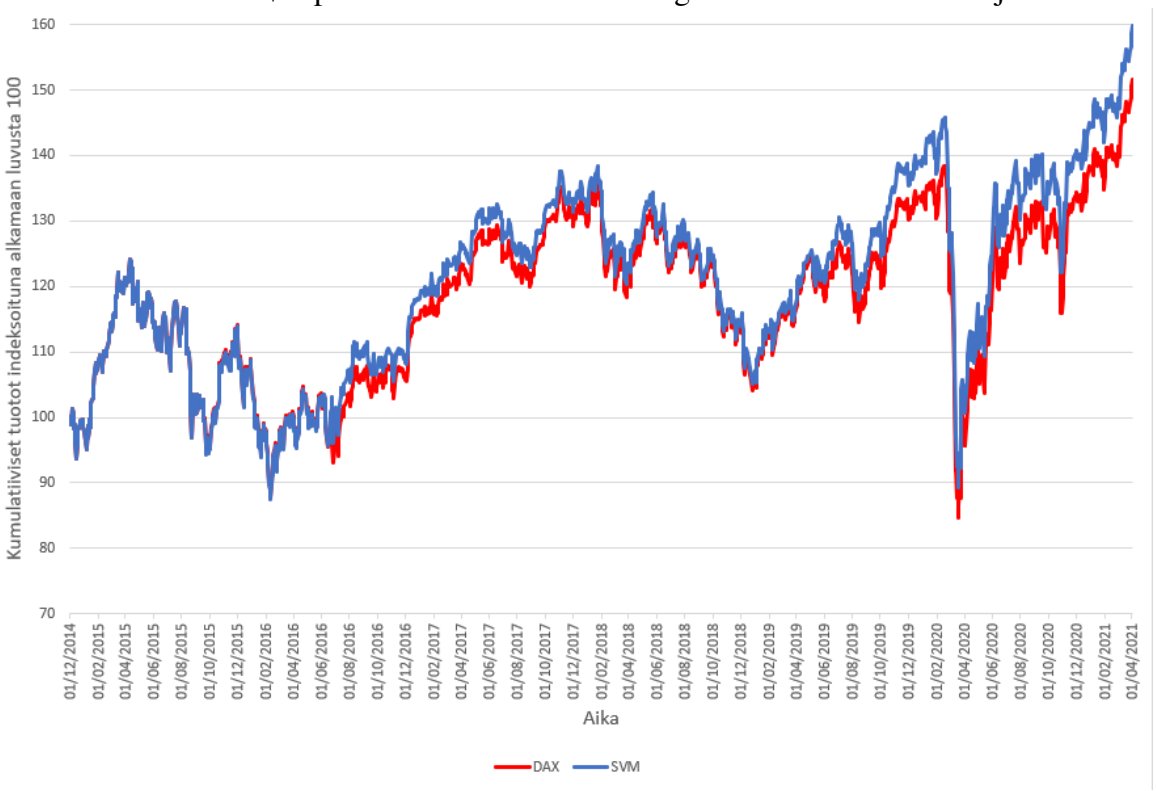
Kuvio 7. Keskihajontasuhteutetun luokitteluongelman tuotot ilman kuluja



Kuvio 8. Keskihajontasuhteutetun luokitteluongelman tuotot kulujen kanssa



Kuvio 9. $N + 1$ päivän suunnan luokitteluongelman tuotot ilman kuluja



Kuvio 10. $N+1$ päivän suunnan luokitteluongelman tuotot kulujen kanssa

Tarkastelemalla kuluttomia tuottoja molemmat luokitteluongelmat ovat tukivektorikoneessa tuottaneet vertailuindeksiä korkeamman kokonaistuoton kyseisellä aikavälillä. Seuraavan päivän suuntaa ennustavaa luokitteluongelmaa hyödyntämällä ovat keskimääräiset tuotot hieman vertailuindeksiä korkeammat, mutta tuottojen keskihajonta on pysynyt samana. Tämä on tuottanut vertailuindeksiä korkeamman Sharpen luvun ja riskikorjattuna saavutettu tuotto on näin ollen myös korkeampi. Kokonaistuoton määrä on kuitenkin hyvin lähellä vertailuindeksiä. Tuottokuvaajia 9 ja 10 tarkastelemalla huomataan, että strategian kumulatiiviset tuotot ovat koko aikavälin ajan hyvin lähellä vertailuindeksiä. Tämä kuvastaa mallin epävarmuutta luokittelusta ja se on suurimman osan ajasta operoinut kuten vakiofunktio ja odottanut kurssin nousevan. Kaupankäynnin määrä on näin ollen ollut matala, joka laskee positiivisen ylituoton tilastollista merkittävyyttä. Isoimmat poikkeamat vertailuindeksiin ovat tapahtuneet ennen ja jälkeen vuoden 2020 COVID-19 pandemian aiheuttamaa pörssiromahdusta, jolloin pörssikurssit ovat nousseet voimakkaasti ja volatiliteetti on yleisesti ollut

korkeaa. Mahdolliset anomaliat ovat näin ollen löytyneet normaalista poikkeavina aikoina, kun kurssiliikkeet ovat olleet voimakkaampia.

Tukivektorikone ei ollut yksinkertaistetuilla syötteillä ja seuraavan päivän kurssiliikkeen suuntaa ennustavalla luokitteluongelmalla saanut luotua hyvää mallia kurssiliikkeestä. Tulokseen liittyy varmasti syötteiden yksinkertaisuus, sillä tilastollisesti merkittäviin tuloksiin on aikaisemmissa tutkimuksissa päästy samalla luokitteluongelmalla. Pienten kurssiliikkeiden ennustaminen seuraavana päivänä voi osoittautua hankalaksi, sillä sen sisältämä kohina etenkin keskihajontaa pienemmissä liikkeissä muodostuu suureksi ja luokittelua on näin ollen vaikea muodostaa normaaleissa oloissa.

Keskihajontasuhteutetun luokitteluongelman kanssa tulokset ovat nousseet kokonaistuotossa yli 45 prosenttiyksikköä ja näin ollen liki kaksinkertaistunut vertailuindeksiin nähden. Oleellinen havainto on, että tuottojen noustessa on myös niiden keskihajonta laskenut eli strategian sisältämä riski on ollut myös vertailuindeksiä matalampi. Tämä olisi aktiivisessa kaupankäynnissä hyvin optimaalinen tilanne, jos tuottoja voidaan kasvattaa samalla kun riskiä lasketaan. Sharpen luku on näin ollen kasvanut yli 50 % vertailuindeksiin nähden ja tuloksia voidaan pitää tilastollisesti merkittävinä.

Tarkastelemalla tuottokuvaajaa 7 on strategia alkanut tuottamaan vertailuindeksiä enemmän 2018 vuoden loppupuolella. Visuaalisesti arvioiden nämä ylituoton muodostumiset ovat ajoittuneet isojen kurssilaskujen kuten COVID-19 pandemian aiheuttama pörssiromahduksen lähettyville, kun volatiliteetti on ollut korkeampaa tai osakekurssit laskeneet pidempään. Ylisuoriutuminen pidempään kestävissä selkeissä kurssilaskuissa tai volatiliteetin ollessa korkeammalla, voidaan mallin todeta löytäneen jotain ennustettavia anomalioita kyseisessä markkinaympäristössä. Tämä vastaa tutkimuskysymyksen alakysymykseen liittyen voiko tukivektorikone löytää anomalioita aikaisemmasta kurssikäytöksestä ja näin ollen havainnollistaen, ettei satunnaiskulku ole toteutunut täydellisesti ainakaan volatiliteetin ollessa markkinoilla korkeampaa. Hypoteesi tälle havainnolle voidaan esittää sijoittajien käyttäytymisestä, sillä markkinoiden laskiessa tai volatiliteetin kasvaessa voi pelkoon liittyvät tuntemukset kasvaa epävarmuuden tulevasta noustessa. Tämä voi ajaa sijoittajat tekemään epärationaalisia sijoituspäätöksiä muodostaen ennustettavaa kurssikäytöstä, joita tukivektorikone on pystynyt havaitsemaan ja hyödyntämään. Normaaleissa olosuhteissa ei anomalioita ole löytynyt

ja näin ollen voidaan kurssikäytöksen olettaa olevan tuolloin tehokkaampaa. Markkinoiden tehokkuus ei tästä näkökulmasta tarkastellen vaikuttaisi olevan vakio, vaan riippuvan markkinatilanteesta ja sen volatiliteetista.

Kulujen vaikutus keskihajontasuhteutettuun luokitteluongelmaan nähdään kuvasta 8, jossa strategia on alisuoriutunut vertailuindeksiin nähden lähes koko aikavälin ajan. Tähän vaikuttaa mallin epätarkkuus, kun markkinaympäristö on vähemmän volatiili ja kurssiliikkeet tasaisempia, jolloin malli ei pysty löytämään anomalioita ja näin ollen suoritetusta kaupankäynnistä muodostuu ylimääräinen kuluerä ilman siitä saatavaa hyötyä.

Vertailemalla luokitteluongelmia toisiinsa on keskihajontasuhteutettu luokitteluongelma tuottanut kurssiliikettä paremmin ennustavan mallin vanhaan $N+1$ päivän suuntaa ennustavaan luokitteluongelmaan verrattuna saavuttaen korkeamman tuoton matalammalla riskillä. Tukivektorikoneiden ominaisuuksien valinta ja hyperparametrit ovat olleet samat ja oleellinen eroavaisuus on ollut käytetty luokitteluongelma. Tulokset puoltavat luokitteluongelman vaikuttavan oleellisesti mallin saamiin tuloksiin ja tässä tutkielmassa esitelty uusi keskihajontasuhteutettu luokitteluongelma on kuvannut kurssiliikettä paremmin. Tutkimuskysymyksen pääkysymykseen luokitteluongelman merkityksestä voidaan todeta luokitteluongelman valinnalla olevan huomattavia vaikutuksia tukivektorikoneella saataviin tuloksiin ja näin ollen tuoden esille jatkotutkimuksen tarpeen eri luokitteluongelmien toimivuudesta ja valinnasta, joita ei aikaisemmin ole merkittävästi tutkittu.

Kaupankäyntikulut huomioiden ovat molemmat tukivektorikoneet hävinneet tuotoissa vertailuindeksille, mutta vanha luokitteluongelma vähemmän. Tämä on johtunut sen vähäisemmästä kaupankäynnin määrästä, joka on seurausta mallin epävarmuudesta luokitteluiden suhteen. Uudella luokitteluongelmalla ovat tulokset romahtaneet oleellisesti, vaikka ne ovat silti pysyneet positiivisena sekä keskihajonta samana kuin vertailuindeksillä. Tämä kuvastaa mallin sisältävän enemmän varmuutta mallinnettavasta kohteesta johtaen useampiin signaaleihin, jotka näin ollen kaupankäyntikulujen kautta ovat laskeneet tuottoja. Kumpikaan strategia ei näillä kaupankäyntikulujen oletuksilla olisi ollut vertailuindeksin omistamista kannattavampi, joka havainnollistaa, ettei toimiva malli riitä yksinään ylituottoa tekevään strategiaan. Tämä vastaa tutkimuskysymyksen alakysymykseen kaupankäyntikulujen merkityksestä, joita voidaan tulosten pohjalta pitää hyvin merkittävinä.

Keskihajontasuhteutetun luokitteluongelman tulokset ilman kuluja olivat tilastollisesti merkittäviä, joka nähdään taulukosta 5. Tulokset ovat ristiriidassa tehokkaiden markkinoiden hypoteesin kanssa, eivätkä tulokset puoltaisi kurssiliikkeen olleen aina satunnaiskulun mukaista, erityisesti markkinalaskujen lähettyvillä, jolloin volatiliteetti on korkeampaa. Luokitteluongelmassa keskihajonnan huomioimisen tuomat parannukset tuloksissa voivat liittyä siihen, että malli ohittaa paremmin satunnaisena pidettävän kohinan, jota on voinut keskittyä alle normaalin keskihajonnan oleviin kurssiliikkeisiin. Keskittymällä aikaisempaa keskihajontaa isompiin liikkeisiin, on malli voinut paremmin onnistua poimimaan isoihin liikkeisiin liittyviä anomalioita ja paremmin ohittamaan kohinaa.

Eroavaisuudet ennen ja jälkeen kaupankäyntikulujen vaikutusta havainnollistavat, että toimiva malli ei itsessään vielä tarkoita toimivaa ylituottoa tekevää strategiaa, vaan strategian muodostamista voidaan pitää omana tutkimuskohteenaan. Strategiaa pystyisi mahdollisesti parantamaan hyödyntämällä havaintoa sen paremmasta suoriutumisesta laskumarkkinoissa ja korkeamman volatiliteetin ympäristöissä, pyrkien rajaamaan aktiivinen näkemyksen ottaminen kyseisiin tilanteisiin. Vaihtoehtoisesti tukivektorikoneen sovittaminen siten että, se muodostaa todennäköisyyden luokittelun varmuudesta voisi mahdollistaa näkemyksen ottaminen vain niihin tilanteisiin, kun malli on varmempi ennusteestaan.

Yleisenä havaintona tuloksista voidaan todeta, että markkinoiden tehokkuutta ei tulisi olettaa vakioksi, vaan se voi vaihdella riippuen markkinaympäristöstä. Näin ollen anomalioita voi löytyä vähemmän, kun markkinat ovat vähemmän volatiliit ja sijoittajat rationaalisempia. Vastaavasti volatiliteetin noustessa ja COVID-19 aiheuttaman pörssiromahduksen kaltaiset tapaukset voivat tehdä markkinoiden liikkeistä ennustettavampaa ja lisätä sen tehottomuutta. Oleellinen haaste on myös markkinoiden jatkuva muuttuminen, joka aiheuttaa sen, etteivät löydetyt anomaliat ole ikuisia, vaan niitä hinnoitellaan pois, jos ne löytyvät ja se on mahdollista kaupankäyntiin liittyvien kulujen jälkeen. Kurssiliikkeiden todellinen luonne tuntuu olevan dynaamisempi ja monimutkaisempi kuin teoria antaisi olettaa. Todellinen käytös vaikuttaa olevan jossain satunnaiskulun ja behavioraalisen rahoituksen olettaman tehottomuuden välissä, riippuen markkinaympäristöstä.

Yksinkertaistettu ominaisuuksien valinta oli suunnattu yksinkertaistamaan luokitteluongelmien eroavaisuuksia ja tulosten perusteella isompien kurssiliikkeiden luokittelu onnistuu tu-

kivektorikoneelta paremmin. Aikaisemmissa tutkimuksissa esitetyt ominaisuuksien valinnat sekä erityisesti hybriditoteutukset näihin liittyen voisivat parantaa tuloksia vielä enemmän, kun luokitteluongelmana käytettäisiin keskihajontasuhteutettua luokitteluongelmaa.

7 Yhteenveto

Tässä tutkielmassa tutkittiin tukivektorikoneen soveltamista osakemarkkinoiden ennustamiseen mallintamalla DAX-osakeindeksin tulevia liikkeitä. Lyhyessä kirjallisuuskatsauksessa aikaisempaan tutkimukseen luvussa 4 havaittiin, että tutkimukset ovat keskittyneet testaamaan uusia koneoppimisen algoritmeja sekä ominaisuuksien valintaa. Tutkimuksissa käytetyt luokitettuongelmat ovat olleet hyvin rajallisia ja luvussa 4.2 esitettiin tulevan päivän kurssiliikkeen suunnan ennustavan luokitteluongelman, joka ei ota kantaa liikkeen määrään, olevan puutteellinen luokitteluongelma. Vähäisen aikaisempien tutkimusten määrän luokitteluongelmien valinnan vaikutuksiin liittyen, valittiin tässä tutkielmassa uusi keskihajontasuhteutettu luokitteluongelma, joka huomioi suunnan lisäksi myös kurssiliikkeen määrän. Liikkeen määrä pohjattiin aikaisempien päivien kurssiliikkeiden keskihajontaan pohjaten se hypoteesiin, että pienet kurssiliikkeet voivat sisältää enemmän satunnaista kohinaa sekä niitä voidaan pitää vähemmän merkittävinä kuin isoja kurssiliikkeitä sijoittamisen näkökulmasta. Uutta luokitteluongelmaa päätettiin verrata aikaisemmissa tutkimuksissa useasti käytettyyn suunnan ennustamiseen.

Aikaisempien tutkimusten pohjalta muodostettiin yksinkertainen ominaisuuksien valinta luvussa 5.2. Näiden valintojen pohjalta muodostettiin kaksi tukivektorikonetta luvussa 5.4 joiden luokitteluongelmat olivat erilaiset. Toisessa käytettiin aikaisemmin suosittua seuraavan päivän suunnan ennustamista ja toisessa tämän tutkielman esittelemää keskihajontasuhteutettua luokitteluongelmaa, joka huomioi liikkeen määrän ennustuksessaan. Tukivektorikoneiden ennustustarkkuus ei ollut tilastollisesti merkittävä, mutta luvut olivat lähellä toisiaan harjoitus- ja testausaineistoissa, joka lisäsi luottamusta, että malleja ei ollut ylisovitettu harjoitusaineistoon ja eroavaisuudet liittyivät aineistojen kurssiliikkeiden eroihin.

Mallien pohjalta rakennettuja kahta aktiivisen kaupankäynnin sijoitusstrategiaa joita, kuvattiin luvussa 5.5 taustatestattiin testausaineistolla. Tuloksissa kuvattiin keskimääräistä tuottoa, tuottojen keskihajontaa, Sharpen lukua sekä kokonaistuottoa. Ilman kaupankäyntikulujen huomiointia oli keskihajontasuhteutetun luokitteluongelman pohjalta rakennettu strategia merkittävästi parempi, kuin vanha seuraavan päivän suuntaa ennustava luokitteluongelma tai vertailuindeksin omistamista kuvaava vakiofunktio. Tutkimuskysymykseen luokitteluongel-

man vaikutuksesta tukivektorikoneen osakekurssien ennustamiskykyyn saatiin vastaus tuloksista. Tulokset osoittivat keskihajontasuhteutetun luokitteluongelman johtavan vakiofunktiota ja seuraavan päivän suuntaa ennustavaa luokitteluongelmaa huomattavasti parempiin riskikorjattuihin tuottoihin, kun kaupankäyntikuluja ei huomioida. Tämä perusteella voidaan jatkotutkimusta pitää tarpeellisena eri luokitteluongelmien tutkimiselle, joka on aikaisemmin ollut vähäistä. Jatkotutkimuksen kohteiksi esitetään tarvetta uusille luokitteluongelmille sekä aikaisemmissa tutkimuksissa esitettyjen ominaisuuksien valinnan toteuttamista uuden luokitteluongelman kanssa, sillä tässä tutkielmassa käytetty ominaisuuksien valinta oli tarkoituksenmukaisesti yksinkertaistettu.

Tulosten merkittävä eroavaisuus vertailuindeksin tuotosta haastaa myös tehokkaiden markkinoiden hypoteesin, jonka mukaisesti aikaisempien kurssiliikkeiden perusteella ei tulevan ennustaminen satunnaista arvausta paremmin pitäisi olla mahdollista. Tulosten pohjalta on perusteltua esittää, että kurssiliikkeet eivät aina noudata täydellisesti satunnaiskulkua, tai ole aikaisemmasta kurssihistoriasta riippumattomia, kun markkinoiden volatiliteetti kasvaa. Erityisesti keskihajontaa isompien liikkeiden ennustamisen esitetään olleen mahdollista satunnaista arvausta paremmin DAX-osakeindeksissä tässä tutkielmassa käsitellyllä aikavälillä. Kurssiliikkeiden anomaliat olivat keskittyneet ajanjaksoihin, kun markkinoiden volatiliteetti oli ollut korkeampi. Näiden tulosten valossa voidaan tutkimuskysymyksen alikysymykseen siitä että, pystyykö tukivektorikoneella löytämään historiallisesta kurssikehityksestä löytämään anomalioita, joka tämän tutkielman taustatestauksen valossa on ollut mahdollista.

Vastaus tutkimuskysymyksen alikysymykseen kaupankäyntikulujen vaikuttamisesta tukivektorikoneeseen pohjattuihin strategioihin saatiin, kun kaupankäyntikulut huomioitiin taustatestauksessa verraten kuluttomaan testitapaukseen. Molemmat aktiiviset strategiat alisuoriutuivat kokonaistuotoillaan verrattuna vakiofunktioon eli kohde-etuuden passiiviseen omistamiseen ilman aktiivista kaupankäyntiä. Tulos havainnollistaa strategian muodostamisen tärkeyttä, sillä pelkästään toimiva malli kurssiliikkeiden ennustamiseen ei olisi riittänyt ylituoton tekemiseen.

Laskentatehon kasvamisen myötä on koneoppiminen osoittautunut yhdeksi lupaavaksi tavaksi mallintaa osakekurssien liikkeitä. Näiden menetelmien saatavuuden ja helpon toteutettavuuden takia on perusteltua olettaa, että niiden avulla helposti löydettävien anomalioiden

tulisi poistua aktiivisen kaupankäynnin seurauksena eivätkä aikaisemmat tulokset takaa mallien toimivuutta myös tulevassa. Tämän perusteella myös tässä tutkielmassa esitellyt tulokset tuottojen suhteen ovat hypoteettisia eikä takeita mallin tarkkuudesta tulevaisuudessa ole.

Lähteet

Ahmadi, Elham, Milad Jasemi, Leslie Monplaisir, Mohammad Amin Nabavi, Armin Mahmoodi ja Pegah Amini Jam. 2018. “New efficient hybrid candlestick technical analysis model for stock market timing on the basis of the Support Vector Machine and Heuristic Algorithms of Imperialist Competition and Genetic”. *Expert Systems with Applications* 94:21–31.

Benartzi, Shlomo, ja Richard H Thaler. 1995. “Myopic loss aversion and the equity premium puzzle”. *The quarterly journal of Economics* 110 (1): 73–92.

Brueckner, Jan K, Paul S Calem ja Leonard I Nakamura. 2012. “Subprime mortgages and the housing bubble”. *Journal of Urban Economics* 71 (2): 230–243.

Burton, Maureen, Reynold F Nesiba ja Bruce Brown. 2015. *An introduction to financial markets and institutions.* Routledge.

Cao, Li-Juan, ja Francis Eng Hock Tay. 2003. “Support vector machine with adaptive parameters in financial time series forecasting”. *IEEE Transactions on neural networks* 14 (6): 1506–1518.

Chen, Yingjun, ja Yijie Hao. 2018. “Integrating principle component analysis and weighted support vector machine for stock trading signals prediction”. *Neurocomputing* 321:381–402.

Choudhry, Rohit, ja Kumkum Garg. 2008. “A hybrid machine learning system for stock market forecasting”. *World Academy of Science, Engineering and Technology* 39 (3): 315–318.

Cornell, Bradford. 2020. “Medallion Fund: The Ultimate Counterexample?” *The Journal of Portfolio Management* 46 (4): 156–159.

Garber, Peter M. 1989. “Tulipmania”. *Journal of political Economy* 97 (3): 535–560.

Henrique, Bruno Miranda, Vinicius Amorim Sobreiro ja Herbert Kimura. 2018. “Stock price prediction using support vector regression on daily and up to the minute prices”. *The Journal of finance and data science* 4 (3): 183–201.

Huang, Wei, Yoshiteru Nakamori ja Shou-Yang Wang. 2005. “Forecasting stock market movement direction with support vector machine”. *Computers & operations research* 32 (10): 2513–2522.

Ince, Huseyin, ja Theodore B Trafalis. 2017. “A hybrid forecasting model for stock market prediction.” *Economic Computation & Economic Cybernetics Studies & Research* 51 (3).

Jabbar, H, ja Rafiqul Zaman Khan. 2015. “Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study)”. *Computer Science, Communication and Instrumentation Devices,* 163–172.

Kim, Kyoung-jae. 2003. “Financial time series forecasting using support vector machines”. *Neurocomputing* 55 (1-2): 307–319.

Kumar, Deepak, Suraj S Meghwani ja Manoj Thakur. 2016. “Proximal support vector machine based hybrid prediction models for trend forecasting in financial markets”. *Journal of Computational Science* 17:1–13.

Leskinen, Jarre. 2019. “Koneoppiminen rahoitusmarkkinoiden ennustamisessa”.

Liu, Lijuan, Bo Shen ja Xing Wang. 2014. “Research on kernel function of support vector machine”. Teoksessa *Advanced technologies, embedded and multimedia for human-centric computing,* 827–834. Springer.

Lo, Andrew W. 2002. “The statistics of Sharpe ratios”. *Financial analysts journal* 58 (4): 36–52.

Lu, Tsung-Hsun. 2014. “The profitability of candlestick charting in the Taiwan stock market”. *Pacific-Basin Finance Journal* 26:65–78.

Malkiel, Burton G. 2003. “The efficient market hypothesis and its critics”. *Journal of economic perspectives* 17 (1): 59–82.

Malkiel, Burton G, ja Eugene F Fama. 1970. “Efficient capital markets: A review of theory and empirical work”. *The journal of Finance* 25 (2): 383–417.

Manahov, Viktor, Robert Hudson ja Bartosz Gebka. 2014. “Does high frequency trading affect technical analysis and market efficiency? And if so, how?” *Journal of International Financial Markets, Institutions and Money* 28:131–157.

Markowitz, Harry. 1952. “PORTFOLIO SELECTION”. *Journal of Finance* 7 (1): 77–91. https://EconPapers.repec.org/RePEc:bla:jfinan:v:7:y:1952:i:1:p:77-91.

Nayak, Rudra Kalyan, Debahuti Mishra ja Amiya Kumar Rath. 2015. “A Naıve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices”. *Applied Soft Computing* 35:670–680.

Nazário, Rodolfo Torıbio Farias, Jéssica Lima e Silva, Vinicius Amorim Sobreiro ja Herbert Kimura. 2017. “A literature review of technical analysis on stock markets”. *The Quarterly Review of Economics and Finance* 66:115–126.

Sharpe, William F. 1966. “Mutual fund performance”. *The Journal of business* 39 (1): 119–138.

Suykens, Johan AK, Tony Van Gestel, Joos Vandewalle ja Bart De Moor. 2003. “A support vector machine formulation to PCA analysis and its kernel version”. *IEEE Transactions on neural networks* 14 (2): 447–450.

Tay, Francis EH, ja Lijuan Cao. 2001. “Application of support vector machines in financial time series forecasting”. *omega* 29 (4): 309–317.

Vapnik, Vladimir N. 1999. “An overview of statistical learning theory”. *IEEE transactions on neural networks* 10 (5): 988–999.

Yang, Haiqin, Laiwan Chan ja Irwin King. 2002. “Support vector machine regression for volatile stock market prediction”. Teoksessa *International Conference on Intelligent Data Engineering and Automated Learning,* 391–396. Springer.

Żbikowski, Kamil. 2015. “Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy”. *Expert Systems with Applications* 42 (4): 1797–1805.