VAASAN YLIOPISTO

LASKENTATOIMEN JA RAHOITUKSEN YKSIKKÖ

Jussi Kettula

KONKURSSIN ENNAKOINTI K: N LÄHIMMÄN NAAPURIN MENETELMÄLLÄ

Laskentatoimen ja rahoituksen
pro gradu -tutkielma

Laskentatoimen ja tilintarkastuksen
maisteriohjelma

VAASA 2019

SISÄLLYSLUETTELO

sivu

KUVIOLUETTELO

TAULUKKOLUETTELO

VAASAN YLIOPISTO
Laskentatoimen ja rahoituksen yksikkö
Tekijä: Jussi Kettula
Pro gradu -tutkielma: Konkurssin ennakointi k:n lähimmän naapurin menetelmällä
Työn ohjaaja: Teija Laitinen
Tutkinto: Kauppatieteiden maisteri
Oppiaine: Laskentatoimi ja rahoitus
Aloitusvuosi: 2017
Valmistumisvuosi: 2019 **Sivumäärä:** 98

TIIVISTELMÄ

Konkurssin ennakointia käsitteleviä kansainvälisiä tutkimuksia on tehty aina 1960-luvulta saakka ja Suomessakin tutkimusta aihepiiriin liittyen on tehty runsaasti. Tutkimuksellisesti aihe on ollut runsaan kiinnostuksen kohteena, koska konkurssin aiheuttamat taloudelliset ja sosiaaliset tappiot ovat niin yrityksen omistajille, kuin sidosryhmillekin suuria. Aihealuetta käsittelevässä tutkimuksessa on havaittu jo varhain, että konkurssin todennäköisyyttä voidaan ennakoida erilaisten kannattavuuden, maksuvalmiuden ja vakavaraisuuden tunnuslukujen avulla yrityksen tilinpäätöstiedoista. Konkurssin ennakointiin on käytetty perinteisesti esimerkiksi lineaarista erotteluanalyysiä ja logistista regressiota. Tietotekninen kehitys ja tietokoneiden laskentatehossa tapahtunut kasvu on lisännyt erilaisten koneoppimisen algoritmien yleisyyttä konkurssin ennakoinnissa ja eri menetelmiä käsitteleviä tutkimuksia on julkaistu viime aikoina runsaasti.

K:n lähimmän naapurin menetelmä on luokittelualgoritmi, joka on kehitetty jo Coverin ja Hartin (1967) tekemässä tutkimuksessa 60-luvun loppupuolella. Menetelmää on hyödynnetty lukuisissa eri käyttökohteissa laskentatoimeen ja rahoitukseen liittyen. Vaikka menetelmä on hyvin tunnettu, ei sitä ole kansainvälisesti hyödynnetty kuin noin reilussa kymmenessä konkurssin ennakointitutkimuksessa. Suomalaisia tutkimuksia on tiettävästi vain yksi (Kiviluoto 1998). K:n lähimmän naapurin menetelmällä on tiettyjä vahvuuksia muihin koneoppimisen algoritmeihin verrattuna, kuten algoritmin intuitiivisuus, laskennallinen keveys ja yksinkertaisuus. Aikaisemman tutkimuksen suhteellisesti vähäinen määrä ja kattavan suomalaisen tutkimuksen puute kannustaa aihealueen ja menetelmän syvempään kartoittamiseen ja suorituskyvyn tutkimiseen.

Tässä tutkimuksessa koostettiin konkurssiyritysaineisto vuonna 2017 konkurssiin menneistä yrityksistä kolmen konkurssihetkeä edeltävän vuoden tilinpäätöstunnusluvuista. Lopullinen konkurssiyritysten määrä aineistossa oli 86 yritystä. Konkurssiyrityksille etsittiin vastinparimenettelyllä niitä rakenteellisesti vastaavat terveet yritykset (Beaver 1966). Terveiden yritysten avulla koostettiin kaksi otossuhteiltaan erilaista aineistoa, joissa konkurssiyritysten määrä oli vakio, mutta terveitä yrityksiä oli ensimmäisessä aineistossa 86 yritystä ja toisessa 744 yritystä. Aineiston otossuhteet vastasivat suhteessa 50/50 % ja 10/90%. Tutkimuksessa verrattiin k:n lähimmän naapurin menetelmän luokittelutarkkuutta ja siihen vaikuttavia tekijöitä logistiseen regressioon, joka on perinteisesti ollut käytetty menetelmä konkurssin ennakoinnissa.

Tutkimuksen tulosten perusteella havaittiin logistisen regression olevan luokittelutarkkuudeltaan k:n lähimmän naapurin menetelmää tarkempi ja erot olivat sitä selvempiä, mitä lähempänä konkurssihetkeä ennustus tehtiin. Menetelmien erot eivät kuitenkaan olleet tilastollisesti merkitseviä. Aineiston rakenteella ja otossuhteilla havaittiin olevan tilastollisesti merkitsevä vaikutus kummankin menetelmän luokittelutarkkuuteen. K:n lähimmän naapurin menetelmän osalta havaittiin lisäksi, että lähimpien naapureiden lukumäärällä ei ollut selkeää tilastollisesti merkitsevää vaikutusta luokittelutarkkuuteen, eikä myöskään selittävien muuttujien keskinäinen painotus lisännyt merkitsevällä tasolla mallien luokittelutarkkuutta.

AVAINSANAT: konkurssi, konkurssin ennustaminen, koneoppiminen, data-analytiikka, k:n lähimmän naapurin menetelmä.

1. JOHDANTO

1.1. Tutkimuksen tausta

Tilastokeskuksen (2019) mukaan vuonna 2018 konkurssiin haettiin Suomessa 2534 yritystä. Määrä on noussut noin 400 yrityksellä verrattuna vuodentakaisiin tietoihin. Keskimäärin konkursseja on laitettu vireille 2010-luvulla 2754 vuosittain (Tilastokeskus 2019). Maantieteellisesti konkursseja tapahtuu eniten Uudellamaalla, jossa yrityksiä on myös määrällisesti eniten. Konkurssien määrän vuosittainen vaihtelu selittyy paljolti taloudellisiin suhdanteisiin liittyvillä muutoksilla, jotka vaikuttavat kaikkiin yrityksiin toimialasta, koosta ja muista tekijöistä riippumatta.

Konkurssi on prosessina raskas, sillä konkurssiin joutuminen aiheuttaa yritykselle, sen omistajille, velkojille ja muille sidosryhmille usein mittavat tappiot. Konkurssiprosessin ajallinen kesto voi myös olla hyvin pitkä, sillä se voi venyä jopa vuosien mittaiseksi (kts. Koulu 2009:23). Laitisen ja Laitisen (2004:18) mukaan yrityksen konkurssista aiheutuu tappioita monille tahoille: yhteiskunta menettää verotuloja, rahoittajat sijoituksia, hankkijat asiakkaita ja asiakkaat tavarantoimittajia. Yrityksen konkurssin seurauksesta myös yrityksen työntekijät menettävät työpaikkansa ja palkkatulonsa. Yrityksen omistajat sitä vastoin menettävät yritykseen sijoittamansa pääoman, koska omistajille voidaan maksaa pääoma takaisin vasta toissijaisesti muiden velkojien jälkeen, jos pääomaa on tuolloin enää jaettavaksi. Tosiasiassa omistajille usein jää konkurssista jäljelle velkaa, sillä varsinkin pienissä ja keskisuurissa yrityksissä omistajat ovat sijoittaneet yritykseen oman pääoman lisäksi vierasta pääomaa, jonka kiinnitykset ovat henkilökohtaisessa omaisuudessa. (Laitinen ja Laitinen 2004:18.)

Konkurssin vaikutukset näkyvät yrityksen ulkopuolella monella tavalla. Konkurssi voi esimerkiksi laukaista toimintaympäristössään muita konkursseja, koska konkurssiyrityksen ostovelat jäävät myyjäyritysten luottotappioksi. Konkurssin taloudelliset vaikutukset usein laajentuvat myös sosiaalisiksi ja henkilökohtaisiksi vaikutuksiksi yrityksen omistajien ja työntekijöiden elämässä. Sosiaalisten vaikutusten on havaittu korostuvan pienten yritysten omistajien keskuudessa (Kalela, Kiander, Kivikuru, Loikkanen ja Simpura 2001). Yrittäjän mieltävät konkurssin usein henkilökohtaisena tappiona tai epäonnistumisena, vaikka toiminnan loppumisen taustalla voivat olla paljon monimutkaisemmat

syyt. Konkurssiin liittyvän epäonnistumisen leiman on mainittu vaikuttaneen yhteisön ajattelutapaan, asenteeseen ja luottamukseen konkurssin tehneestä yrittäjästä. (Stokes ja Blackburn 2002:3.) Sosiaalinen selviytyminen konkurssista on usein riippuvainen yrittäjän kyvystä suhtautua konkurssiin opettavaisena kokemuksena henkilökohtaisen epäonnistumisen sijaan. Suomessa on varsinkin 90-luvun laman ja siitä seuranneiden konkurssien jälkeen on tutkittu konkurssin vaikutuksia yksityiselämään (kts. Kalela, Kiander, Kivikuru, Loikkanen ja Simpura 2001). Useissa tapauksissa konkurssin havaittiin johtavan mielen epätasapainoon, syrjäytymiseen, työttömyyteen ja perheen menettämiseen (Lampela-Kivistö, Sorri ja Kiiski 2001:462-478).

Konkurssiin johtaneita syitä ja konkurssin ennakoimista on tutkittu useissa eri tutkimuksissa (mm. Altman 1968; Meyer ja Pifer 1970; Laitinen 1991). Mitä aikaisemmin konkurssi pystytään ennakoimaan, sitä todennäköisemmin se olisi vältettävissä ja sen vaikutukset minimoitavissa. Riittävän aikainen hälytysjärjestelmä voisi mahdollistaa liiketoiminnan ja yrityksen rakenteellisen mukauttamisen, jotta konkurssia kohti vievä kehitys voitaisiin kääntää. Riittävän aikainen varoitus konkurssista mahdollistaisi liiketoiminnan hallitun alasajon tilanteessa, jossa yrityksen elinkelpoisuus todetaan peruuttamattomasti menetetyksi. (Laitinen ja Laitinen 2004:19.)

Konkurssiin johtavien syiden ymmärtäminen ja maksukyvyttömyyden ennakoiminen on tärkeää sekä yrityksen sisäisestä, että ulkoisesta näkökulmasta. Yrityksen sisällä tietoa tarvitaan johdon päätöksenteon tueksi ja yrityksen ulkopuolella taas sidosryhmien, kuten sijoittajien ja lainanantajien, päätöksentekoa tukevana elementtinä. Yrityksen johdolla on aina lähtökohtaisesti käytettävissään enemmän ja ajankohtaisempaa tietoa liiketoiminnan tilasta kuin muilla yrityksen sidosryhmillä. Tästä syystä johdolla on parhaat lähtökohdat elinkelpoisuuden arviointiin. Hälytysjärjestelmä voisi auttaa johtoa arvioimaan yrityksen elinkelpoisuutta ja suunnittelemaan korjaavia toimenpiteitä ennen kuin konkurssiin viittaava kehityskulku on sidosryhmien nähtävillä tilinpäätösinformaatiossa (Laitinen ja Laitinen 2004:20.)

Konkurssin ennakointimenetelmiä ja syitä konkurssin taustalla on pyritty tutkimuksissa kartoittamaan aina 60-luvulta lähtien. Ensimmäisiä konkurssin ennakointia tilinpäätöstunnuslukujen avulla käsitteleviä tutkimuksia olivat muun muassa Beaverin (1966) ja Altmanin (1968) tutkimukset. Beaver (1966) kartoitti tilastollisesti yksittäisissä

tilinpäätöstunnusluvuissa esiintyviä muutoksia konkurssin lähestyessä. Altman (1968) taas tiivisti usean tunnusluvun informaation yhteen yhdistelmälukuun ja havaitsi tunnuslukujen ennakoivan yhdessä konkurssin todennäköisyyttä paremmin, kuin yksin. Ensimmäisenä merkittävänä suomalaisena tutkimuksena on Prihtin (1975) tutkimus, jossa hän aikaisemmista empiriaan pohjautuvista tutkimuksista poiketen rakensi teoreettista viitekehystä konkurssin taustalla olevista syistä. (Laitinen ja Laitinen 2004:94-96).

Perinteiset konkurssin ennakointimenetelmät pohjautuvat tyypillisesti tilastotieteellisiin menetelmiin, kuten erotteluanalyysit ja logistinen regressio. Nykyaikaiset menetelmät perustuvat koneoppimisen algoritmeihin, joissa tilastotieteellisiä oletuksia konkurssin taustatekijöistä ei vaadita. Data-analytiikan ja koneoppimisen saralla on tehty useita tutkimuksia, joissa erilaisia luokittelualgoritmeja on tutkittu konkurssin ennakoinnissa. Esimerkkejä käytetyistä menetelmistä ovat neuroverkot, CBR (Case-Based Reasoning), geneettiset algoritmit, päätöspuut, SVM (Support Vector Machine), k:n lähimmän naapurin menetelmä, bayesian algoritmit ja sumea logiikka (Fuzzy logic) (Amani ja Fadlalla 2017:34). Uusia menetelmiä on valtavasti ja niillä saadut luokittelutarkkuudet ovat hyviä, mutta yksiselitteisesti yhtä ylivertaista menetelmää perinteisten ennakointimenetelmien korvaajaksi ei ole löydetty. Jotkin uudet luokittelumenetelmät ovat konkurssin ennakoinnissa hyvin tehokkaita, mutta intuitiivisesti hankalasti ymmärrettäviä, kuten esimerkiksi neuroverkot (Chen, Yang, Wang, Liu, Xu, Wang ja Liu 2011:1). Neuroverkkojen käyttöä konkurssin ennakoinnissa ovat tutkineet muun muassa Odom ja Sharda 1990; Laitinen ja Kankaanpää 1999 ja Atiya 2001. Myös eri algoritmien parhaita puolia yhdistävän Random Forest -algoritmin on havaittu olevan luokittelutarkkuudeltaan lupaava menetelmä konkurssin ennakoimisessa (kts. Breiman 2001; Barboza, Kimura ja Altman 2017).

K:n lähimmän naapurin algoritmi on suhteellisen vanha luokittelualgoritmi, jonka taustaajatus hahmoteltiin ensimmäisen kerran Coverin ja Hartin (1967) tutkimuksessa. Algoritmin toimintaperiaate on yksinkertainen ja helpommin intuitiivisesti ymmärrettävissä kuin monen muun algoritmin (Chen ym. 2011:1). Luokittelualgoritmeja vertailevassa tutkimuksessa k:n lähimmän naapurin algoritmi on valittu 10 parhaimman joukkoon (Wu, Kumar, Quinlan, Ghosh, Yang, Motoda, McLachlan, Ng, Liu, Yu, Zhou, Steinback, Hand ja Steinberg 2008). Algoritmin perustuu tuntemattoman havainnon luokittelemiseen suoraan lähimpänä piirreavaruudessa olevien havaintojen painotettuna keskiarvona (Haara 2002:6). Menetelmällä on saatu aikaisemmassa tutkimuksessa lupaavia tuloksia

konkurssin ennakoinnissa (Park ja Han 2002; Bian ja Mazlack 2003; Yip 2004; Ribeiro, Vieira, Duarte, Silva, Das Neves, Liu ja Sung 2008; Chen ym. 2011, Serrano-Cinca ja Gutiérrez-Nieto 2013). Suomessa on tiettävästi menetelmää käytetty yhdessä tutkimuksessa (kts. Kiviluoto 1998).

1.2. Tutkimusongelma

Tämän tutkimuksen tavoitteena on tutkia konkurssin ennakoimisen luokittelutarkkuutta k:n lähimmän naapurin algoritmia käyttäen. Aineistona tutkimuksessa käytetään suomalaisten yritysten tilinpäätösdataa Asiakastiedon tietokannasta. menetelmän luokittelutarkkuutta ja suorituskykyä verrataan tutkimuksessa logistiseen regressioon, joka on menetelmänä perinteinen, intuitiivinen ja suorituskykyinen.

Tässä tutkimuksessa pyritään myös kartoittamaan algoritmin suorituskykyyn vaikuttavia tekijöitä ja mittaamaan ovatko tekijöiden vaikutus luokittelutuloksiin merkittävä. Tutkielman empiriaosiossa tutkitaan, miten lähimpien naapurien määrä, etäisyysfunktion muoto, muuttujien painotus, erilaiset datan esikäsittelytavat ja aineiston rakenne vaikuttavat menetelmän suorituskykyyn. Tämä tutkimus ei suoraan perustu mihinkään yksittäiseen aikaisempaan tutkimukseen, vaan tukittavat asiat ja ennakko-odotukset ovat koostettu useasta aikaisemmasta tutkimuksesta. Menetelmävaihtoehtojen ja tekijöiden vaikutusta mallin hyvyyteen verrattiin yleisesti käytössä olevilla tunnuslukujen avulla, kuten luokittelutarkkuudella, AUC-arvolla ja virheluokitusten määrillä. Luokittelutarkkuuksissa havaitun eron tilastollisen merkitsevyyden määrittämiseen käytettiin z-testisuuretta.

1.3.Tutkimuksen rakenne

Tutkielma koostuu luvuissa 2-4 käsiteltävästä teoriaosuudesta ja luvussa 5-6 käsiteltävästä empiriaosuudesta. Teoriaosiossa kuvataan luvun 2 alussa lyhyesti maksukyvyttömyyden seurauksia, joita ovat yrityssaneeraus ja konkurssi. Yrityssaneerauksen ja konkurssin määritelmien jälkeen kuvataan erilaisia mikro- ja makrotaloudellisia tekijöitä maksukyvyttömyyden taustalla, jotka edesauttavat maksukyvyttömyyden syntymistä.

Luvun loppupuolella tarkastellaan laskentatoimen näkökulmasta yksittäisen yrityksen mittakaavassa maksukyvyttömyyden taustalla vaikuttavia tekijöitä.

Luvussa 3 käsitellään hyvin lyhyesti konkurssin ennakoimisessa perinteisesti käytettyjä menetelmiä, kuten yhden- ja usean tunnusluvun malleja. Koska tämän tutkimuksen tarkoituksena on käsitellä k:n lähimmän naapurin luokittelumenetelmää, joka on koneoppimisen algoritmi, on luvussa koostettu tiivis yhteenveto muista yleisistä koneoppimisen algoritmeista. Luvun lopussa on kuvattu konkurssin ennakointitutkimuksessa käytössä olleita aineistojen rakenteellisia piirteitä ja selittävien muuttujien valinnassa käytettyjä menetelmiä. Luvussa 4 käsitellään tässä tutkimuksessa käytettävää knn-menetelmää hieman yksityiskohtaisemmin, sekä käydään läpi aikaisempia konkurssinennakointitutkimuksia, joissa menetelmää on käytetty. Luvun lopussa on lyhyt yhteenveto ja kooste tässä tutkimuksessa tutkittavista hypoteeseista.

Luvussa 5 esitellään tutkimuksessa käytetty aineisto ja sen keräämisessä käytetyt tietokannat ja oletukset. Aineiston rakennetta ja aineiston suodattamista lopulliseen muotoon pyritään kuvaamaan mahdollisimman kattavasti. Luvussa esitellään periaatteet, joita käyttämällä malleihin valittiin selittävät muuttujat ja miten selittävien muuttujien hyvyyttä arvioitiin suhteessa vastemuuttujaan ja muihin selittäviin muuttujiin. Tulosten laskentaa edeltävät prosessit ja menetelmät ovat esitelty erillisissä alaluvuissa k:n lähimmän naapurin menetelmälle ja logistiselle regressiolle. Luvussa 5 kuvataan myös erilaiset suorituskykyä mittaavat tunnusluvut, joilla mallien suorituskykyä, hyvyyttä ja tilastollista merkitsevyyttä mitattiin.

Luvussa 6 esitellään k:n lähimmän naapurin menetelmän osalta optimaalisen parametrivalinnan tulokset ja lopullisilla malleilla lasketut luokittelutulokset. Samassa yhteydessä esitellään myös logistisella regressiolla lasketut tulokset ja vertaillaan niitä k:n lähimmän naapurin menetelmän tuloksiin. Tulosten tarkastelun osalta palataan aikaisempaan tutkimukseen ja tarkastellaan tutkimuksessa saatuja tuloksia asetettujen hypoteesien näkökulmasta. Lopussa pohditaan tulosten yleistettävyyttä ja mahdollisia yleistettävyyteen vaikuttavia asioita. Luvussa 7 seuraa lyhyt yhteenveto koko tutkimuksesta.

2. MAKSUKYVYTTÖMYYS JA SEN SYNTYYN VAIKUTTAVAT TEKIJÄT

2.1. Maksukyvyttömyyden määritelmä

Maksukyvyttömyys määritellään yleisesti yrityksen kyvyttömyydeksi selviytyä maksuvelvoitteista niiden tullessa maksuun (Laitinen ja Laitinen 2014:108). Maksukyvyttömyys voi olla luonteeltaan lievää, jolloin yritys selviää siitä korkeintaan maksuviiveellä (Laitinen ja Laitinen 2004:16). Tällainen yritys tyypillisesti maksaa maksut, mutta myöhässä. Maksuviiveitä voi tapahtua myös silloin, kun yritys on maksukykyinen, mutta jostain syystä maksuhaluton (Laitinen ja Laitinen 2014:10). Yrityksen tahtotilasta riippumattoman maksukyvyttömyyden myötä syntyvien maksuviiveiden on havaittu muuttuvan vähitellen vakavimmiksi. Mikäli yritys ei kykene saattamaan rahoitustaan kuntoon, maksukyvyttömyys pahenee ja voi johtaa maksuhäiriöön. Maksuhäiriöitä voi kertyä yritykselle useita peräkkäin, kun maksusitoumukset jätetään maksamatta. Yritys voi pyrkiä sopimaan maksuhäiriöiden kasaantuessa vaihtoehtoisista maksujärjestelyistä velkojien kanssa. Jos yhteisymmärrystä velkojien kanssa ei saavuteta tai lisärahoitusta saada, on seurauksena yleensä joko yrityssaneeraus tai konkurssi. (Laitinen ja Laitinen 2004:16.)

2.2. Maksukyvyttömyyden seuraukset yritykselle

2.2.1. Yrityssaneeraus

Saneerauksesta säädetään yrityksen saneerauslaissa (YSL 47/1993). Saneerauksen tarkoitus on pyrkiä tervehdyttämään ja mahdollistamaan maksuvaikeuksissa olevan velallisen yritystoiminnan jatkuvuus virallisen saneerausmenettelyn kautta (YSL 1:1§). Edellytyksenä saneerausohjelman aloittamiselle on, että velallista uhkaa maksukyvyttömyys tai velallisella on vähintään kaksi velkojaa, joiden yhteenlasketut saatavat edustavat vähintään 20 % velallisen tunnetuista veloista (YSL 2:6§). Lisäksi saneerauksen edellytyksenä on, että velallinen yritys nähdään liiketoiminnaltaan elinkelpoisena, jonka pitkäaikaisen maksukyvyn parantamiseen saneerausmenettelyllä on positiivinen vaikutus.

Osallistumista saneerausmenettelyyn haetaan tuomioistuimelta ja hakijana voi olla velallinen itse tai velkoja (YSL 2:5§). Tuomioistuimen päätöksellä alkaneen saneerausmenettelyn myötä velallinen asetetaan perintäkieltoon, jolloin velalliseen ei saa kohdistaa toimenpiteitä maksukiellon piiriin kuuluvan saneerausvelan perimiseksi ja ennen saneerausta aloitetut perintätoimenpiteet joudutaan keskeyttämään (YSL 4:19§). Saneerauksen alkamisen jälkeen velalliseen ei myöskään voida kohdistaa maksuviivästyksen seurauksia (YSL 4:19§). Edellytykset ja esteet saneerausmenettelyn suhteen johtavat siihen, että yrityksen tila saneerausmenettelyyn haettaessa ei voi olla toivoton (Laitinen ja Laitinen 2004:64).

Hakemus saneerausmenettelyyn voidaan tehdä silloinkin, vaikka velallinen olisi jo haettu konkurssiin (YSL 24§). Lain mukaan tällöin konkurssihakemusta ei saa ratkaista ennen kuin päätös saneerausmenettelyn aloittamisesta on tehty (YSL 24§). Saneeraushakemukset jätetään tästä johtuen usein vasta viime hetkellä konkurssihakemuksen ollessa jo vireillä, koska keskeneräinen saneerausmenettely antaa suojan konkurssia vastaan (Laitinen ja Laitinen 2004:63).

Laakson, Laitisen ja Vennon (2010:149) mukaan saneerausmenettelyn kesto vaihtelee 6-8 vuoden välillä. Saneerausohjelmien onnistumista on kritisoitu niin Suomessa, kuin muuallakin maailmalla (Laakso ym. 2010:148). Saneerausohjelmien onnistumistodennäköisyytenä on yleisesti pidetty noin 50 % osuutta saneeraukseen osallistuneista yrityksistä (Laakso ym. 2010:148). Tutkimusten mukaan vain noin 25 % tervehtymistä tavoittelevista yrityksistä onnistuu siinä saneerauksen kautta ja alimmillaan onnistumisen on havaittu olevan noin 7 % tasolla (Laakso ym. 2010:148). Saneerausohjelman epäonnistuminen tai saneeraushakemuksen hylkääminen johtaa tyypillisesti poikkeuksetta yrityksen konkurssiin (Laitinen ja Laitinen 2004:17).

2.2.2. Konkurssi

Konkurssiin liittyvistä asioista säädetään konkurssilaissa (KonkL 120/2004). Määritelmänsä mukaisesti konkurssissa on kyse maksukyvyttömyysmenettelystä, joka koskee velallisen kaikkia velkoja ja jossa velallisen omaisuus käytetään yhdellä kertaa velkojen maksuun (KonkL 1:1§). Päämääränä konkurssissa on yrityksen varojen

oikeudenmukainen ja tasapuolinen jakaminen velkojien kesken (Laitinen ja Laitinen 2004:17). Konkurssi syntyy maksukyvyttömyyden seurauksena, kun yritys ei enää kykene vastaamaan veloistaan. Konkurssissa yrityksen omaisuus siirtyy velkojien määräysvaltaan. Tuomioistuin määrää konkurssipesälle pesänhoitajan, joka vastaa omaisuuden hoitamisesta, myymisestä sekä konkurssipesän hallinnosta. Konkurssiin hakeudutaan joko yrityksen omasta aloitteesta tai velkojan hakemuksesta. Tilastokeskuksen (2012) mukaan vuonna 2012 vireille asetetuista konkursseista 22 prosentissa hakijana oli velallinen itse ja 78 prosentissa hakijana oli velkoja. Velkojan osalta verottaja oli konkurssiin hakijana 53 prosentissa tapauksista, vakuutusyhtiöt 31 prosentissa ja muut hakijat 16 prosentissa konkursseista.

Verottajan suuri osuus konkurssiin hakijoista johtuu niin kutsutuista vaarallisista veloista (Laitinen ja Laitinen 2014:23). Näitä ovat kaikki yritystoiminnasta juoksevasti syntyvät ja verottajalle tilitettävät verot, jotka ovat ulosottokelpoisia heti eräpäivän täytyttyä ilman tuomioistuimen päätöstä ja joiden kertymiseen velkoja ei voi vaikuttaa. Tämän vuoksi verottaja hakee herkemmin velallista konkurssiin, koska sitä kautta velkaantuminen saadaan katkaistua. Vaarallisiksi veloiksi luokitellaan niin ikään yrityskiinnityksin solmitut pankkivelat, koska konkurssin tapahtuessa velkoja todennäköisesti saa ensimmäisten joukossa suorituksen saatavalleen kiinnitykseen perustuen. Muita velkoja pidetään yleisesti vaarattomampina, koska ne ovat toisiinsa nähden yhtäläisessä asemassa konkurssissa ja siksi niiden maksuun paneminen konkurssia käyttämällä ei lisää velallisen todennäköisyyttä maksaa velka kokonaisuudessaan takaisin. (Laitinen ja Laitinen 2014:23.)

Konkurssin ajallisen keston on mainittu olevan keskimäärin noin 3 vuotta ja 7 kuukautta (Koulu 2009:23). Konkurssimenettely lähtee liikkeelle konkurssihakemuksesta ja tuomioistuimen päätöksestä asettaa yritys konkurssiin. Konkurssiin asettamisen yhteydessä tuomioistuin välittää tiedon Patentti- ja rekisterihallitukselle kaupparekisteriin tehtävää merkintää varten (Konkurssilaki 22:65§). Kun konkurssipesän omaisuus on muutettu rahaksi, tehdään lopputilitys, jossa jako-osuudet maksetaan velkojille ja konkurssipesä puretaan (KonkL 19:1§).

Koulun (2009:23-24) mukaan tyypillinen konkurssi on lähes tuloksetonta velkojien näkökulmasta, sillä konkurssipesästä jää jako-osuutta keskimäärin vain noin 10 % velkojille suhteessa alkuperäisten velkojen määrään. Vuonna 2000 85 %:ssa konkursseista jako-

osuus tavallisille velkojille jäi alle 10 prosenttiin velkojen määrästä ja 50 %:ssa konkurssitapauksissa jako-osuutta ei jäänyt ollenkaan. Konkurssi on ulosottoa huomattavasti tehottomampi, sillä ulosotossa päästään keskimäärin 20 % perimistulokseen. Saneerausmenettelyssä perimistuloksen on mainittu olevan keskimäärin noin 40 % velkojen kokonaismäärästä. (Koulu 2009:23-24.)

Sundgrenin (1995:83) mukaan 25-45 % konkurssiin haetuista yrityksistä etenee koko konkurssiprosessin läpi ja saa lopuksi konkurssituomion. Konkurssiprosessissa kustannuksia syntyy konkurssipesän hallinnosta, kuten kirjanpidosta, asianajajista, pesänhoitajasta ja muista hallinnollisista lähteistä. Kustannusten on havaittu olevan keskimäärin noin 13,6 % konkurssiyritysten varallisuudesta. Jos konkurssiyritykset pilkottiin ja myytiin paloittain ja kokonaisuuksittain, olivat kustannukset siitä huolimatta 15,7 % yrityksen varallisuudesta. (Sundgren 1995:83-105.)

2.3. Maksukykyä vaarantavat tekijät toimintaympäristössä

2.3.1. Mikrotaloudelliset tekijät

Laitisen ja Laitisen (2014:24) mukaan konkurssiuhkan suuruus riippuu yrityksen ominaisuuksista. Mikrotaloudellisia riskiryhmiä ja riskitekijöitä voidaan tunnistaa esimerkiksi seuraavien tekijöiden suhteen: toimiala, yhtiömuoto, koko, ikä ja sijainti. **Toimialan** suhteen riskin suuruus riippuu toimialalla vallitsevasta kilpailusta, kysynnästä ja toimialan herkkyydestä suhdanteisiin. Jotkin toimialat ovat suhdanneherkempiä kuin toiset, mikä näkyy tilauskannan, kysynnän, vaihto-omaisuuden ja kapasiteetin käyttöasteen erilaisina muutoksina. Esimerkiksi rakentamisen toimialalla konkurssiriski on ollut huomattavasti korkeampi kuin muilla toimialoilla (Laitinen ja Laitinen 2014:27). Suhteellisesti toimialaan linkittyvä riski on 0-26,4 % välillä ja alhaisin riski on sähkö-, kaasu- ja lämpöhuollon toimialalla ja korkein rakentamisessa. Rakentamista on yleisesti pidetty suhdanteita hyvin ilmentävänä toimialana, jossa suuret heilunnat ovat suhdanteiden muuttuessa yleisiä. (Laitinen ja Laitinen 2014:24-27.)

Yhtiömuodolla on myös vaikutusta konkurssiriskin suuruuteen (Laitinen ja Laitinen 2014:29). Yhtiömuoto valikoituu yrityksen perustamishetkellä, kun optimoidaan vaihtoehtoja suhteessa toiminnan luonteeseen, verotukseen ja toimintaan sisältyvään riskiin. Jos

yhtiö ei tavoite kasvua ja toiminta on luonteeltaan vähäriskistä, toiminimi tai henkilöyhtiö on tavallisesti paras valinta yhtiömuodoksi. Henkilöyhtiöissä yksi tai useampi omistajista vastaa täysimääräisesti yhtiön veloista. Toiminnan ollessa riskipitoisempaa ja yhtiön tavoitellessa kasvua, valitaan yhtiömuodoksi tavallisesti pääomayhtiö, joka on juridinen oikeussubjekti ja vastaa ottamistaan veloista. Pienten osakeyhtiöiden tapauksessa tyypillistä on, että omistaja tosiasiassa takaa yrityksen ottamia velkoja, joten koko velkataakka ei välttämättä ole yhtiön vastuulla. Laitisen ja Laitisen (2014:29) mukaan vuosina 2007-2011 suurin konkurssiriski oli osakeyhtiöillä (1,7 %) ja pienin konkurssiriski toiminimiyrittäjillä (0,3 %). Henkilöyhtiöiden riski sijoittuu toiminimiyrittäjien ja osakeyhtiöiden välille noin 1 % konkurssiriskillä. (Laitinen ja Laitinen 2014:29.)

Yrityksen koon vaikutus konkurssiriskiin näkyy siten, että riski on alhainen (0,8-1%) pienissä alle 5 henkilöä työllistävissä mikroyrityksissä ja riski kasvaa suurimmillaan 1-2,5 % tasolle yrityksen koon kasvaessa aina 5-50 henkilömäärään saakka (Laitinen ja Laitinen 2014:30). Suuremmissa yli 50 henkilöä työllistävissä yrityksissä konkurssiriski alenee alle 1 % tasolle. Syynä tähän on esimerkiksi mikroyritysten tapauksessa niiden työvoimavaltaisuus, palvelutoimialapainotus ja toiminimi- ja henkilöyhtiömuotojen yleisyys. Näiden ominaisuuksien vuoksi mikroyrityksiin ei kerry varallisuutta kovin runsaasti, minkä seurauksesta konkurssin todennäköisyys alenee, kun velkojille ei riitä jaettavaa jako-osuutta. Pienyrityksissä (5-50 henkilöä) konkurssit ovat todennäköisimpiä, sillä yrityksiin ei ole sitoutunut riittävästi varallisuutta, jotta sitä voitaisiin tiukkoina aikoina realisoida saneeraustoimien rahoittamiseen. Suuremmat yritykset (yli 50 henkilöä) ovat toimintansa suhteen joustamattomampia, mutta niihin on sitoutunut niin paljon varallisuutta, että sitä voidaan realisoida ja kattaa näin saneerauksessa syntyviä kustannuksia. Suurempiin yrityksiin on myös sitoutunut niin paljon varallisuutta, että niitä harvemmin päästetään konkurssiin. (Laitinen ja Laitinen 2014:30.)

Yrityksen ikä on yksi merkittävimmistä konkurssiriskeistä. Konkurssiriski on suurimmillaan (n. 1,3 %) yrityksen ollessa alle 4 vuoden ikäinen. Vanhemmilla yrityksillä riski lähtee alenemaan toiminnan vakiintuessa, kunnes yli 15-vuotiaiden yritysten tapauksessa se vakiintuu noin 0,5 % tasolle. Suurempi konkurssiriski elinkaaren alussa selittyy aloittavan yrityksen pienellä koolla ja suhteellisen suurella velkaisuudella. (Laitinen ja Laitinen 2014:32.)

Yrityksen **maantieteellinen sijainti** vaikuttaa osaltaan konkurssien lukumäärään, koska yritykset ovat jaottuneet epätasaisesti Suomen sisällä painottuen pääkaupunkiseudulle. Yritysten toimintaedellytykset voivat maantieteellisen sijainnin vuoksi olla epätasaiset johtuen paikallisesta kysynnästä, kilpailutilanteesta tai paikallisen tukiverkoston ja verkostoitumisen laajuudesta. Maantieteellinen sijainti vaikuttaa jonkin verran suhteellisen konkurssiriskin suuruuteen. Riskin on havaittu olevan suurimmillaan Uudellamaalla 1,10 % tasolla ja alimmillaan Ahvenanmaalla 0,38 % tasolla. (Laitinen ja Laitinen 2014:34.)

2.3.2. Makrotaloudelliset tekijät

Makrotaloudelliset tekijät ovat yrityksestä riippumattomia tekijöitä, jotka määrittelevät yrityksen toimintaympäristössä olevat olosuhteet, joissa yritys harjoittaa toimintaa. Näihin kuuluvat esimerkiksi suhdanteet, vienti ja tuonti, inflaatio ja rahoitusmarkkinat (Laitinen ja Laitinen 2014:34). **Suhdanne** on makrotaloudellinen tekijä, joka vaikuttaa kansantalouden kasvuvauhtiin noususuhdanteessa nopeuttavasti ja vastaavasti laskusuhdanteessa kasvua alentavasti. Yritykselle suhdannevaihtelu vaikuttaa samalla tavalla, kuin koko kansantaloudellekin. Noususuhdanteessa kysyntä on keskimäärin suurempaa, mikä näkyy suurempana myyntinä sekä parempana kannattavuutena. Vastaavasti laskusuhdanteessa myynnin määrä vähenee, jolloin myös kannattavuus heikkenee. Noususuhdanteessa yrityksen on mahdollista käyttää ansiokkaasti velan vipuvaikutusta hyväkseen, kun vieraan pääoman kustannus on alhaisempi kuin oman pääoman kustannus (Laitinen ja Laitisen 2014:36). Laskusuhdanteessa velkapainotteinen pääomarakenne saattaa kuitenkin kääntyä yritystä vastaan, koska vieraan pääoman kustannukset nousevat ja kysyntä alenee. Jos yrityksen rahoitus- ja kustannusrakenne on laskusuhdanteessa epäedullinen, yrityksen kannattavuus usein heikkenee voi heiketä ja tulorahoitus laskea. Tämä saa yrityksen turvautumaan vahvemmin vieraaseen pääomaan, joka heikentää yrityksen maksuvalmiutta ja maksukykyä entisestään. Lopulta vieraan pääoman kustannukset nousevat liian suuriksi ja yritys joutuu konkurssiin (Laitinen ja Laitinen 2014:37).

Ulkomaan viennin ja tuonnin määrä vaikuttaa suoraan konkurssiriskin kehittymiseen. Esimerkiksi viennin kasvaessa tämä merkitsee yritykselle kasvavia markkinoita ja kasvavaa kysyntää. Vaikutus on tulorahoitusta ja kasvua lisäävä ja siten konkurssiriskiä alentava. Viennin vaikutus voi myös osaltaan olla negatiivinen, sillä vientiin liittyy riskejä ja

viennin lisääntyminen voi kasvattaa kilpailua. Tuonnilla on sekä positiivisia, että negatiivisia vaikutuksia. Tuonnin myötä yritys voi saada edullisempia tuotannontekijöitä, mikä parantaa yrityksen kustannusrakennetta. Tuonti voi myös vaikuttaa negatiivisesti, kun ulkomailta virtaa yrityksen tuotteita vastaavia substituutteja, jotka tiukentavat kotimaisten markkinoiden kilpailua. (Laitinen ja Laitinen 2014:37)

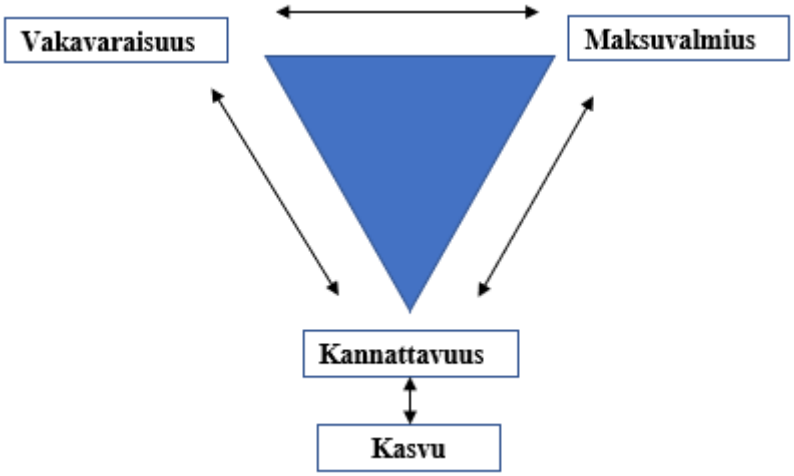
Kolmas makrotaloudellinen tekijä on **inflaatio**. Sen vaikutus yrityksen konkurssiriskiin on osin positiivinen ja negatiivinen. Positiivisesti inflaatio vaikuttaa yrityksen näkökulmasta esimerkiksi vieraan pääoman kautta, kun lainan reaaliarvo laskee inflaation seurauksesta. Tästä seuraa alemmat lainanlyhennyskulut ja hetkellisesti myös lainan korot ovat alemmat. Toisaalta inflaatio voi vaikuttaa yritykseen negatiivisesti, kun tuotantotekijöiden nimellinen arvo nousee. Jos nousu on nopeaa, niin yritys ei välttämättä ehdi siirtää tuotantotekijöiden hinnannousua riittävän nopeasti myytävien tuotteiden hintoihin. (Laitinen ja Laitinen 2014:37)

Rahoitusmarkkinat vaikuttavat yrityksen konkurssiriskiin luotonsaannin ehtojen kireyden ja luoton kustannusten kautta (Laitinen ja Laitinen 2014:38). Kun luotonsaannin ehdot kiristyvät, yritys ei saa tarvittaessa yhtä helposti lisärahoitusta ja lisärahoituksen kustannukset lisäksi nousevat. Mitä velkaantuneempi yritys on, sitä tiukempia ehtoja ja korkeampia kustannuksia sen kohdalla sovelletaan – mikä ajaa yrityksen lopulta konkurssiin. Rahoitusmarkkinoiden kiristyminen voi kuitenkin vaikuttaa konkurssiriskiä alentavasti, kun yritykset eivät saa yhtä helposti lainaa riskipitoisiin investointeihin, joiden on nimenomaan havaittu olevan eräs konkurssin keskeisimmistä syistä. (Laitinen ja Laitinen 2014:38)

2.4. Laskentatoimi ja maksukyvyttömyys

Yrityksen taloudellisia toimintaedellytyksiä kuvataan laskentatoimen kirjallisuudessa usein terveyskolmion avulla (kts. Laitinen ja Laitinen 2004:242-245). Terveyskolmion kulmat koostuvat kannattavuudesta, maksuvalmiudesta ja vakavaraisuudesta (kuvio 1). Kaikki nämä tekijät yhdessä ovat yrityksen liiketoiminnan jatkuvuuden kannalta tärkeitä. Yhdenkin tekijän poistaminen saa kolmion kaatumaan ja yrityksen ennen pitkää maksukyvyttömyyden partaalle. Neljäntenä komponenttina terveyskolmiossa esitetään lisäksi

kasvu, joka toimii ikään kuin katalyyttina muille tekijöille. Kannattavuuden ja kasvun yhdistelmä määrittää tulorahoituksen riittävyyden ja siten yrityksen vakavaraisuuden ja maksuvalmiuden.



Kuvio 1. Yrityksen terveyskolmio. (kts. Laitinen ja Laitinen 2014: 131)

Kun kannattavuus on heikko ja kasvu on voimakasta, päädytään tilanteeseen, jossa maksuvalmius vaarantuu, koska heikko kannattavuus ei takaa riittävää tulorahoitusta. Jos kannattavuus on heikko ja kasvu maltillista, voidaan päätyä tilanteeseen, jossa vakavaraisuus vaarantuu. Tämä johtuu heikon kannattavuuden kuluttamista omista pääomista, joka johtaa lopulta vakavaraisuuden alentumiseen. Rahoituskriisin ennakoinnin kannalta keskeisimmät tunnusluvut ovat juuri yrityksen kasvua ja kannattavuutta kuvaavat tunnusluvut (Laitinen ja Laitinen 2004:244).

2.4.1. Kannattavuus ja kasvu

Kannattavuudella tarkoitetaan yrityksen pitkän aikavälin tulontuottamiskykyä, kun huomioidaan menojen ja niiden kautta saatujen tulojen aikaviive (Laitinen ja Laitinen 2004:245). Yritys kannattaa sitä paremmin, mitä nopeammin ja enemmän voittoa se pystyy tuottamaan suhteessa kustannuksiin (Laitinen ja Laitinen 2004:245). Kannattavuutta voidaan mitata joko absoluuttisesti rahamääräisenä tai suhteellisesti vertaamalla voitontuottokykyä yrityksen kokoon (Laakso ym. 2010:19-26). Suhteuttamiseen voidaan käyttää liikevaihtoa, jolloin kannattavuuden vertailua voidaan tehdä paremmin yrityksen

koosta riippumatta. Toimialojen vertailu tuloslaskelmasta lasketuilla kannattavuusluvuilla on kuitenkin haastavaa, sillä eri toimialoilla on erilainen kustannusrakenne ja sen myötä lähtökohtaisesti erilainen katteiden taso (Laakso ym.2010:24). Joillakin toimialoilla pääosa kustannuksista syntyy raaka-aineista, toisilla henkilöstökuluista ja osalla taas liiketoiminnan muista kuluista. Toimialojen vertailussa toimii paremmin pääoman tuottosuhteet. Tuottosuhde voidaan laskea esimerkiksi koko pääoman suhteen, korollisen pääoman suhteen tai oman pääoman suhteen. (Laakso ym. 2010:24.)

Kannattavuudessa erotetaan lähtökohtaisesti erilleen liiketoimintaan perustuva kannattavuus ja rahoitukseen perustuva kannattavuus. Liiketoiminnan kannattavuus syntyy suoraan yrityksen varsinaisen liiketoiminnan seurauksesta ja rahoituksen kannattavuus taas toissijaisesti esimerkiksi sijoituksista tai koroista saaduista tuotoista, jotka eivät ole suoraan riippuvuudessa yrityksen varsinaiseen liiketoimintaan. Kannattavuuteen liittyy tiiviisti myös muuttuvien ja kiinteiden kulujen käsite. Tyypillisesti muuttuvat kulut kehittyvät suhteessa liikevaihtoon, eli kulut tyypillisesti kasvavat, kun liikevaihto kasvaa. Rahoitustuotot ja -kulut ovat sitä vastoin luonteeltaan usein kiinteitä. Merkittävä osa liiketoiminnan riskistä syntyy kulurakenteesta, koska yritys ei kykene mukauttamaan kustannusrakennettaan tulorahoituksessa tapahtuvien muutosten mukaisesti (Laitinen ja Laitinen 2014:114-115). Tästä esimerkkinä voidaan mainita lainan kustannukset, jotka eivät riipu siitä, miten paljon tuotteita yritys saa myydyksi. Mitä suurempi osuus yrityksen kulurakenteesta muodostuu kiinteistä kustannuksista ja mitä pienempi on yrityksen katetuotto, sitä suuremmaksi kohoaa tappiollisen toiminnan riski suhdanteiden seurauksena.

Yrityksen kasvua voidaan mitata käyttämällä liikevaihdon muutosta kasvun mittarina (Laakso ym. 2010:37). Kasvun mittaaminen voi perustua myös resursseihin, joita ilmentävät esimerkiksi taseen loppusumma tai työntekijöiden määrä. Kasvun laskennassa voidaan käyttää myös usean tilikauden keskikasvua, jolloin saadaan tilikausien välistä vaihtelua tasattua ja tunnusluku kuvaa silloin myös keskimääräisenä paremmin yrityksen pidemmän aikavälin kasvustrategiaa. (Laakso ym. 2010:37.) Kasvutekijän vaikutus maksukyvyttömyyteen perustuu esimerkiksi tulojen ja menojen eriaikaisuuteen. Kustannukset sitoutuvat tuotteisiin jo valmistusvaiheessa, toisin kuin niitä vastaavat tuotot, jotka saadaan vasta kun asiakas on maksanut laskun. Yrityksen kasvaessa nopeasti kuluja

sitoutuu jatkuvasti enemmän, kuin niitä rahoittamaan saadaan tulorahoitusta myynnin kautta. Tällöin seurauksena voi olla maksukyvyttömyys ennen pitkää tulorahoituksen riittämättömyyden vuoksi. (Laakso ym. 2010:36.)

Yrityksen kasvu on vain harvoin tasaista vuosien välillä. Varsinkin projektiluontoiset työt aiheuttavat tuloja epätasaisesti, koska projektien määrä ja laajuus usein vaihtelee. Syklisyys lisää liikeriskiä ja on haaste toiminnan suunnittelulle ja rahoitukselle. Yritys voi yhtenä vuonna tehdä huipputuloksen ja toisena vuonna olla rahoituskriisissä. Kasvu voi lisätä konkurssiriskiä kasvavan tuotevalikoiman kautta. Suuremmalla tuotevalikoimalla voidaan saada enemmän myyntiä, mutta tuotevalikoima sitoo pääomia ja voi aiheuttaa varastojen kiertonopeuden hidastumista. Jos kasvu kääntyy negatiiviseksi, on sen vaikutus positiivista kasvua huomattavasti epäedullisempi. Negatiivinen kasvu supistaa liikevaihtoa ja sitä kautta tulorahoitusta. Tulorahoituksen pieneneminen on hankalaa siirtää yrityksen kustannuspuoleen riittävän nopeasti, jolloin kustannusrakenteesta voi muodostua yritykselle epäedullinen. (Laakso 2010:37.)

2.3.2. Vakavaraisuus ja maksuvalmius

Vakavaraisuus määritellään yrityksen rahoitusrakenteen terveydeksi, jossa vieraan pääoman osuus ei ole hallitseva (Laitinen ja Laitinen 2004:255). Vakavaraisuus kuvaa yrityksen tarvitseman rahoituksen syntytapaa. Rahoitus voi olla peräisin tulorahoituksesta, omasta pääomasta tai vieraasta pääomasta (Laakso 2010:26-27). Tulorahoitus syntyy yrityksen liiketoiminnasta, vieras pääoma rahoittajilta ja oma pääoma sijoittajilta sekä kertyneistä voittovaroista. Tyypillisesti yrityksen rahoitus käsittää jossain suhteessa näitä kaikkia rahoitusmuotoja. Mitä paremmin yritys kannattaa, sitä suurempi osuus menoista voidaan kattaa tulorahoituksella. (Laakso ym. 2010:26-27.)

Yrityksen vakavaraisuutta mitataan varantoperusteisesti ja virtaperusteisesti. Varantoperusteisella vakavaraisuudella tarkoitetaan tietyllä ajanhetkellä laskettua oman ja vieraan pääoman suhdetta. Vakavaraisuuden tunnuslukuna käytetään tyypillisesti omavaraisuusastetta, joka lasketaan oman pääoman osuutena koko taseen loppusummasta. (Laakso ym. 2010:28-29.) Virtaperusteinen vakavaraisuus huomioi vieraan pääoman rasittavan vaikutuksen yrityksen tulorahoituksessa. Tyypillisesti sen laskenta tehdään suhteuttamalla

tulorahoitusta kuvaava rahavirta, kuten rahoitustulos tai rahoitusjäämä, vieraan pääoman maksuvelvoitteisiin ja kustannuksiin (Laakso ym. 2019:30-31).

Maksuvalmiudella tarkoitetaan rahan riittävyyttä maksuvelvoitteiden maksamiseen tietyllä hetkellä. Yrityksen maksuvalmiuden ollessa kunnossa yritys kykenee maksamaan maksuvelvoitteet aina silloin, kun ne erääntyvät maksuun. Vakavaraisuuden tapaan myös maksuvalmiutta voidaan kuvata sekä varantoperusteisesti, että virtaperusteisesti. (Laitinen ja Laitinen 2004:248; Laakso ym. 2010:32.)

Varantoperusteisena tunnusluvut perustuvat nopeasti rahaksi muutettavan, eli likvidin omaisuuden, ja lyhyellä aikavälillä maksettavaksi tulevan vieraan pääoman maksuvelvoitteiden suhteeseen. Tunnuslukuina käytetään quick ratiota ja current ratiota. Tunnuslukujen kriittinen taso riippuu yrityksen toimialasta ja taseen sisältämän vaihto-omaisuuden määrästä. Virtaperusteinen mittaamistapa siihen, miten tulorahoitus riittää kattamaan maksuvelvoitteet (Laakso ym. 2010:33). Mitä enemmän yrityksellä kulloinkin on tulorahoitusta, sitä paremmassa kunnossa on sen virtaperusteinen maksuvalmius (Laitinen ja Laitinen 2004:248). Jos taas yrityksen tulorahoitus ei kata lyhyellä ajanjaksolla voitonjakoa tai lyhytvaikutteisia menoja, sen toiminnan jatkuvuutta voidaan pitää pitkällä aikavälillä heikkona. Tulorahoituksen mittarina voidaan käyttää tuloslaskelmasta suoriteperusteisesti laskettuna rahoitustuloksen suhdetta liikevaihtoon ja rahoituslaskelmasta rahoitusjäämän suhdetta myynnin kassaanmaksujen määrään (Laitinen ja Laitinen 2004:249).

3. KONKURSSIN ENNAKOINNISSA KÄYTETYT MENETELMÄT JA AINEISTOT

3.1.Yhden tunnusluvun mallit

Yhden tunnusluvun mallit olivat ensimmäisiä konkurssin ennakoimiseen käytettyjä menetelmiä. Ensimmäisen konkurssin ennakoimista koskevan tutkimuksen näitä menetelmiä käyttämällä teki Beaver vuonna 1966 (Laitinen ja Laitinen 2004:75). Yhden tunnusluvun mallit ovat nimensä mukaisesti malleja, joissa pyritään yhden tunnusluvun kehitystä seuraamalla ennakoimaan konkurssin todennäköisyyttä. Yhden tunnusluvun analyysi pohjautuu oletukseen, että yksittäisen tunnusluvun jakauma konkurssiyritysten ja terveiden yritysten populaatiossa on systemaattisesti erilainen ja siten tunnusluvun arvo antaa viitteitä yrityksen maksukyvyn tilasta (Laitinen ja Laitinen 2004:127). Mitä suuremmat ovat konkurssiyrityksen ja terveen yrityksen välisten jakaumien erot, sitä parempaan luokittelutarkkuuteen päästään. Jos taas jakaumat ovat päällekkäisiä, seurauksena on todennäköisemmin luokitteluvirheitä. Konkurssia ennakoiviksi tunnusluvuiksi on tärkeää valita sellaisia, jotka kehittyvät tasaisesti ja johdonmukaisesti konkurssin lähestyessä (Beaver 1966). On myös toivottavaa, että tunnusluvun arvo indikoisi konkurssin todennäköisyyttä mahdollisimman aikaisessa vaiheessa (Laitinen ja Laitinen 2004:80).

3.2. Usean tunnusluvun mallit

Usean tunnusluvun malleissa on tarkoituksena sisällyttää usean eri tunnusluvun sisältämä informaatio yhteen yhdistelmälukuun (Laitinen ja Laitinen 2004:131). Yhdistelmätunnuslukuun sisällytettävät yksittäiset tunnusluvut voivat olla täysin samoja, joita yksittäisen tunnuslukumallin tapauksessa on tulkittu erikseen. Muuttujien käyttäytyminen ennen konkurssia noudattaa siis myös lähtökohtaisesti samoja periaatteita, joita yksittäisen tunnusluvun malleissa käytettäviltä muuttujilta edellytetään. Konkurssin lähestyessä yhdistelmäluvun arvon tulee muuttua johdonmukaisesti johonkin tiettyyn suuntaan. Usean

tunnusluvun mallin on havaittu toimivan paremmin ja antavan luotettavampia tuloksia kuin yksittäisen tunnusluvun malli (Laitinen ja Laitinen 2004:131).

Altmanin (1968) tutkimus oli ensimmäisiä, jossa kartoitettiin usean tunnusluvun mallin käyttökelpoisuutta konkurssin ennakoinnissa. Altman (1968) käytti tutkimuksessaan lineaarisen erotteluanalyysia, joka yhdistää usean tilinpäätöstunnusluvun informaation yhteen painotettuna summana laskettavaan lukuun. Altman (1968) nimitti tätä lukua Z-luvuksi. Tutkimus oli luonteeltaan empiirinen ja malliin valittiin selittävät muuttujat sen perusteella, miten voimakkaasti ne korreloivat konkurssiriskin kanssa.

Lineaarinen erotteluanalyysi on menetelmänä yksinkertainen, mutta sen toimivuus perustuu vahvoihin tilastollisiin taustaoletuksiin. Eräs niistä on multinormaalisuusoletus, eli muuttujien jakaumien ja niiden yhteysjakauman oletetaan noudattavan normaalijakaumaa. Normaalijakautuneisuus on edellytys mallin käyttämiselle (Laitinen ja Laitinen 2014:161). Toinen keskeinen tilastollinen olettama mallin taustalla on terveiden yritysten ja konkurssiyritysten välinen kovarianssimatriisin samansuuruisuus. Kovarianssilla mitataan muuttujien yhteisvaikutuksen suuruutta. Jos muuttujat ovat riippuvuudessa keskenään ja riippuvuus voimistuu konkurssiyritysten ja terveiden yritysten välillä, se voi aiheuttaa multikollineaarisuutta. Siinä selittävät muuttujat selittävät osin samaa ilmiötä, mikä vaimentaa niiden selityskykyä (Ranta ym. 2012:420). Tunnuslukujen kertoimet voivat myös kasvaa, jolloin ne indikoivat todellisuutta vahvempaa riippuvuutta (Laitinen ja Laitinen 2004:135).

3.2.1. Logistinen regressio

Logistinen regressio on suosittu ja tehokas luokittelumenetelmä, jota voidaan käyttää samaan tapaan kuin lineaarista regressiota. Logistinen regressio laajentaa kuitenkin lineaarisen regressiomallin luokkamuuttujien tasolle ja binäärisiin luokitteluongelmiin. Logistista regressiota voidaan käyttää aineiston luokitteluun tai profilointiin, jolloin aineistosta etsitään vastemuuttujia kuvaavia riippuvuussuhteita selittävien muuttujien suhteen. (Shmueli, Bruce, Stephens ja Patel 2017:211.)

Logistinen regressio on luonteeltaan ehdollisen todennäköisyyden malli, jossa vastemuuttujaa painotetaan samaan tapaan kuin lineaarisen regressiossa (kaava 1). Vastemuuttujana mallissa on logistisen jakauman linkkifunktio, jonka kautta mallin antama arvo muunnetaan ehdolliseksi todennäköisyydeksi (kaava 2). Mallissa voidaan käyttää lineaarisen regression tapaan useita selittäviä muuttujia. Malli sovitetaan sen epälineaarisen luonteen vuoksi suurimman uskottavuuden menetelmällä, eikä pienimmän neliösumman menetelmällä, kuten lineaarisessa regressiossa. (Laitinen ja Kankaanpää 1999:70.)

$$logit = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

(1)

Missä:

logit= Vastemuuttuja, joka saa arvoja väliltä [-∞ – ∞]
$\beta_0 \ldots \beta_n$ = Yhtälön kertoimet
$X_1 \ldots X_n$ = Yhtälön selittävät muuttujat

Linkkifunktio perustuu logistisen jakauman kertymäfunktioon, josta saatu kertymäfunktion ehdollista todennäköisyyttä kuvaava arvo voidaan pyöristää binäärisiksi luokiksi, jotka saavat joko arvon konkurssi (1) tai terve yritys (0) (kuvio 2; kaava 2). Logistisessa regressiossa ei ole taustalla yhtä vahvoja tilastollisia jakaumaoletuksia kuin lineaarisessa erotteluanalyysissä. Malli on käyttäytymiseltään ja ymmärrettävyydeltään yksinkertainen ja läpinäkyvä. Ymmärrettävyyttä lisää vastemuuttujan muuntaminen ehdolliseksi todennäköisyydeksi. (Laitinen ja Laitinen 2014:161-162.)

$$p = \frac{1}{1 + e^{-logit}}$$

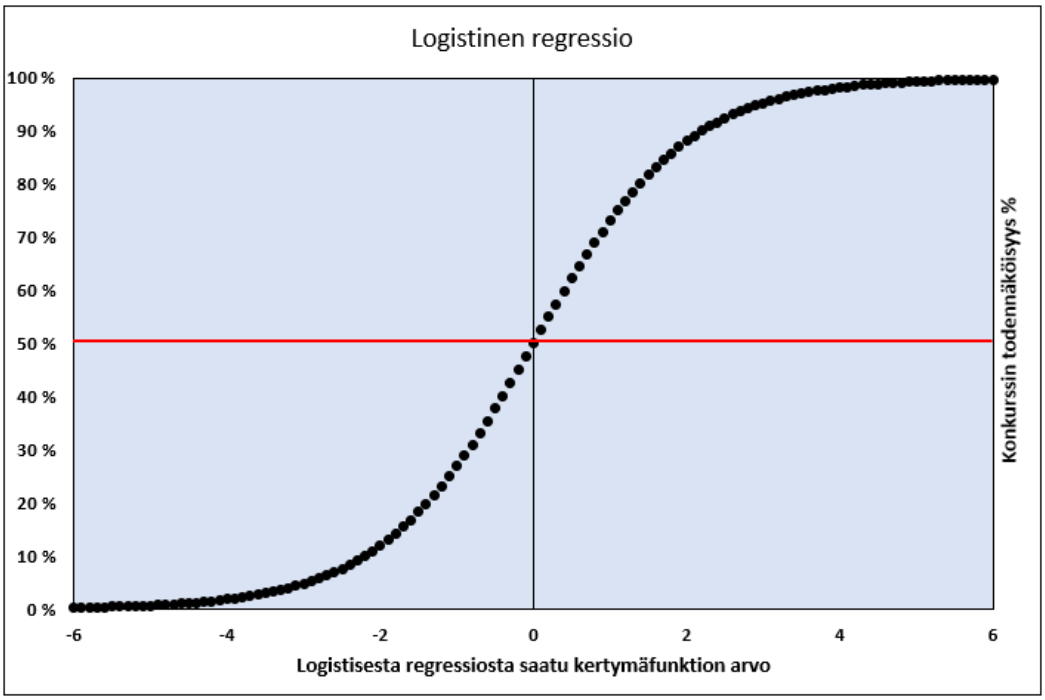
(2)

Missä:

p = todennäköisuus [0-1]
e = Neperin luku (2,718)
logit = vastemuuttujan arvo

Binäärisessä luokittelussa logistisen regression luokittelutarkkuus on korkeimmillaan silloin, kun aineistosta 50 % on konkurssiyrityksiä ja 50 % on terveitä yrityksiä. Tämä

johtuu siitä, että logistinen kertymäfunktio on kohdassa 0,5 laakeimmillaan. Kuviossa 2 tätä 50 % kriittistä todennäköisyyttä on kuvattu punaisella viivalla. Jos malli antaa yritykselle yli 50 % ehdollisen todennäköisyyden olla konkurssiyritys niin yritys luokitellaan tällöin konkurssiyritykseksi. (Laitinen ja Laitinen 2014:162-163.)



Kuvio 2. Logistinen regressio havainnollistettuna.

3.3. Yleisimmät luokittelualgoritmit

Tietotekniikan ja tietokoneiden laskentatehon kehittyminen on kasvattanut luokittelussa käytössä olevien menetelmien kirjoa. Yhtenä erona perinteisten ja koneoppimisen luokittelualgoritmien välillä on, että luokittelualgoritmit sovittautuvat mallinnusaineistoon iteratiivisesti pala palalta, kun taas perinteiset luokittelijat sovitetaan kerralla käyttäen koko aineistoa. Luonnollisesti iteratiivinen sovittaminen vaatii lähtökohtaisesti enemmän laskentatehoa kuin koko aineistoon tehty sovitus. Koneoppimisen algoritmien vahvuudet näkyvät erityisesti juuri luokitteluongelmien tapauksissa. Jos verrataan saatavilla olevien luokittelualgoritmien lukumäärää muihin algoritmeihin, on luokittelualgoritmien osuus noin 67 % kaikista algoritmeista (Amani ja Fadlalla 2017:38).

Neuroverkot ovat yksi tutkimuksissa eniten käytetyistä algoritmityypeistä konkurssin ennakoinnissa. Neuroverkkojen toimivuutta ovat Suomessa kartoittaneet muun muassa Laitinen ja Kankaanpää (1997;1999). Neuroverkkojen tausta-ajatus pohjautuu nimensä mukaisesti biologisen neuroverkon toimintaan, jossa useat neuronit ovat toisistaan riippuvaisia laskennallisia elementtejä ja tasoittain yhteydessä toisiinsa. Neuroverkossa on aina vähintään kolmitasoinen arkkitehtuuri, jossa on informaation syöttötaso (input), piilevä taso (hidden) ja ulostulotaso (output). Tasot ovat laskennallisesti riippuvuudessa keskenään. Informaatio virtaa neuroverkossa eteenpäin ja kulkiessaan se aktivoi eri piirteitä vastaavia neuronien solmukohtia. Neuroverkko koulutetaan halutun luokituksen tunnistamiselle käyttämällä opetusaineistoa. Solmukohdissa informaatioon syötetään muun muassa erilaisia epälineaarisia muunnoksia ennen tiedon siirtymistä seuraavalle tasolle. Seuraavalla tasolla informaatio aktivoi taas muuttuneen painotuksen vuoksi eri solmukohdat, kunnes viimeisellä tasolla aktivoituvan neutronin aktivaatiotaso ilmaisee joko konkurssin tai terveen yrityksen. (Laitinen ja Laitinen 2004:151-155.) Neuroverkkoja on kritisoitu piilevistä tasoista johtuvasta algoritmin tulkinnallisuuden puutteesta, jonka vuoksi algoritmin luokittelun tunnistusmekanismi on erityisen hankala hahmottaa (Chen ym. 2011). Vaikka algoritmin intuitiivisuus onkin heikko, on sillä saatu hyviä tuloksia konkurssin ennakointitarkkuudessa (kts. Laitinen ja Kankaanpää 1997; Laitinen ja Kankaanpää 1999).

Ehdollisiin todennäköisyyksiin pohjautuva luokittelualgoritmi on esimerkiksi Naiivi Bayes -luokittelija. Siinä algoritmi laskee opetusaineistoon pohjautuvasti jokaiselle eri piirteelle esiintymistodennäköisyyden ja opettaa sitä kautta algoritmin tunnistamaan todennäköisyyksien avulla piirteiden ja vastemuuttujan suhteen. Tuntemattomalle havainnolle estimoidaan todennäköisyys piirteiden ja piirreyhdistelmien avulla ehdollisena todennäköisyytenä tuntemattoman havainnon kuulumisesta tiettyyn luokkaan. Shmueli ym. (2017:167-173.) Konkurssin ennakoinnissa Bayesian menetelmällä on saatu luokittelutarkuuden suhteen lupaavia tuloksia aikaisemmassa tutkimuksessa (kts. Aghaie ja Saeedi 2009).

Lineaariset vektoripohjaiset luokittelualgoritmit ovat olleet verrattain suosittuja konkurssin ennakointitutkimuksissa. Tähän ryhmään kuuluvat esimerkiksi SVM (Support Vector Machine) -algoritmit. SVM-algoritmi on toimintaperiaatteeltaan vähän vastaavanlainen kuin runsaasti konkurssin ennakoimisessa perinteisesti käytetty lineaarinen

erottelufunktio. Poikkeuksena SVM-algoritmin toimintaperiaatteessa on kuitenkin se, että yhden lineaarisen erottelufunktion sijaan algoritmi sovittaa aineistoon useita lineaarisia rajoitefunktioita ja pyrkii niillä lokeroimaan aineiston mahdollisimman tehokkaasti. Joustavuutta SVM-algoritmin toimintaan voidaan tuoda helposti lisää hyödyntämällä algoritmissa kernel-painotuksia, jotka muuntavat lineaarisen luokitteluongelman epälineaariseen muotoon. (Barboza ym. 2017.)

Päätöspuupohjaiset (classification tree) algoritmit ovat olleet myös konkurssin ennakoimistutkimuksissa runsaasti käytettyjä menetelmiä. Tähän luokkaan kuuluu muun muassa rekursiivinen osittaminen. Päätöspuuperusteisessa menetelmässä yritykset luokitellaan binäärisesti erikseen jokaisella tunnusluvulla aloittaen parhaiten konkurssia ennakoivasta tunnusluvusta. Tuloksena syntyy haaroittuva rakenne, joka muistuttaa puuta, ja on siten myös helposti ymmärrettävissä (Laitinen ja Laitinen 2004:144.) Haasteena päätöspuissa on mallin ylisovittuminen, jolloin malli ei kuvaa enää itse ilmiötä, vaan sovituksessa käytettyä aineistoa. Kun luokittelu tehdään testiaineistoon, ei malli suoriudu enää yhtä hyvin kuin sovitusaineistossa. (Shmueli ym. 2017:203).

Ylisovittumiseen on Breiman (1996) esittänyt ratkaisuna Bagging-algoritmin, jonka hän myöhemmin yhdisti päätöspuu-menetelmään Random Forest- algoritmissa (Breiman 2001). Random Forest valitsee opetusaineistosta satunnaisen määrän satunnaisen suuruisia ositteita, joihin päätöspuut sovitetaan. Päätöspuun sovittamisessa käytettävät muuttujat valitaan myös jokaisen päätöspuun kohdalla satunnaisesti. Päätöspuita voidaan generoida lukumääräisesti valtavasti ja menetelmän käytännöllisyys syntyy siitä, että lopullinen päätöspuumalli on ikään kuin yhteenveto kaikista satunnaisesti kasvatetuista päätöspuista. Päätöspuiden keskiarvoistuminen vähentää mallin ylisovittumista, mutta mahdollistaa silti aineiston ja itse ennustettavan ilmiön ominaispiirteiden riittävän kattavan huomioimisen osana luokittelumallia. (Shmueli ym. 2017:202-204). Random forest -algoritmia pidetään välinpitämättömänä luokittelijana poikkeavien havaintojen ja aineiston kohinan suhteen (Yeh, Chi ja Lin 2014).

3.4. Aineiston rakenteen vaikutus mallintamiseen ja konkurssin ennakointiin

Konkurssin ennakoinnissa käytettyjä menetelmiä on lukuisia erilaisia. Pääsääntöisesti tutkimuksissa luokittelumenetelmiä on vertailtu keskenään joko luokittelutarkkuuksien suhteen (Barboza ym. 2017) tai luokittelutarkkuuksissa olevan eron tilastollista merkitsevyyden kautta (Laitinen ja Kankaanpää 1997). Yhtenä tärkeänä komponenttina konkurssinennakointimenetelmien hyvyydessä ja mallien suorituskyvyssä on myös käytetty aineisto ja sen rakenne.

Yhtenä konkurssin ennakoinnin tunnuspiirteenä on kiinnostuksen kohteena olevan ilmiön suhteellinen harvinaisuus koko yrityspopulaation näkökulmasta. Tilastokeskuksen (2017) mukaan yritysten määrä Suomessa vuonna 2016 oli noin 350 000 kpl ja vastaavasti samana vuonna konkurssiin haettuja yrityksiä oli Tilastokeskuksen (2016) mukaan noin 2400 kpl. Konkurssiyritysten suhteellinen määrä koko populaatiossa oli siten vain 0,69 %. Monen tilastollisen mallin tapauksessa pätee, että malli ennustaa luotettavimmin populaation odotusarvoa lähinnä olevien havaintojen arvot. Kun otos on kerätty satunnaisesti koko populaatiosta, malli ennustaa silloin parhaiten aineiston keskimääräisen yrityksen ominaisuudet, mutta heikommin populaation ääriarvoihin kuuluvat yritykset. (Ranta ym. 2012:369-378.) Konkurssiyritykset ovat luonteeltaan nimenomaan populaation ääriarvoja.

Tämä tilanne on huomioitu konkurssitutkimuksessa käyttämällä niin kutsuttua **vastinparimenettelyä** aineiston hankinnassa (kts. Beaver 1966; Laitinen 1990:40). Vastinparimenettelyssä lähdetään liikkeelle etsimällä mahdollisimman paljon konkurssiyrityksiä ja vasta tämän jälkeen löydetyille konkurssiyrityksille etsitään terveiden yritysten joukosta kokoluokaltaan ja toimialaltaan mahdollisimman samankaltaiset yritykset. Mallintamisen kannalta siis ollaan ideaalitilanteessa, kun mallinnusaineisto sisältää konkurssiyrityksiä ja terveitä yrityksiä suhteessa 50/50 %. Mallin soveltaminen varsinaisessa käytännön tilanteessa edellyttää, että päätöksentekijällä on käytössään ennakkotietoa yrityksestä ja perusteltu epäily yksittäisen yrityksen todennäköisyydestä olla konkurssiyritys tulevaisuudessa. Sen sijaan, että päätöksentekijä ratkaisee luokitteluongelman heittämällä siitä kolikkoa, voi päätöksentekijä tehdä valistuneen arvauksen konkurssin ennakointimallin avulla. Ongelmana sovellustilanteessa on kuitenkin ennakkotiedon laatu, eli antaako ennakkotieto riittävästi informaatiosta sen pohjaksi, että yksittäisestä yrityksestä tiedetään sen olevan 50 % todennäköisyydellä konkurssiyritys. Vastinparimenettelyä käyttämällä populaation rakenteen vääristyminen voi vaikuttaa osaltaan luokitteluvirheiden

harhaisuuteen. Harhaisuuden määrän on kuitenkin havaittu olevan suhteellisen pientä (Zmijewski 1984 viitattu Laakso ym. 2010:68).

Toinen konkurssitutkimuksen haaste on vaatimus konkurssiyrityksen tilinpäätöstietojen täydellisestä saatavuudesta. Yritykset pyrkivät lähtökohtaisesti kaunistelemaan tai jopa jättämään joitain tietoja julkaisematta tilanteessa, jossa tilinpäätöksen osoittama elinkelpoisuus alkaa heikentyä. Konkurssitutkimuksen näkökulmasta tämä voi aiheuttaa ongelman tulosten yleistettävyyden näkökulmasta, koska mallintamisessa käytetyt tunnusluvut ja aineisto voi olla vääristynyt. (Laitinen ja Laitinen 2004:126.)

Tutkimuksissa on havaittu, että tunnuslukuihin perustuvat mallit eivät kykene yksiselitteisesti erottamaan toisistaan konkurssiyritystä ja heikosti menestyvää yritystä (Gilbert, Menon ja Schwartz 1990). Gilbert ym. (1990:167-168) mukaan mallinnusaineistossa on mallin luokittelutarkkuuden kannalta suotavampaa käyttää konkurssiyritysten lisäksi satunnaisesti valittuja terveitä yrityksiä. Jos mallinnusaineistossa on ylikorostuneesti konkurssin partaalla olevia yrityksiä konkurssiyritysten ohella, mallin selityskyky kärsii merkittävästi ja tuloksissa esiintyy runsaasti luokitteluvirheitä, joissa terve yritys luokitellaan konkurssiyritykseksi. Mallin luokittelukyvyn heikkeneminen johtuu siitä, että konkurssiyrityksen ja terveen yrityksen tunnusluvut voivat olla lähellä toisiaan. (Taffler 1982:355; Laitinen ja Laitinen 2004:126).

Aineistoon sisällytettävien yritysten lukumäärä aikaisemmissa konkurssitutkimuksissa on vaihdellut melko runsaasti. Merkittävä osa aikaisemmista tutkimuksista on pohjautunut suhteellisen pieniin aineistokokoihin, kuten alle 100 konkurssiyrityksen määriin (kts. Beaver 1966; Altman 1968; Edminister 1972; Deakin 1972; Blum 1974; Diamond 1976; Martin 1977; Hamer 1983; Tam ja Kiang 1992). Laadukkaan konkurssitutkimuksen edellytyksenä ei siten välttämättä ole yritysten suuri määrä aineistossa. Toki laajempiakin aineistoja, jotka ovat sisältäneet useita satoja tai tuhansia konkurssiyrityksiä, on ollut käytössä (kts. Park ja Han 2002; Ribeiro ym. 2008; Barboza ym. 2017). Suomalaisissa tutkimuksissa on ollut pääasiassa alle 100 konkurssiyritystä aineistossa (Prihti 1975; Laitinen 1991; Back, Laitisen ja Seren 1996). Suurimpia aineistoja oli käytössä Kiviluodolla (1998) ja Kaskella, Sinkkosella ja Peltosella (2001), joilla yrityksiä oli yhteensä yli 1000 kpl, mutta konkurssiyrityksiä 150-300 kpl. Kumarin ja Ravin (2007) tekemän koonnostutkimuksen perusteella tutkimuksissa käytetyn aineiston koon mediaani sisältäen

sekä konkurssiyritykset, että terveet yritykset, on ollut noin 200 yritystä. Tiettävästi laajimmassa aineistossa on ollut 8977 kpl yrityksiä (Kolari, Glennon, Shin ja Caputo 2002).

Aineistoon sisällytettävien yritysten lukumäärän vaikutusta mallin hyvyyteen ja luotettavuuteen on selvitetty esimerkiksi Pietruszkiewiczin (2008) tutkimuksessa. Menetelmänä Pietruszkiewicz (2008) käytti k:n lähimmän naapurin menetelmää. Siinä havaittiin, että konkurssimallin tarkkuus paranee systemaattisesti opetusaineiston koon kasvaessa 20 yritykseen asti. Tämän jälkeen suuremmalla opetusaineiston koolla ei havaittu olevan luokittelutarkkuuteen enää vaikutusta. Tutkimuksen perusteella luotettava ja yleistettävissä oleva konkurssin ennakointimalli voitiin saavuttaa jo hyvin pienellä aineistokoolla.

Muuttujavalinnan tulisi aina lähtökohtaisesti perustua teoriaan konkurssiyrityksen tunnuslukujen kehittymisestä konkurssin lähestyessä. Teoreettinen tausta auttaa muuttujavalinnan perustelussa ja niiden avulla johdetut tulokset ovat lisäksi lähtökohtaisesti sattumasta ja otoksesta riippumattomia, taustalla olevan teoreettisen pohjan vuoksi. Käytännössä kuitenkin suurin osa konkurssitutkimuksista ei perusta muuttujavalintaa teoreettisiin perusteisiin vaan empiriaan. (Laitinen ja Laitinen 2004:168-169.) Muuttujavalinnassa tulisi pystyä suodattamaan vaihtoehtoisten muuttujien joukosta esiin kaikkein parhaiten konkurssia ennakoivat muuttujat. Kaikkia muuttujia ei tule sisällyttää malliin, sillä suuri määrä muuttujia aiheuttaa nk. dimensiokirouksen (*curse of multidimensionality*), joka vääristää yksittäisten muuttujien ennustuskykyä (Tsai 2009).

Selittävien muuttujien valinnassa voidaan käyttää esimerkiksi erilaisia korrelaatiokertoimia, kuten Pearsonin tai Spearmanin korrelaatioita, joiden avulla voidaan selvittää muuttujien riippuvuus kiinnostuksen kohteena olevaan vastemuuttujaan. Tunnuslukujen riippuvuutta voidaan kartoittaa myös esimerkiksi faktorianalyysillä tai yksisuuntaisella varianssianalyysillä (Park ja Han 2002). Faktorianalyysillä voidaan suodattaa sellaiset selittävät muuttujat mallin ulkopuolelle, jotka ovat korreloituneita keskenään ja lisäävät sitä kautta riskiä multikollineaarisuuteen (Laitinen ja Laitinen 2004:173).

Logistisen regression tapauksessa on myös suhteellisen helppoa valita soveltuvat muuttujat askeltavaa (stepwise) regressiota hyödyntäen, jossa muuttujat voidaan valita malliin mukaan automaattisesti niiden tilastolliseen merkitsevyyteen perustuen. Rannan ym. (2012:422) mukaan taaksepäin kulkevaan regressiomalliin lisätään kaikki saatavilla

olevat muuttujat. Tämän jälkeen malli poistaa automaattisesti aina yhden sellaisen muuttujan mallista, jota vastaava t-testin arvo on kaikista pienin. Malli jatkaa iteraatioita niin kauan, kunnes muuttujan poistaminen ei enää paranna mallia kokonaisuutena.

Koneoppimisen algoritmeja on kehitetty muuttujavalintaa tueksi ja niiden hyödyt korostuvat varsinkin tilanteessa, jossa selittävien muuttujien määrä on hyvin suuri (Kursa ja Rudnicki 2010). Esimerkiksi Random forest -algoritmiin (Breiman 2001) perustuvan Kursan ja Rudnickin (2010) kehittämän Boruta-algoritmin on todettu olevan tehokas suodattamaan mallinnuksen kannalta merkityksettömät muuttujat ulos mallista. Algoritmi lisäksi validoi ja laskee vaihtoehtoisille muuttujille niitä vastaavat z-testin arvot, jotka helpottavat muuttujien keskinäisen hyvyyden vertailussa.

Laitisen ja Laitisen (2014:164) mukaan on tärkeää, että tilinpäätösmuuttujat valitaan siten, että ne kattavat yhdessä kuusi yhteistä taloudellisten toimintaedellytysten ulottuvuutta. Näitä ovat kannattavuus, vakavaraisuus, varantoperusteinen maksuvalmius, virtaperusteinen maksuvalmius, koko ja kasvu. Muuttujien suhteen on myös tärkeää, että ne eivät ole korreloituneita keskenään, eli ne eivät sisällä päällekkäistä informaatiota. Kannattavuuden mittarina suositellaan käytettäväksi **sijoitetun pääoman tuottosuhdetta**, sillä se on eniten käytetty kannattavuuden mittari ja suhteuttaa tuloslaskelman tuloksen taseen sitoutuneeseen pääomaan. Vakavaraisuuden tunnusluvuista paras on **omavaraisuusaste**, joka mittaa varantoperusteista vakavaraisuutta. Varantoperusteisena maksuvalmiuden tunnuslukuna voidaan käyttää **quick ratiota**, joka kuvaa helposti realisoitavissa olevan varallisuuden riittämistä lyhytaikaisiin velkoihin. Virtaperusteisena maksuvalmiuden tunnuslukuna voidaan käyttää **rahoitustulosprosenttia**, joka osoittaa liikevaihdosta lyhytvaikutteisten kulujen ja voitonjaon jälkeen yritykselle jäävän osuuden. Yrityksen koon osalta paras tunnusluku on yrityksen **liikevaihto**, joka kuvaa yrityksen toiminnan laajuutta. Liikevaihdon jakauma on tunnuslukuna vinoutunut, joten on suositeltavaa käyttää mallissa liikevaihdon luonnollista logaritmia muuttujana. Vastaavasti kasvun tunnuslukuna voidaan käyttää **liikevaihdon positiivista tai negatiivista muutosta**. (Laitinen ja Laitinen 2014:164.) Malliin voidaan sisällyttää myös tärkeimpien tase-erien kiertonopeutta kuvaavia tunnuslukuja (Strang 2000:81-82).

4. K:N LÄHIMMÄN NAAPURIN MENETELMÄ JA KONKURSSIN ENNAKOINTI

4.1. Menetelmän toimintaperiaate

K:n lähimmän naapurin menetelmän tausta-ajatus hahmoteltiin ensimmäisen kerran Coverin ja Hartin (1967) tutkimuksessa. Algoritmi on toimintaperiaatteeltaan yksinkertainen ja luonteeltaan ei-parametrinen, jolloin sen käyttäminen ei edellytä oletuksia vastemuuttujan tai selittävien muuttujien tilastollisista jakaumista (Lantz 2013:66). Algoritmi on hyvin monipuolinen sovellettavuutensa puolesta ja sitä voidaan käyttää suhteellisten, numeeristen, binääristen ja luokkamuuttujien kanssa sekä luokittelussa, että regressiossa. Konkurssitutkimuksen ohella k:n lähimmän naapurin menetelmää on käytetty laskentatoimeen liittyvässä tutkimuksessa aikaisemmin esimerkiksi ennustettaessa osakemarkkinoiden ja valuuttakurssien kehittymistä, pankkien asiakasprofilointiin, rahanpesuepäilyjen analysointiin ja luottoluokituksen ennustamiseen (Imandoust ja Bolandraftar 2013:609).

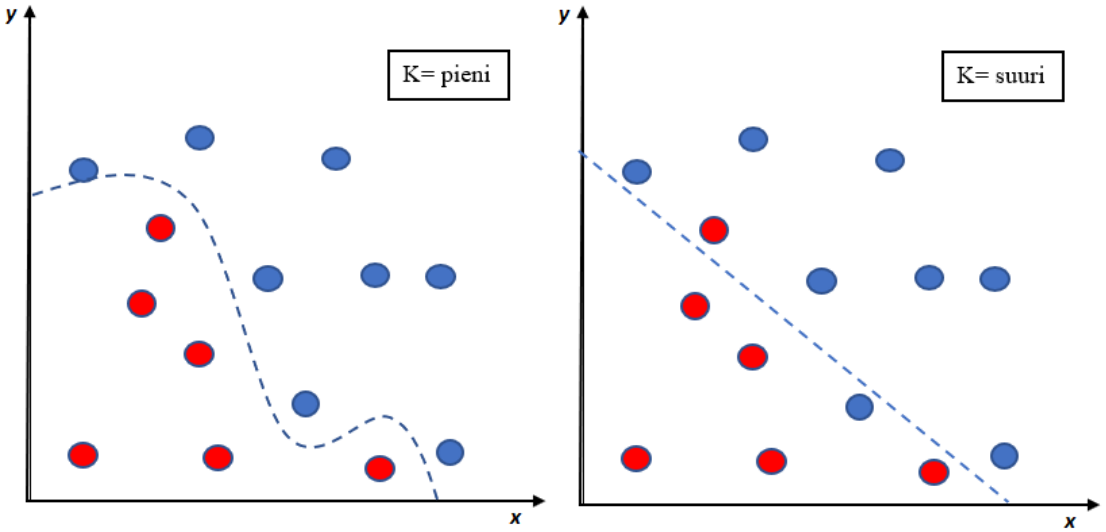
Konkurssin ennakoimisessa k:n lähimmän naapurin menetelmää on tutkittu suhteellisen vähän sekä kansainvälisissä tutkimuksissa, että Suomessa. Luokittelumenetelmänä se on osittain jäänyt muiden koneoppimisen luokittelumenetelmien, kuten SVM:n (Support Vector Machine), Random forest -algoritmin ja neuroverkkojen varjoon. K:n lähimmän naapurin menetelmässä on kuitenkin useita hyviä puolia, kuten algoritmin intuitiivisuus ja toimintaperiaatteen helppo ymmärrettävyys. Algoritmi toimii lisäksi hyvin joustavasti ja tehokkaasti alhaisilla muuttujamäärillä ja k:n arvoilla, jolloin se ei vaadi suuria määriä laskentatehoa (Lantz 2013:66).

K:n lähimmän naapurin menetelmä luokittelee tuntemattoman havainnon hyödyntäen tiettyä määrää piirteiltään riittävän samankaltaisia tunnettuja havaintoja. Algoritmi koulutetaan mallinnusaineistoon vastaavalla tavalla kuin tavallinen lineaarinen tai logistinen regressio. Ero tavalliseen regressioon ilmenee k:n lähimmän naapurin menetelmällä tehdyssä ennustamisessa siten, että algoritmi käyttää ennustamisessa kiinteää määrää piirreavaruudessa olevia lähimpiä tunnettuja havaintoja. Siinä missä logistinen regressio

käyttää tuntemattoman havainnon ennustamiseen koko aineistoa, johon se on sovitettu, k:n lähimmän naapurin menetelmä käyttää vain muutamaa kaikkein lähimpänä olevia havaintoja. Tässä mielessä k:n lähimmän naapurin menetelmä voi toimia hyvin joustavasti ja lokaalisti. (Lantz 2013:67-69.)

K:n lähimmän naapurin menetelmässä on sen toiminnan kannalta kolme tärkeää mukautettavissa olevaa parametria. Ensimmäinen on k:n arvo, joka tarkoittaa ennustamisessa käytettyä naapureiden määrää. Naapureiksi nimitetään piirreavaruudessa tuntematonta havaintoa lähinnä olevia tunnettuja havaintoja. Toisena keskeisenä algoritmin parametrina on käytetty etäisyysfunktion muoto, jolla lähin etäisyys naapureihin määritetään. Kolmantena parametrina on lähimpien naapureiden painotus, jolla määritetään, miten suuri painoarvo tunnettujen havaintojen arvoille ennustamisessa annetaan. Varsinainen tuntemattoman havainnon ennustaminen perustuu mainittujen parametrien perusteella tehtyyn äänestykseen, jossa tuntematon havainto saa arvon lähimpien naapureiden perusteella. (Hechenbichler ja Schliep 2004).

Naapureiden määrä voi käytännössä vaihdella välillä 1-N, jossa N tarkoittaa populaation kokoa. Käytettäessä naapureiden määränä arvoa 1, algoritmi toimii hyvin joustavasti, kun tuntematon havainto saa arvonsa suoraan piirreavaruudessa lähinnä sitä olevalta tunnetulta havainnolta (Lantz 2013:71). Jos k:n arvo on suuri, luokittelu keskiarvoistuu ja algoritmin joustavuus alenee, kun tuntemattomalle havainnolle lasketaan arvo usean naapurin perusteella (kuvio 3). Ääriesimerkkinä naapureiden määrästä on koko aineiston sisältävien havaintojen lukumäärä, jolloin tuntematon havainto saa arvonsa keskiarvona koko tunnetun aineiston perusteella. Lähtökohtaisesti k:n arvona voidaan käyttää lukua väliltä 3-10 kpl. K:n arvon tulisi lisäksi aina lähtökohtaisesti olla pariton luku. Tämä johtuu siitä, että binäärisessä luokittelussa algoritmi toimii ikään kuin äänestämällä kahden vaihtoehdon väliltä. (Lantz 2013:72.)



Kuvio 3. K:n lähimpien naapureiden määrän vaikutus luokittelun joustavuuteen.

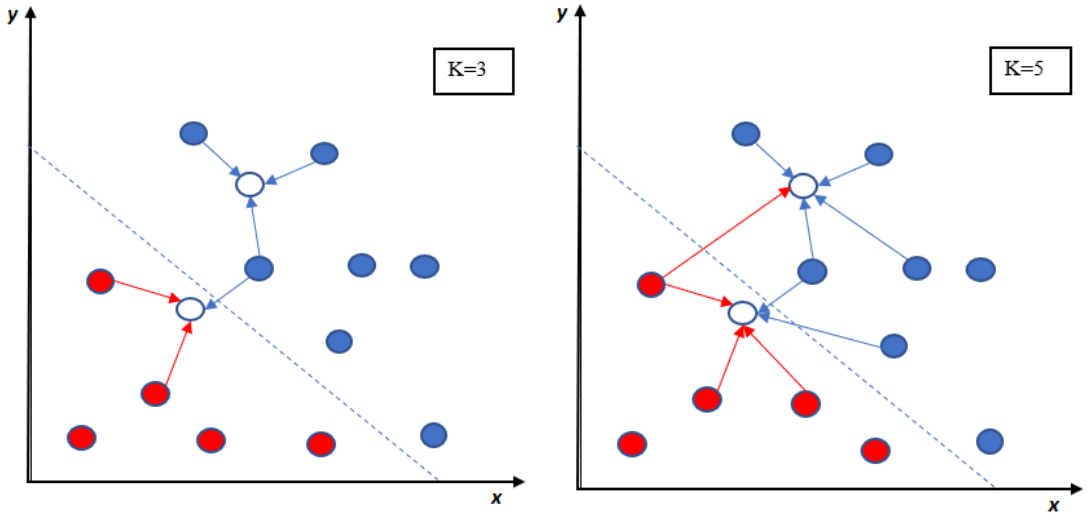
Naapureiden määrän lisäksi toinen keskeinen asia on, miten algoritmi etsii lähimmät naapurit piirreavaruudesta. K:n lähimmän naapurin menetelmässä yleisin etäisyysfunktion typpi on euklidinen etäisyys, jolla algoritmi laskee etäisyyden piirreavaruudessa havaintojen välillä (Lantz 2013:70). Euklidinen etäisyys perustuu Pythagoraan lauseeseen ja se voidaan ilmaista matemaattisesti kaavan 3 osoittamalla tavalla. Siinä alaindeksi 1 kuvaa ensimmäistä selittävää muuttujaa, alaindeksi 2 toista selittävää muuttujaa ja niin edelleen aina viimeiseen selittävään muuttujaan (n) saakka. Jotta kaikki selittävät muuttujat vaikuttaisivat etäisyysfunktion toimintaan samalla tavalla, täytyy muuttujat ennen etäisyyden laskemista normalisoida, jolloin sisäinen vaihteluväli on kaikilla muuttujilla sama (Lantz 2013:73).

$$d = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \quad (3)$$

Missä:

d = Euklidinen etäisyys
q_n= Tuntematonta havaintoa vastaavan piirteen (n) arvo.
p_n= Tunnettua havaintoa vastaavan piirteen (n) arvo.

Kuviossa 4 on havainnollistettu, miten tuntemattomalle havainnolle on etsitty 3 tai 5 lähintä naapuria. Kuvaajassa esitetyt värilliset pisteet kuvaavat tunnettuja havaintoja, jotka ovat joko konkurssiyrityksiä (punainen) tai terveitä yrityksiä (sininen). Värittömät pisteet ovat tuntemattomia havaintoja. Tuntemattoman havainnon luokittelussa naapureiden merkittävyyttä voidaan painottaa erilaisilla kernel-painotuksilla. Kernel-funktiot ovat matemaattisia yhtälöitä, joiden avulla johdetaan naapureille painotuskertoimet perustuen naapureiden väliseen etäisyyteen (Hechenbichler ja Schliep 2004).



Kuvio 4. Algoritmin toimintaperiaate tuntemattoman havainnon luokittelussa

4.2. K:n lähimmän naapurin menetelmä aikaisemmassa tutkimuksessa

Vanhin tutkimus, joka tätä tutkielmaa varten oli tietokannoista löydettävissä ja jossa lähimmän naapurin menetelmää tutkittiin konkurssin ennakoinnissa, oli Tamin ja Kiangin (1992) tutkimus. Siinä aineisto koostui 59 konkurssipankista ja niille vastinparimenettelyllä etsityistä terveistä pankeista. Luokittelutarkkuudeksi (kaava 4) saatiin vuotta ennen konkurssia 77,2 % (k=1) ja 77,2 % (k=3). K:n arvon kasvattamisella ei siten havaittu olevan merkittävää vaikutusta luokittelutarkkuuteen vuotta ennen konkurssia. Kahta vuotta ennen konkurssia sen sijaan luokittelutarkkuudeksi saatiin 77,5 % (k=1) ja 80 % (k=3). Vastaavasti samassa aineistossa logistisella regressiolla luokittelutarkkuudeksi saatiin 81,8 % vuotta ennen konkurssia ja 92,5 % kahta vuotta aikaisemmin (taulukko 1).

$$Luokittelutarkkuus\ (ACC) = \left(\frac{A+B}{N}\right) * 100 \quad (4)$$

missä

A = oikein luokitellut konkurssiyritykset (kpl)
B= oikein luokitellut terveet yritykset (kpl)
N = Kaikki yritykset yhteensä (kpl)

Yksi lukumäärältään suurin tutkimusaineisto oli käytössä Parkin ja Hanin (2002) tekemässä tutkimuksessa. Konkurssiyrityksiä oli 1072 kpl, joille oli vastinparimenettelyllä etsitty vastaavat terveet yritykset. Tutkimuksessa verrattiin perusmuotoista k:n lähimmän naapurin menetelmää analyyttisen hierarkiaprosessin avulla painotettuun k:n lähimmän naapurin menetelmään. Analyyttisen hierarkiaprosessin rakennetta on kuvannut muun muassa Pukkala (2007:198-200). Menetelmässä päätösongelma jaotellaan hierarkiatasoihin, joissa jokaisen tason kriteerit voidaan arvottaa niin, että ne summautuvat hierarkiatasoittain arvoon 1. Tulosten perusteella AHP:llä painotetulla k:n lähimmän naapurin menetelmällä saatiin luokittelutarkkuudeksi 84,5 %, kun vertailussa käytetyllä painottamattomalla menetelmällä saatiin 74,0 %. Tutkimuksen perusteella muuttujien keskinäistä hyvyyttä koskevalla ennakkotiedolla näyttäisi siten olevan vaikutusta luokittelutarkkuuteen (taulukko 1).

Tuhansia konkurssiyrityksiä sisältävä tutkimusaineisto oli käytössä myös Ribeiron ym. (2008) tutkimuksessa. Aineistossa oli tilinpäätöstietoja 60000 yrityksen verran vuosilta 2002-2006, joista vuonna 2007 oli mennyt konkurssiin 3000 yritystä. Paras luokittelutarkkuus saatiin käyttämällä kolmea lähintä naapuria. Vuotta ennen konkurssia luokittelutarkkuus oli 85,8 % ja kahta vuotta ennen 76,9 %. Vertailussa käytetyllä SVM-menetelmällä (*Support vector machine*) saatiin vuotta ennen konkurssia luokittelutarkkuudeksi 90,5 % ja kahta vuotta ennen 81,6 %. (taulukko 1.)

Myös monissa muissa tutkimuksissa k:n lähimmän naapurin menetelmällä on saatu lupaavia tuloksia luokittelutarkkuuksien suhteen. Yli 70 % luokittelutarkkuuksia saatiin muun muassa Pietruszkiewiczin (2008) tutkimuksessa ja yli 85 % tarkkuuksia Yun ym. (2009) tutkimuksessa. Serrano-Cinca ja Gutiérrez-Nieto (2013) sen sijaan raportoivat yli 94 % luokittelutarkkuuksia, joihin verrattuna referenssinä käytetyllä logistisella

regressiolla saatiin luokittelutarkkuudeksi hieman yli 95 %. Parhaana naapureiden määränä tutkimuksissa on ollut esimerkiksi 3 kpl (Tam ja Kiang 1998; Ribeiro ym. 2008), 15 kpl (Kiviluoto 1998) ja 30 kpl (Yu ym. 2009). Optimaalinen naapurien lukumäärä on riippuvainen tutkittavasta aineistosta ja tulee siksi määrittää aina aineistokohtaisesti. (taulukko 1.)

Kiviluoto (1998) on tehnyt ainoan löydetyn konkurssitutkimuksen suomalaisella aineistolla, jossa yhtenä vertailtavana menetelmänä käytettiin k:n lähimmän naapurin menetelmää. Luokittelutarkkuudeksi saatiin tutkimuksessa k:n lähimmän naapurin menetelmällä 91,5 %, joka oli muihin tutkittuihin menetelmiin verrattuna luokittelutarkkuudeltaan paras. Referenssinä käytetyllä lineaarisella erotteluanalyysillä saatiin tarkkuudeksi 89,5%, joten suurta eroa menetelmien välillä ei kuitenkaan ollut. (taulukko 1.)

Osassa tutkimuksia on pyritty kehittelemään yhdistelmämenetelmiä, joissa eri luokittelualgoritminen parhaat puolet on pyritty yhdistämään saman menetelmän sisään. Esimerkiksi Gao, Cui ja Po ym. (2008) selvittivät k:n lähimmän naapurin yhdistettävyyttä SVM-algoritmiin (*Support Vector Machine*). He huomasivat, että lähimmän naapurin menetelmä paransi pelkällä SVM:llä laskettua luokittelutarkkuutta 86,8 % tasolta 92,2 % tasolle (taulukko 1).

Chen ym. (2011) sen sijaan yhdisti k:n lähimmän naapurin menetelmään PSO-optimointialgoritmin (*Particle Swarm Optimization*) ja globaalin optimin etsintää tehostavan adaptiivisen lisäosan, joka estää algoritmia kiinnittymästä lokaaleihin optimeihin. Yhdistelmäalgoritmissa käytetty k:n lähimmän naapurin algoritmi hyödynsi lisäksi sumeaa logiikkaa (fuzzy logic). Sumeaa logiikkaa on käytetty aikaisemminkin osana k:n lähimmän naapurin menetelmää (Keller, Gray ja Givens 1985; Bian ja Mazlack 2003). Optimointialgoritmit pyrkivät adaptiivisesti etsimään optimaalista lähimpien naapureiden määrää ja samaan aikaan sumea logiikka lisää etäisyysfunktion toimintaan enemmän joustavuutta ja mahdollistaa joustavammat naapureiden määrät. Tutkimuksen tulokset osoittivat, että yhdistelmäalgoritmilla (PTVPSO-FKNN) saatiin luokittelutarkkuudeksi 87,1 %, kun tavallisella lähimmän naapurin menetelmällä tarkkuus jäi 85,8 %:iin (Chen ym. 2011). Suurta eroa menetelmien välillä ei siten todellisuudessa havaittu (taulukko 1).

Zhao, Huang, Wei, Yu, Wang ja Chen (2017) käyttivät tutkimuksessaan samaa yhdistelmäalgoritmia ja mittasivat sille luokittelutarkkuudeksi 81,6 %. Luokittelutarkkuutta verrattiin Random forest-algoritmiin ja logistiseen regressioon, joiden tarkkuudet olivat samalla aineistolla 80,8 % ja 74,6 %. Aikaisemmin muun muassa Bian ja Mazlack (2003) ovat tutkineet sumean logiikan integroimista osaksi k:n lähimmän naapurin menetelmää. Heillä ei kuitenkaan tuolloin ollut algoritmissa käytössä erillistä optimointiosaa. Tutkimuksessa saatiin siitä huolimatta luokittelutarkkuudeksi FKNN-menetelmälle 84 % (taulukko 1).

Taulukko 1. KNN-algoritmia käsitteleviä tutkimuksia ja niiden tuloksia. Virhetyyppi 1 kuvaa terveitä yrityksiä, jotka luokiteltiin konkurssiyrityksiksi. Virhetyyppi 2 kuvaa konkurssiyrityksiä, jotka luokiteltiin terveiksi yrityksiksi. ACC on luokittelutarkkuus.

Tutkimus	**Menetelmä**	**Virhe-tyyppi 1 %**	**Virhe-tyyppi 2 %**	**ACC %**
Tam ja Kiang (1992)	Knn (k=1)	30,5 %	14,8 %	77,4 %
Tam ja Kiang (1992)	Knn (k=3)	33,2 %	9,6 %	78,6 %
Tam ja Kiang (1992)	Logit	23,4 %	2,3 %	87,2 %
Kiviluoto (1998)	Knn	75,2 %	1,5 %	91,5 %
Park ja Han (2002)	Knn			74,1 %
Park ja Han(2002)	AHP-Knn			84,5 %
Bien ja Mazlack (2003)	fuzzy knn			84,0 %
Ribeiro ym. (2008)	Knn	11,0 %	26,8 %	87,1 %
Pietruszkiewicz (2008)	Knn	11,7 %	16,7 %	71,7 %
Yu ym. (2009)	Knn			87,5 %
Chen ym. 2011	Knn	12,1 %	16,2 %	85,8 %
Chen ym. 2011	PTVPSO-FKNN	13,2 %	12,26 %	87,1 %
Serrano-Cinca ja Gutiérrez-Nieto (2013)	Knn	36,7 %	5,3 %	94,0 %
Serrano-Cinca ja Gutiérrez-Nieto (2013)	Logit (Stepwise)	38,3 %	3,9 %	95,4 %
Zhao ym. (2016)	PTVPSO-FKNN	20,3 %	17,0 %	81,3 %
Zhao ym. (2016)	Logit	21,1 %	30,4 %	74,6 %
Alrashed ja Che (2018)	Knn	0,0 %	6,0 %	97,0 %
Alrashed ja Che (2018)	Logit	37,0 %	12,0 %	78,0 %

Alrasheed ja Che (2018) ovat tutkineet konkurssiyritysten ja terveiden yritysten määrien vaikutusta eri luokittelumenetelmillä. He käyttivät tutkimuksessaan kolmea eri aineistoa, joissa jokaisessa oli suhteellisesti eri määrät konkurssiyrityksiä ja terveitä yrityksiä. Yhdessä suhde vastasi konkurssiyritysten suhdetta kaikkiin yrityksiin (6,9/93,1 %), toisessa aineistossa konkurssiyrityksiä oli 30 % - ja kolmannessa 50 % suhteessa koko aineiston yritysmäärään. Tutkimuksessa oli mukana k:n lähimmän naapurin menetelmä ja logistinen regressio. Luokittelun tarkkuutta mitattiin virhetyyppien 1 ja 2 suhteellisilla osuuksilla (taulukko 2). Virhetyyppi 1:llä tarkoitetaan tervettä yritystä, joka on luokiteltu konkurssiyritykseksi ja virhetyyppi 2:lla taas konkurssiyritystä, joka on luokiteltu terveeksi yritykseksi. Taulukosta 2 voidaan nähdä, että virhetyyppien 1 suhteellinen määrä korostuu aineistossa, jossa konkurssiyritysten suhde vastaa koko yrityspopulaatiossa oleva konkurssimäärää. K:n lähimmän naapurin menetelmällä parhaimpaan luokittelutarkkuuteen päästään jo 30 % konkurssiyrityksiä sisältävällä otossuhteella. Logistisen regression tapauksessa paras luokittelutarkkuus saavutetaan otoksella, jossa on 50 % konkurssiyrityksiä.

Taulukko 2. Otossuhteiden vaikutus luokittelun tuloksiin. Virhetyyppi 1 kuvaa terveitä yrityksiä, jotka luokiteltiin konkurssiyrityksiksi. Virhetyyppi 2 kuvaa konkurssiyrityksiä, jotka luokiteltiin terveiksi yrityksiksi. (Alrasheed ja Che 2018)

Menetelmä	**Otossuhde %**	**Virhetyyppi 1 %**	**Virhetyyppi 2 %**	**Luokittelutarkkuus %**
Knn	6,9/93,1	76,0 %	1,4 %	61,3 %
Knn	30/70	0,1 %	6,0 %	96,9 %
Knn	50/50	0,0 %	6,0 %	97,0 %
Logit	6,9/93,1	95,0 %	0,2 %	52,4 %
Logit	30/70	65,0 %	3,0 %	66,0 %
Logit	50/50	37,0 %	12,0 %	75,5 %

Muuttujamäärillä havaittiin olevan vain vähän vaikutusta mallin luokittelutarkkuuteen Alrasheed ja Che 2018). Tutkimuksessa testattiin k:n lähimmän naapurin menetelmällä kahta eri suuruista selittävien muuttujien yhdistelmää. Ensimmäinen testatuista muuttujavalinnasta sisälsi kaikki saatavilla olevat tunnusluvut, joita oli lukumääräisesti 64 kpl. Toisessa muuttujavalinnassa käytettiin viittä Pearsonin korrelaatioiden pohjalta valittua muuttujaa. Selittävien muuttujien lukumäärällä havaittiin olevan suurin merkitys luokittelutarkkuuteen, kun aineistossa oli konkurssiyritysten suhde terveisiin yrityksiin pienimmillään (6,9 %). Tällöin ero virhetyyppi 1 suhteellisessa osuudessa oli noin 12 % suurempi käytettäessä kaikkia saatavilla olevia selittäviä muuttujia mallissa. Vaikutus virhetyyppi 2:n osuuteen oli hyvin marginaalinen, mutta päinvastainen. Kaikkia muuttujia käytettäessä oli virhetyyppi 2 osuus noin 0,6 % vähemmän kuin viidellä parhaalla selittävällä muuttujalla. Kun aineiston konkurssiyritysten suhdetta nostettiin ensin 30 %:iin ja sitten 50 %:iin, erot virhetyyppien suhteiden välillä hävisivät käytännössä kokonaan. Tutkimuksen mukaan selittävien muuttujien määrällä ei näytä olevan k:n lähimmän naapurin menetelmää käytettäessä erityisen merkittävää vaikutusta. (Alrasheed ja Che 2018.)

4.3. Yhteenveto ja tutkimuksen hypoteesit

K:n lähimmän naapurin menetelmää konkurssin ennakoinnissa käsitteleviä tutkimuksia on tehty määrällisesti suhteellisen vähän, mutta kuitenkin riittävästi esikuvatutkimusten löytämisen näkökulmasta. Aikaisemman tutkimuksen läpikäynnin yhteydessä onnistuttiin löytämään reilu 10 kansainvälistä tutkimusta ja yksi suomalaisella aineistolla tehty tutkimus. Monet kansainvälisistä tutkimuksista ovat yli kymmenen vuoden takaa. Kiviluodon (1998) ainoana Suomessa tekemä tutkimus on yli 20 vuoden takaa. Uuden ajankohtaisen tutkimuksen tarve suomalaisella aineistolla voidaan tästä syystä nähdä perusteltuna.

Aikaisemmissa tutkimuksissa esitetyt tulokset niin virhetyyppien suhteellisten määrien sekä luokittelutarkkuuden osalta osoittavat k:n lähimmän naapurin menetelmän olevan konkurssin luokittelijana lupaava (taulukko 1). Myös verrattaessa k:n lähimmän naapurin suorituskykyä muihin luokittelumenetelmiin, kuten esimerkiksi logistiseen regressioon, on k:n lähimmän naapurin menetelmä luokittelutarkkuudeltaan hyvin samalla tasolla. Tutkimusten perusteella on kuitenkin hankala ottaa kantaa siihen, kumpi menetelmä on

toistaan parempi, sillä niistä esitetyt tulokset ovat osittain ristiriitaisia johtuen oletettavasti erilaisista aineistojen rakenteista, menetelmien selittävistä muuttujista ja parametreista. Osassa tutkimuksia on saatu paremmat luokittelutulokset k:n lähimmän naapurin menetelmällä ja osassa logistisella regressiolla. Tässä tutkimuksessa on tarkoitus selvittää, kumpi menetelmistä kykenee ennakoimaan konkurssia tarkemmin. K:n lähimmän naapurin luokittelutarkkuutta ja suorituskykyä verrataan logistiseen regressioon, koska sitä on käytetty runsaasti konkurssin ennakoinnissa suorituskykynsä ja intuitiivisuutensa vuoksi. Tutkimuksen ensimmäiseksi hypoteesiksi muodostettiin aikaisemman tutkimuksen perusteella nollahypoteesi, koska aikaisempi tutkimus ei yksiselitteisesti osoita kummankaan menetelmän paremmuutta toiseen verrattuna.

> **H1**: K:n lähimmän naapurin menetelmän ja logistisen regression luokittelutarkkuuksissa ei ole tilastollisesti merkitsevää eroa menetelmien välillä.

Yhtenä haasteena aikaisemmassa tutkimuksessa on esitettyjen tulosten vertailukelpoisuus tutkimusten välillä. Tutkimuksissa esitetyt luokittelutulokset sisältävät runsaasti hajontaa ja menetelmien keskinäinen paremmuus ei ole aina selkeästi todennettavissa. Pääosin tämän voidaan olettaa johtuvan aineistoon liittyvistä rakenteellisista tekijöistä. Aineistoissa voi olla toimialaltaan ja kooltaan erilaisia yrityksiä. Myös konkurssiyritysten määrän suhde terveisiin yrityksiin on vaihdellut tutkimusten välillä runsaasti (kts. Beaver 1966; Altman 1968; Kiviluoto 1998; Ribeiro ym. 2008). Osassa tutkimuksia on käytetty vastinparimenettelyä, jolla otossuhde on ollut 50/50 % (kts. Altman 1968; Tam ja Kiang 1992) ja osassa tutkimuksista terveitä yrityksiä on ollut enemmän suhteessa konkurssiyrityksiin (kts. Kiviluoto 1998; Ribeiro ym. 2008; Alrasheed ja Che 2018). Alrasheed ja Che (2018) ovat tutkineet aineiston rakenteen ja otossuhteen vaikutusta luokittelutarkkuuteen ja heidän tuloksensa osoittivat, että luokittelutarkkuus on osaltaan riippuvainen aineiston rakenteesta ja otossuhteesta. He havaitsivat, että k:n lähimmän naapurin menetelmä on vähemmän herkkä otossuhteen muutoksille verrattuna logistiseen regressioon. Tämän pohjalta muodostettiin tutkimuksen toinen tutkittava hypoteesi:

H2: K:n lähimmän naapurin menetelmän luokittelutarkkuus on vähemmän riippuvainen otossuhteen vaihtelusta kuin logistisen regression luokittelutarkkuus

K:n lähimmän naapurin menetelmän toimintaan ja suorituskykyyn voidaan vaikuttaa esimerkiksi lähimpien naapureiden lukumäärällä, etäisyysfunktion tyyppiä muuttamalla tai erilaisia naapureiden painotuksia (kernel) käyttämällä. Aikaisempaa tutkimusta, joka kartoittaisi näiden eri parametrien vaikutusta luokittelutarkkuuteen ei juurikaan ole. Parametreistä ainoastaan lähimpien naapureiden optimaalinen lukumäärä on sellainen, joka tutkimusten tulosten yhteydessä on mainittu. Esimerkiksi parhaimpana naapurien määränä on pidetty 3 kpl (Tam ja Kiang 1998; Ribeiro ym. 2008), 15 kpl (Kiviluoto 1998) ja 30 kpl (Yu ym. 2009). Varsinaisesti kuitenkaan tutkimuksissa ei ole raportoitu sitä, miten suuri vaikutus naapureiden lukumäärällä on tuloksiin. Tähän perustuen muodostettiin kolmas tämän tutkimuksen hypoteesi, joka asetettiin nollahypoteesina, koska aikaisempi tutkimus ei selkeästi osoita tiettyä optimaalista naapurien lukumäärää:

H3: Lähimpien naapureiden lukumäärällä ei ole tilastollisesti merkitsevää vaikutusta k:n lähimmän naapurin menetelmän luokittelutarkkuuteen.

Yhtenä mahdollisuutena tehostaa k:n lähimmän naapurin menetelmän suorituskykyä on esitetty myös muuttujien painotus (Park ja Han 2002). Painottaminen on mahdollista silloin, jos on olemassa ennakkotietoa muuttujien keskinäisestä paremmuudesta. Esimerkiksi Park ja Han (2002) tutkivat vaihtoehtoisia tapoja painotuksen pohjaksi, joissa huomioitiin muun muassa analyyttisen hierarkiaprosessin (AHP) kautta johdettu painotus ja logistisen regression kertoimien t-testin arvojen kautta johdettu painotus. Parkin ja Hanin (2002) tutkimuksen mukaan muuttujien painotus paransi luokittelutarkkuutta 68,3 %:sta (painottamaton) 83,0 %:iin (AHP-painotus). Tämän pohjalle muodostettiin tutkimuksen neljäs hypoteesi:

H4: Selittävien muuttujien ennakkotietoon pohjautuvalla painotuksella on tilastollisesti merkitsevä vaikutus k:n lähimmän naapurin menetelmän luokittelutarkkuuteen.

5. AINEISTO JA MENETELMÄT

5.1. Yritysaineiston hankinta ja rakenne

Tutkimuksessa käytetyt tilinpäätösaineistot ja tunnusluvut koostettiin Orbis-tietokannasta. Aineiston keräämisessä yhtenä merkittävänä haasteena oli konkurssiyritysten tunnistaminen ja konkurssiajankohdan todentaminen. Yliopiston tunnuksilla käytettävissä olevissa tietokannoissa (Orbis ja Voitto+) oli mahdollista suodattaa yrityksistä esiin ”lakanneet” yritykset. Konkurssiyritysten osuus näistä lakanneista yrityksistä havaittiin Yritys- ja yhteisötietojärjestelmässä (YTJ) tehdyn validoinnin perusteella olevan ainoastaan noin 2 %. YTJ:ssä näkyvä yrityksen tämänhetkinen status ei lisäksi kaikkien yritysten kohdalla ollut todenmukainen ja näytti jostain syystä yrityksen tilan olevan aktiivinen, vaikka kolmannesta lähteestä voitiin todentaa yrityksen olevan konkurssissa.

Edellä mainituista syiden takia päätettiin aineiston keräämiseen liittyen kysyä apua Asiakastieto Oyj:stä, joka toimii Suomessa yritystiedon kerääjänä, analysoijana ja yksityisen yritysrekisterin ylläpitäjänä. Asiakastieto Oyj luovutti tämän tutkimuksen käyttöön kaikki osakeyhtiö- ja kommandiittiyhtiömuotoiset yritykset, jotka olivat menneet konkurssiin vuonna 2017 ja joista oli saatavilla tilinpäätöstietoja. Näitä yrityksiä oli kaiken kaikkiaan 1402 kpl. Asiakastiedon luovuttama aineisto sisälsi konkurssiyritysten nimet, y-tunnukset, sekä toimialat. Tutkimuksen kannalta Asiakastiedosta saatu tieto konkurssiyrityksistä oli merkittävä etu, sillä sen perusteella voitiin konkurssistatusta ja sen tarkkaa voimaantuloajankohtaa pitää varmana.

Kun konkurssin alkamisajankohdan tiedettiin olevan vuosi 2017, päätettiin yrityksille hakea tilinpäätöstunnusluvut aikavälille 2013-2016. Yritykset etsittiin Orbiksesta käyttäen tietokannan joukkohakutoimintoa. Kaikki saatavilla olevat tunnusluvut ajettiin yrityskohtaisesti tietokannasta ulos. Yritysten välillä havaittiin olevan suuria eroja tunnuslukujen saatavuudessa (taulukko 3). Läheskään kaikilla konkurssiyrityksillä ei ollut halutulle ajankohdalle saatavilla kaikkia tunnuslukuja. Tunnuslukujen täydellisyysvaatimuksen ollessa ehdoton, aiheutti tunnuslukujen alhaisten saatavuuksien yhteisvaikutus merkittävän rajoittavan tekijän saatavilla olevan aineiston kokoon. Lopullinen konkurssiyritysaineisto supistui alkuperäisestä 1402 konkurssiyrityksestä 86 yritykseen.

Taulukko 3. Tilinpäätöstunnuslukujen saatavuus suhteessa konkurssiyrityksiin. Suhteet kuvaavat osuuksia konkurssiyrityspopulaatiosta, joista tunnusluku on ollut saatavilla.

Tunnusluku	2016	2015	2014	2013	Keskimäärin
Liikevaihto €	56,0 %	77,1 %	68,9 %	61,2 %	65,8 %
Tase €	46,6 %	66,4 %	63,1 %	54,7 %	57,7 %
Tulos €	46,4 %	66,3 %	63,1 %	54,7 %	57,6 %
Current ratio €	45,9 %	64,8 %	61,6 %	53,0 %	56,3 %
Myyntisaamisten kiertoaika pv	44,1 %	63,8 %	60,5 %	52,4 %	55,2 %
Ostovelkojen kiertoaika pv	43,9 %	63,3 %	60,0 %	52,1 %	54,8 %
Liikevaihto %	42,2 %	60,4 %	57,9 %	50,4 %	52,7 %
ROA %	36,0 %	58,0 %	58,0 %	51,3 %	50,8 %
Omavaraisuusaste %	30,1 %	51,9 %	53,9 %	48,4 %	46,0 %
Rahoitustulos %	35,3 %	52,9 %	49,9 %	43,2 %	45,3 %
Ebitda %	33,6 %	50,3 %	48,2 %	41,7 %	43,5 %
ROCE %	16,6 %	34,1 %	36,1 %	33,4 %	30,0 %
ROE %	11,1 %	26,6 %	31,8 %	31,2 %	25,2 %
Quick ratio %	7,6 %	19,3 %	24,0 %	23,5 %	18,6 %

Moni tunnusluku, joka oli aikaisemmassa tutkimuksessa havaittu hyväksi konkurssin ennakoijaksi, osoittautui voimakkaasti rajoittavaksi, kuten esimerkiksi Quick ratio ja sijoitetun pääoman tuotto % (ROCE) (taulukko 3). Quick ration osalta tunnusluvun rajoittava vaikutus oli niin voimakas, että se jouduttiin jättämään tästä syystä kokonaan aineiston ulkopuolelle. Lopulliseen aineistoon valittiin tunnusluvuiksi taulukossa 4 esitetyt 14 kpl tunnuslukuja. Tunnusluvut kerättiin pääasiassa vuosilta 2014-2016, mutta liikevaihdon osalta mukaan otettiin myös yrityskohtaiset liikevaihdot vuodelta 2013. Vuoden 2013 liikevaihtoa hyödynnettiin kasvumuuttujan laskemisessa vuodelle 2014.

Taulukko 4. Tutkimukseen valitut tunnusluvut.

Tunnusluvun luonne	Tunnusluku	2016	2015	2014	2013
Yrityksen koko	Liikevaihto €	x	x	x	x
	Tase €	x	x	x	
Kasvu	Liikevaihdon kasvu %	x	x	x	
	Liikevaihdon kasvu €	x	x	x	
Kannattavuus	Tulos €	x	x	x	
	Tulos %	x	x	x	
	ROCE %	x	x	x	
	Ebitda %	x	x	x	
	Ebitda €	x	x	x	
	Rahoitustulos €	x	x	x	
	Rahoitustulos %	x	x	x	
	ROA %	x	x	x	
Maksuvalmius	Current ratio	x	x	x	
Vakavaraisuus	Omavaraisuusaste %	x	x	x	

Liikevaihto ja tase kuvaavat yrityksen suuruusluokkaa. Liikevaihdon avulla voitiin laskea kasvumuuttuja tilikausien välille. Kasvumuuttuja on mukana tunnusluvuissa sekä euromääräisenä, että liikevaihtoon suhteutettuna. Kannattavuutta kuvaavista tunnusluvuista mukana on tilikauden tulos euromääräisenä ja liikevaihtoon suhteutettuna. Sijoitetun pääoman tuotto % (ROCE) kuvaa yrityksen suhteellista kannattavuutta, eli tuottoa, joka on saatu yritykseen sijoitetulle korolle tai muulle tuottoa vaativalle pääomalle (kaava 5). Käyttökate (EBITDA) lasketaan lisäämällä liiketulokseen tilikauden aikaiset poistot ja arvonalentumiset (kaava 6). Muuttujissa on käyttökate mukana euromääräisenä, sekä liikevaihtoon suhteutettuna. Rahoitustulos kuvaa tuottoa, jonka tulee riittää lyhyellä aikavälillä lainan lyhennyksiin, investointien rahoitukseen ja voitonjakoon omalle pääomalle (kaava 7). Muuttujissa on rahoitustulos mukana sekä euromääräisenä, että liikevaihtoon suhteutettuna. Kokonaispääoman tuotossa (ROA) verrataan tulosta ennen rahoituskuluja ja veroja koko siihen pääomaan, joka taseen kautta on sitoutunut yritystoimintaan (kaava 8). Oikaistussa taseen loppusummassa huomioidaan mahdolliset taseessa olevat poistoerot, vapaaehtoiset varaukset ja oman pääoman oikaisut. Tunnusluku mittaa yrityksen kykyä tuottaa tulosta kaikelle toimintaan sitoutuneelle pääomalle. (Yritystutkimus 2011.)

$$ROCE\ \% = \frac{Nettotulos + rahoituskulut + verot}{Sijoitettu\ p\ddot{a}\ddot{a}oma\ keskim\ddot{a}\ddot{a}rin\ tilikaudella} * 100$$

(5)

$$Ebitda\ € = Liiketulos + poistot\ ja\ arvonalentumiset$$

(6)

$$Rahoitustulos\ € = Nettotulos + poistot\ ja\ arvonalentumiset$$

(7)

$$ROA\ \% = \frac{Nettotulos + rahoituskulut + verot}{Oikaistu\ taseen\ loppusumma\ keskimäärin\ tilikaudella}$$

(8)

Maksuvalmiutta kuvaavana tunnuslukuna tutkimuksessa käytetään Current ratiota, jossa oletetaan yrityksen rahoitusomaisuuden lisäksi vaihto-omaisuuden olevan helposti realisoitavissa lyhyellä ajanjaksolla käytettäväksi velkojen maksuun (kaava 9). Tunnusluvussa jakajana käytetään taseen lyhytaikaista vierasta pääomaa, joka erääntyy maksettavaksi alle vuoden aikajänteellä. Aikaisemman tutkimuksen perusteella current ratiota parempana konkurssia ennakoivana muuttujana on todettu olevan quick ratio, joka on muuten sama, kuin current ratio, mutta osoittajassa ei ole mukana vaihto-omaisuutta (Laitinen ja Laitinen 2014:164). Quick ratio toimii lähtökohtaisesti current ratiota paremmin tilanteessa, jossa vertaillaan eri toimialojen yrityksiä. Koska mahdollisuutta quick ration käyttämiseen ei ollut, jouduttiin tyytymään current ratioon. Vakavaraisuuden tunnusluvuista saatiin tunnuslukuihin mukaan omavaraisuusaste, joka lasketaan suhteuttamalla yrityksen oma pääoma taseen loppusummaan oikaistuna saaduilla ennakkomaksuilla (kaava 10). Omavaraisuusasteen on todettu olevan tunnusluvuista parhaiten konkurssia ennakoiva (Laitinen ja Laitinen 2014:164). (Yritystutkimus 2011.)

$$Current\ ratio = \frac{Vaihto.omaisuus + rahoitusomaisuus}{Lyhytaikainen\ vieras\ pääoma}$$

(9)

$$Omavaraisuusaste = \frac{Oikaistu\ taseen\ oma\ pääoma}{Taseen\ loppusumma - saadut\ ennakot}$$

(10)

Terveiden yritysten otantaa varten konkurssiyritykset jaettiin liikevaihdon suhteen kymmeneen eri liikevaihtoluokkaan (taulukko 5). Konkurssiyrityksiä oli mukana kahta

luokkaa lukuun ottamatta kaikissa kokoluokissa. Frekvenssiltään suurimpina liikevaihtoluokkina olivat 1-3 M€ (18 kpl), 750-1000 t€ (11 kpl), 250-750 t€ (29 kpl) ja 100-250 t€ (11 kpl), jotka sisälsivät 80 % kaikista aineiston konkurssiyrityksistä.

Taulukko 5. Konkurssiyritysaineiston rakenne liikevaihtoluokittain.

Liikevaihtoluokka	**Konkurssiyrityksiä luokassa (kpl)**	**Osuus %**
yli 100 M€	1	1 %
50-100 M€	0	0 %
10-50 M€	5	6 %
6-10 M€	0	0 %
3-6 M€	7	8 %
1-3 M€	18	21 %
750-1000 t€	11	13 %
250-750 t€	29	34 %
100-250 t€	11	13 %
alle 100 t€	4	5 %
Yhteensä	**86**	**100 %**

Terveiden yritysten valinta suoritettiin vastaaviin liikevaihtoluokkiin perustuvasti, jolla voitiin varmistua aineiston rakenteellisesta samankaltaisuudesta. Otanta mukailee tältä osin aikaisemmassa konkurssitutkimuksessa käytettyä vastinparimenettelyä, jossa konkurssiyrityksille pyritään löytämään vastinparit koon ja toimialan mukaan (Beaver 1966; Laitinen 1990:40). Terveet yritykset etsittiin Orbis-tietokannasta erikseen jokaista liikevaihtoluokkaa vastaavilla kokorajoituksilla ja niille tallennettiin samat tunnusluvut kuin konkurssiyrityksille (taulukko 4). Hakuehdoksi haettaville terveille yrityksille asetettiin lisäksi tilikauden tuloksen keskimääräinen positiivisuus viimeisen 4 vuoden aikana, jolla voitiin alentaa riskiä, että terveiden yritysten joukkoon tulisi mukaan konkurssiyrityksiä. Muita rajoitteita haettavien terveiden yritysten kannattavuudelle, maksuvalmiudelle tai pääomarakenteelle ei asetettu, jotta haettavien terveiden yritysten rakenteessa ja tunnusluvuissa olevaa luonnollista hajontaa ei rajoiteta liikaa. Liian voimakkaat rajoitukset haettavissa yrityksissä voivat eriyttää terveitä ja konkurssiyrityksiä liikaa ja aiheuttaa sen vuoksi harhaiset luokittelutulokset.

Terveitä yrityksiä tallennettiin tietokannasta jokaista liikevaihtoluokkaa kohden tuhansia, jotta yritysten valinta voitiin tämän jälkeen suorittaa satunnaisotannalla. Terveitä yrityksiä poimittiin satunnaisotannalla kahden erillisen aineiston koostamiseksi ensin 86

yritystä ja tämän jälkeen 774 yritystä (taulukko 6). Tällöin valituista terveistä yrityksistä voitiin muodostaa konkurssiyritysten kanssa kaksi aineistoa, joista toisessa konkurssiyritysten määrä suhteessa terveisiin yrityksiin on 50 % ja toisessa 10 %. Konkurssiyritysten määrä kummassakin aineistossa on vakio (86 kpl). Lukumääräisesti pienempi aineisto sisältää 172 kpl yrityksiä ja suurempi 860 kpl yrityksiä (taulukko 6).

Taulukko 6. Terveiden yritysten otanta.

Liikevaihtoluokka	**Terveiden yritysten otanta (kpl)**	
yli 100 M€	1	9
50-100 M€	0	0
10-50 M€	5	45
6-10 M€	0	0
3-6 M€	7	63
1-3 M€	18	162
750-1000 t€	11	99
250-750 t€	29	261
100-250 t€	11	99
alle 100 t€	4	36
Terveitä yrityksiä (kpl)	**86**	**774**
Aineisto yhteensä (kpl)	**172**	**860**
Konkurssiyrityksiä (%)	**50 %**	**10 %**

Taulukossa 7 on havainnollistettu aineistoissa olevien konkurssiyritysten rakennetta tunnuslukujen keskiarvoilla mitattuna terveiden yritysten vastaaviin keskiarvoihin. Taulukosta voidaan nähdä joitain selkeitä eroavaisuuksia konkurssiyritysten ja terveiden yritysten välillä. Konkurssiyrityksillä esimerkiksi tilikauden tulos, omavaraisuusaste, sijoitetun pääoman tuotto, kokonaispääoman tuotto ja rahoitustulos ovat keskimäärin negatiivisia koko kolmen vuoden ajan ennen konkurssia. Myös liikevaihdon, taseen ja käyttökatteen voidaan havaita olevan laskusuuntaisia konkurssia edeltävästi. Omavaraisuusaste on terveillä yrityksillä noin 40 % tuntumassa, kun konkurssiyrityksillä se on alle 10 % ja vuotta ennen konkurssia negatiivinen.

Kasvumuuttujan käyttäytyminen konkurssiyrityksillä on sen sijaan mielenkiintoinen. Kasvunopeus on liki kolminkertainen kolme vuotta ennen konkurssia verrattuna terveisiin yrityksiin. Korkean kasvun todettiin aineiston tarkemmalla tarkastelulla johtuvan muutamasta yrityksestä, jotka kasvavat noin 10-12 kertaisella kasvuvauhdilla verrattuna muihin yrityksiin. Vaikka aineistossa on mukana paljon nääntyviäkin yrityksiä, joilla

vuosittainen kasvu on negatiivista, nostaa muutaman yrityksen nopeampi kasvuvauhti koko aineiston keskimääräisen kasvun positiiviseksi. Jos aineistosta laskettaisiin keskiarvon sijasta mediaani, se olisi negatiivinen.

Vastinparimenettelyn onnistumista voidaan tarkastella yritysten kokoa kuvaavalla liikevaihdolla ja taseella. Taseen osalta molemmat terveiden yritysten aineistot ovat hyvin samassa kokoluokassa kuin konkurssiyritykset. Liikevaihdolla tarkasteltuna 50/50 % aineiston terveet yritykset kuvaavat keskimäärin paremmin konkurssiyrityksiä kuin 10/90 % aineiston terveet yritykset. Havaittu ero keskimääräisissä liikevaihdoissa on tässäkin tapauksessa peräisin muutamasta poikkeavasta havainnosta, jotka nostavat keskimääräistä liikevaihtoa.

Taulukko 7. Aineiston rakennetta kuvaavat keskiarvot. Eri tunnuksista konkurssiaineistossa ja kahdessa terveiden yritysten aineistossa.

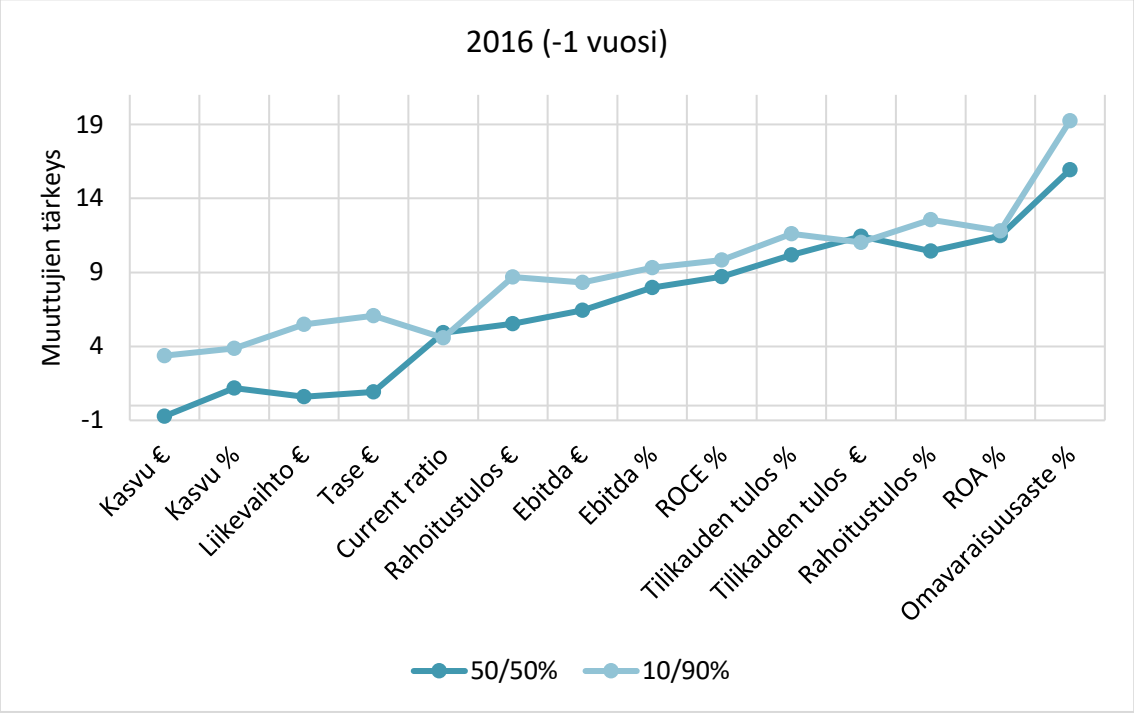
	Konkurssiyritykset			**Terveet yritykset (50 %)**			**Terveet yritykset (90 %)**		
Muuttuja	**2016**	**2015**	**2014**	**2016**	**2015**	**2014**	**2016**	**2015**	**2014**
Liikevaihto t€	3301,7	3762,0	4562,8	4168,9	4187,3	3995,8	5059,3	4742,2	4584,6
Tilikauden tulos t€	-850,0	-71,0	-338,6	231,2	247,7	154,2	210,1	226,4	153,0
Tase t€	3177,8	4133,0	3745,4	3530,0	3223,4	3335,4	3760,7	3654,9	3739,8
Omavaraisuusaste %	-2,3 %	10,9 %	13,9 %	42,9 %	38,5 %	35,1 %	40,8 %	38,2 %	35,8 %
Current ratio	1,18	1,49	1,52	2,27	2,16	2,20	2,30	2,19	2,24
Roce %	-51,2 %	-3,4 %	-8,5 %	16,9 %	18,4 %	8,6 %	12,3 %	14,4 %	12,0 %
ROA %	-14,8 %	-2,6 %	-5,5 %	8,3 %	8,9 %	5,2 %	7,6 %	7,6 %	6,8 %
Ebitda %	-5,4 %	3,0 %	2,2 %	13,3 %	14,0 %	11,8 %	13,9 %	14,6 %	13,7 %
Rahoitustulos %	-10,7 %	-0,8 %	-1,1 %	10,4 %	11,6 %	9,1 %	11,0 %	11,6 %	10,6 %
Kasvu %	1,8 %	29,1 %	33,6 %	2,0 %	7,9 %	10,5 %	5,3 %	6,6 %	18,0 %

5.2. Selittävien muuttujien valinta

Selittävät muuttujat valittiin erikseen kolmelle ajankohdalle hyödyntäen R-ohjelmaa (R Development Core Team 2018) ja Boruta-algoritmia (Miron, Kursa, Witold ja Rudnicki 2010). Boruta-algoritmi pohjautuu Random forest -luokittelualgoritmiin. Algoritmi

monistaa aineiston sisältämät selittävät muuttujat, asettaa monistetut muuttujat satunnaiseen järjestykseen ja käyttää niitä tunnuslukujen validoinnissa varjomuuttujina. Algoritmi sovittaa Random forest -luokittelijan aineistoon ja etsii päätöspuiden avulla konkurssia parhaiten selittävät muuttujat. Algoritmi laskee tunnusluvuille tilastollisen merkitsevyyden z-testin avulla ja asettaa ne tällä perusteella paremmuusjärjestykseen. (kuviot 5, 6, 7).

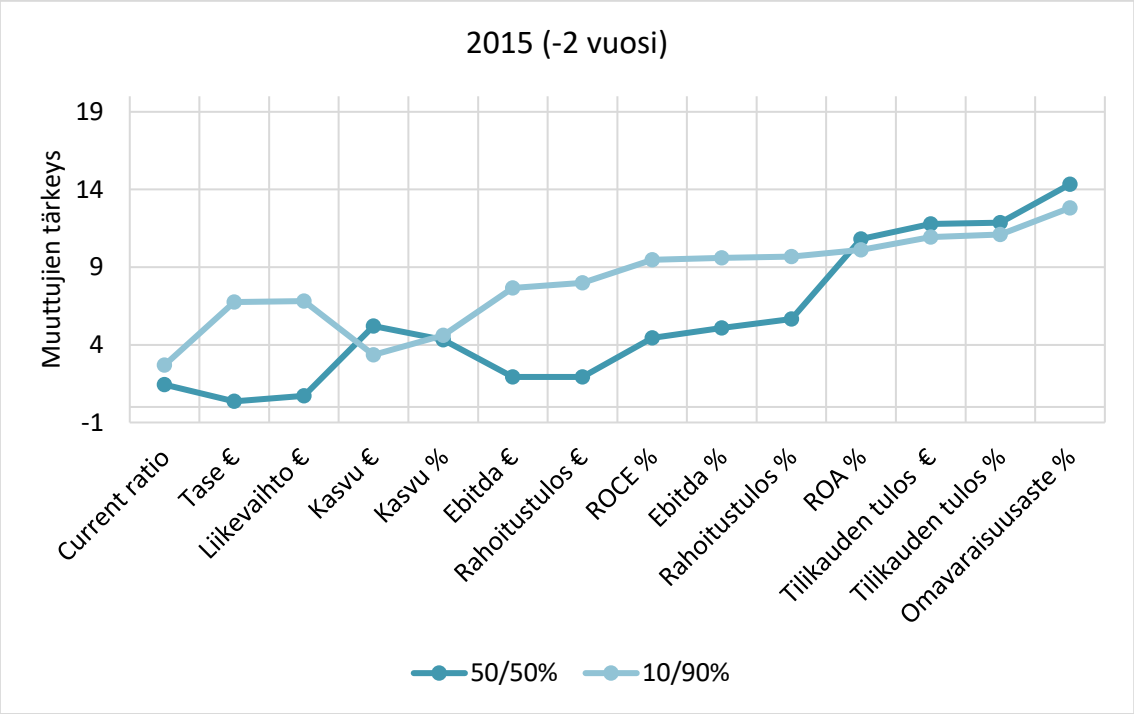
Kuviossa 5 on esitetty muuttujavalidoinnin tulokset vuodelle 2016, mikä vastaa yhtä vuotta ennen konkurssihetkeä. Selkeästi parhaiten konkurssia ennakoivana muuttujana on omavaraisuusaste-%, joka on linjassa aikaisemman tutkimuksen kanssa (kts. Laitinen ja Laitinen 2014:164). Toiseksi parhaimpana muuttujana on rahoitustulos-%, kolmanneksi parhaimpana kokonaispääoman tuotto-% (ROA) ja neljänneksi parhaimpana tilikauden tulos euroina.



Kuvio 5. Muuttujavalidointi vuotta ennen konkurssihetkeä. Muuttujat järjestetty siten, että niiden keskimääräinen tärkeys kasvaa kuvaajan oikeaan laitaan.

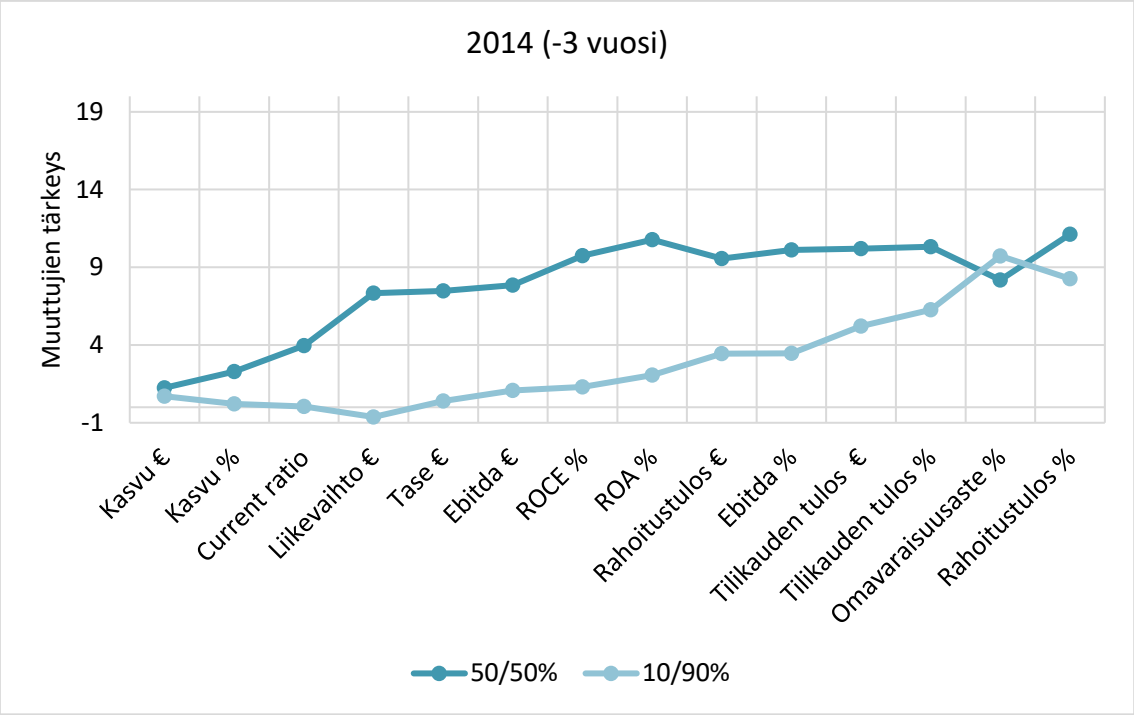
Kuviossa 6 on muuttujavalidoinnin tulokset vuodelle 2015, joka vastaa kahta vuotta ennen konkurssihetkeä. Selkeästi parhaiten konkurssia ennakoivana muuttujana on

omavaraisuusaste-%. Toiseksi ja kolmanneksi parhaimpina muuttujina ovat tilikauden tulos euromääräisenä ja suhteellisena. Neljänneksi ja viidenneksi parhaimmat muuttujat ovat kokonaispääoman tuottoprosentti-% (ROA) ja rahoitustulos-%.



Kuvio 6. Muuttujavalidointi kahta vuotta ennen konkurssihetkeä. Muuttujat järjestetty siten, että niiden keskimääräinen tärkeys kasvaa kuvaajan oikeaan laitaan.

Kuviossa 7 on muuttujavalinnan tulokset vuodelle 2014, eli kolmea vuotta ennen konkurssihetkeä. Muuttujien paremmuuksia tarkasteltaessa voidaan selkeästi huomata, että mitä kauempana konkurssihetkestä ollaan, sitä vähemmän muuttujien välisissä tilastollisissa merkitsevyyksissä on havaittavissa eroa. Otossuhteet selkeästi myös vaikuttavat lisäämällä kohinaa muuttujavalidoinnin tuloksiin. 10/90 % aineiston osalta muuttujat saivat korkeampia z-arvoja, mutta niiden keskinäinen hyvyys ei juurikaan poikennut 50/50 % aineistosta. Parhaimpina konkurssia ennakoivina muuttujina näyttäisivät olevan rahoitustulos-% ja omavaraisuusaste-%. Kolmanneksi ja neljänneksi parhaimmat muuttujat olivat tilikauden tulos euroina ja suhteellisena. Viidenneksi paras muuttuja oli käyttökate-% (EBITDA).



Kuvio 7. Muuttujavalinta kolmea vuotta ennen konkurssihetkeä. Muuttujat järjestetty siten, että niiden keskimääräinen tärkeys kasvaa kuvaajan oikeaan laitaan.

Selittävien muuttujien joukossa on yksi vakavaraisuuden tunnusluku ja yksi maksuvalmiuden tunnusluku. Lisäksi kokoa ja kasvua mittaavia tunnuslukuja on neljä kappaletta, mutta loput muuttujat ovat kannattavuuden tunnuslukuja. Muuttujien keskinäisen paremmuuden vertailussa omavaraisuusasteen ohella kannattavuuden tunnusluvut osoittautuivat selkeästi olevan parhaiten konkurssia ennakoivia. Kannattavuuden tunnusluvut selittävät kuitenkin osin samaa asiaa, minkä vuoksi muuttujat voivat olla keskenään korreloituneita. Keskinäinen korrelaatio voi aiheuttaa multikollineaarisuutta, jossa muuttujien selityskyky vääristyy ja vaimenee. Tästä syystä muuttujien keskinäistä riippuvuutta kartoitettiin Pearsonin korrelaatiokertoimien avulla. Korrelaatioiden laskennassa hyödynnettiin ainoastaan 50/50 % aineistoa.

Vuotta ennen konkurssihetkeä muuttujat asettuivat korrelaatioiden perusteella lähes samaan hyvyysjärjestykseen kuin boruta-algoritmilla (taulukko 8). Korrelaatioista voitiin todeta, että lähes kaikki selittävät muuttujat ovat jossain määrin korreloituneita keskenään. Minimoimalla muuttujien välisiä korrelaatioita päädytään vuotta ennen konkurssia

olevaan malliin valitsemaan omavaraisuusasteen lisäksi rahoitustulos-% ja sijoitetun pääoman tuotto-%. Kokonaispääoman tuottoa kuvaava ROA-% jätetään mallin ulkopuolelle, koska se on vahvasti korreloitunut omavaraisuusasteen kanssa. Maksuvalmiutta tai kasvua kuvaavia tunnuslukuja ei lopullisiin malleihin otettu mukaan, koska niiden selityskyvyt havaittiin niin heikoiksi boruta-algoritmin- ja Pearsonin korrelaatioiden perusteella.

Taulukko 8. Pearsonin korrelaatiokertoimet muuttujien väliselle korrelaatiolle vuotta ennen konkurssihetkeä. Lopulliseen malliin valikoidut muuttujat merkitty tähdellä (*).

2016	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Luokka	-0,63	-0,53	-0,42	-0,42	-0,34	-0,29	-0,33
1. Omavaraisuusaste *	1,00	0,64	0,49	0,50	0,47	0,38	0,45
2. ROA%	0,64	1,00	0,62	0,61	0,74	0,41	0,21
3. Rahoitustulos % *	0,49	0,62	1,00	0,92	0,36	0,84	0,15
4. Ebitda %	0,50	0,61	0,92	1,00	0,34	0,69	0,17
5. ROCE % *	0,47	0,74	0,36	0,34	1,00	0,20	0,16
6. Tilikauden tulos %	0,38	0,41	0,84	0,69	0,20	1,00	0,08
7. Current ratio	0,45	0,21	0,15	0,17	0,16	0,08	1,00

Kahta vuotta ennen konkurssia selittävien muuttujien joukkoon valittiin omavaraisuusaste, sillä sen selityskyky oli Boruta-algoritmin ja Pearsonin korrelaatioiden perusteella paras (taulukko 9). Toiseksi muuttujaksi valitaan tilikauden tulos-% ja kolmanneksi muuttujaksi kokonaispääoman tuotto-%, sillä ne olivat Boruta-algoritmin perusteella hyvässä riippuvuudessa konkurssin kanssa. Korrelaatioiden perusteella on havaittavissa kuitenkin melko korkeita riippuvuuksia myös selittävien muuttujien välillä. Lopulliset muuttujat pyrittiin valitsemaan mahdollisimman hyvin tämä huomioiden.

Taulukko 9. Pearsonin korrelaatiokertoimet muuttujien väliselle korrelaatiolle kaksi vuotta ennen konkurssihetkeä. Lopulliseen malliin valikoidut muuttujat merkitty tähdellä (*).

2015	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Luokka	-0,49	-0,36	-0,32	-0,32	-0,31	-0,24	-0,18
1. Omavaraisuusaste *	1,00	0,46	0,38	0,34	0,43	0,25	0,37
2. Tilikauden tulos %*	0,46	1,00	0,90	0,77	0,73	0,63	0,14
3. Rahoitustulos %	0,38	0,90	1,00	0,82	0,63	0,55	0,13
4. Ebitda %	0,34	0,77	0,82	1,00	0,65	0,59	0,11
5. ROA % *	0,43	0,73	0,63	0,65	1,00	0,69	0,08
6. ROCE %	0,25	0,63	0,55	0,59	0,69	1,00	0,05
7. Current ratio	0,37	0,14	0,13	0,11	0,08	0,05	1,00

Muuttujien selityskyky huononee entisestään siirryttäessä kolmea vuotta konkurssia edeltävään ajankohtaan (taulukko 10). Tälle ajankohdalle valittiin malliin mukaan ensimmäiseksi muuttujaksi rahoitustulos-%, sillä se oli Borutan perusteella keskimäärin selityskyvyltään paras muuttuja. Toisena muuttujana malliin otetaan mukaan omavaraisuusaste ja kolmanneksi muuttujaksi tilikauden tulos-%. Selityskyvyltään Ebitda-% vaikutti olevan korrelaatioiden perusteella tilikauden tulosta parempi muuttuja, mutta sen vahva riippuvuus rahoitustuloksesta oli syy siihen, että se jätettiin mallin ulkopuolelle.

Taulukko 10. Pearsonin korrelaatiokertoimet muuttujien väliselle korrelaatiolle kolme vuotta ennen konkurssihetkeä. Lopulliseen malliin valikoidut muuttujat merkitty tähdellä (*).

2014	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Luokka	-0,37	-0,32	-0,28	-0,27	-0,21	-0,16	-0,14
1. Omavaraisuusaste *	1,00	0,31	0,27	0,40	0,29	0,41	0,25
2. Rahoitustulos % *	0,31	1,00	0,85	0,68	0,58	0,04	0,44
3. Ebitda %	0,27	0,85	1,00	0,61	0,56	0,04	0,50
4. Tilikauden tulos % *	0,40	0,68	0,61	1,00	0,59	0,10	0,41
5. ROA %	0,29	0,58	0,56	0,59	1,00	0,07	0,67
6. Current ratio	0,41	0,04	0,04	0,10	0,07	1,00	0,05
7. ROCE %	0,25	0,44	0,50	0,41	0,67	0,05	1,00

5.3. K:n lähimmän naapurin menetelmä

Mallintamista edeltävästi aineistojen (50/50 % ja 10/90 %) muuttujat normalisoitiin, jotta muuttujat saatiin yhteismitallisiksi k:n lähimmän naapurin menetelmää varten. Yleisimmät normalisointimenetelmät ovat min-max -normalisointi (kaava 11) ja z-score-standardointi (kaava 12). Min-max -normalisoinnissa jokainen selittävä muuttuja suhteutetaan tunnuslukua vastaavasti populaation minimi- ja maksimiarvoihin, jolloin muuttujat saadaan vaihtelun osalta välille 0-1. Z-score -standardoinnissa muuttujat suhteutetaan käyttämällä keskiarvoa ja keskihajontaa, jolloin standardoidut muuttujat vaihtelevat ilman kiinteää ylä- tai alarajaa nollan molemmin puolin normaalijakautuneesti (Lantz 2013:73).

$$X_{uusi} = \frac{X - \min(X)}{\max(X) - \min(X}$$

(11)

missä:

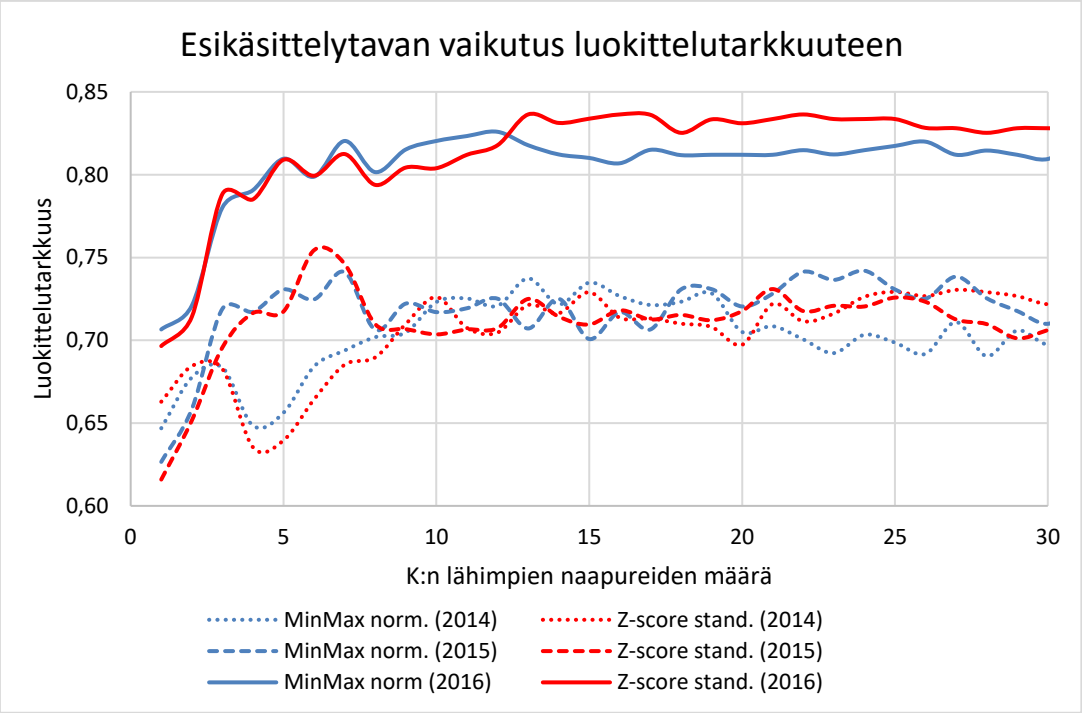
X_{uusi} = Normalisoitu X-muuttuja
X= Alkuperäinen X:n arvo
min (X)= pienin X-havainnon arvo aineistossa
max (X)= suurin X-havainnon arvo aineistossa

$$X_{uusi} = \frac{X - \mu}{\sigma} \quad (12)$$

missä:

X_{uusi} = Standardoitu X-muuttuja
X= Alkuperäinen X:n arvo.
μ= X-havaintojen keskiarvo aineistossa
σ= X-havaintojen keskihajonta aineistossa

Kummankin vaihtoehtoisen normalisointimenetelmän vaikutusta luokittelutarkkuuteen testattiin lukuisilla erisuurilla k:n arvoilla ja 50/50 % aineistolla (kuvio 8). Normalisointimenetelmien välillä ei havaittu olevan suurta vaikutusta luokittelutarkkuuden suhteen. Alhaisilla k:n arvoilla saatiin keskimäärin hieman parempia luokittelutarkkuuksia min-max -normalisoinnilla. Tästä syystä tutkimuksen muut laskennat päätettiin suorittaa käyttäen sitä aineiston esikäsittelytapana.



Kuvio 8. Normalisointimenetelmien vaikutus luokittelutarkkuuteen. X-akselilla kuvattu naapureiden lukumäärä. Kuvaajassa on eri viivatyyleillä kuvattu eri ajankohtien aineistoista lasketut luokittelutarkkuudet lähimpien naapureiden suhteessa. Konkurssiyritysten suhde terveisiin yrityksiin 50/50 %.

Aineistot jaettiin mallin sovittamiseen (70 %) ja mallin testaamiseen (30 %) käytettävään ositteeseen. K:n lähimmän naapurin algoritmin sovitus tehtiin sovitusaineistoon k-ositetun (k=10) ristiinvalidoinnin avulla, jolloin aineisto voitiin käyttää sovitukseen tehokkaasti. Sovittamisen jälkeen mallilla ennustettiin konkurssi/terve yritys -luokitus 30 % testiaineistoon. Testiaineistoon tehdystä luokituksesta tallennettiin lopulliset tulokset vertailua varten.

5.3.1. Mallin parametrien valinta

K:n lähimmän naapurin menetelmässä on algoritmin optimaalisen toiminnan kannalta kolme keskeistä luokittelun tarkkuuteen vaikuttavaa parametria. Näitä ovat lähimpien naapureiden määrä, etäisyysfunktion muoto ja naapureiden painotus. Jokaisen osatekijän osalta testattiin ennen lopullisten tulosten laskentaa erilaisia parametrivaihtoehtoja, jotta

optimaalinen k:n lähimmän naapurin malli voitaisiin löytää. Algoritmin parametrien optimointi suoritettiin käyttäen 70 % mallinnusaineistoa, johon parametritestauksen tulokset laskettiin k-ositetun ristiinvalidoinnin avulla. Lähimpien naapureiden vaihteluvälinä päätettiin optimoinnissa käyttää 1-15 kpl naapureita.

Etäisyysfunktiona käytettiin Minkowskin etäisyyttä, joka on yleistetty muoto etäisyysfunktiosta (kaava 13; kts. Hechenbichler ja Schliep 2004). Minkowskin etäisyysfunktiossa *m*-parametrin suuruudella voidaan muokata etäisyysfunktion astetta ja siten sen toimintaa. Jos *m*-parametrin arvo on 1, etäisyysfunktio toimii Manhattanin etäisyytenä ja jos taas m-parametrin arvona käytetään lukuarvoa 2, lasketaan etäisyys euklidiseen etäisyyteen perustuvasti. Erilaisilla etäisyysfunktion asteilla voidaan vaikuttaa etäisyyden laskentaan naapureiden välillä ja siten optimoida lähimpien naapureiden valintaa. Parametrivalidoinnissa etäisyysfunktion vaikutusta luokittelutarkkuuteen testattiin 5 eri etäisyysfunktion asteella (0.5, 1, 1.5, 2, 3).

$$d = (|q_1 - p_1|^m + |q_2 - p_2|^m + |q_n - p_n|^m)^{\frac{1}{m}}$$

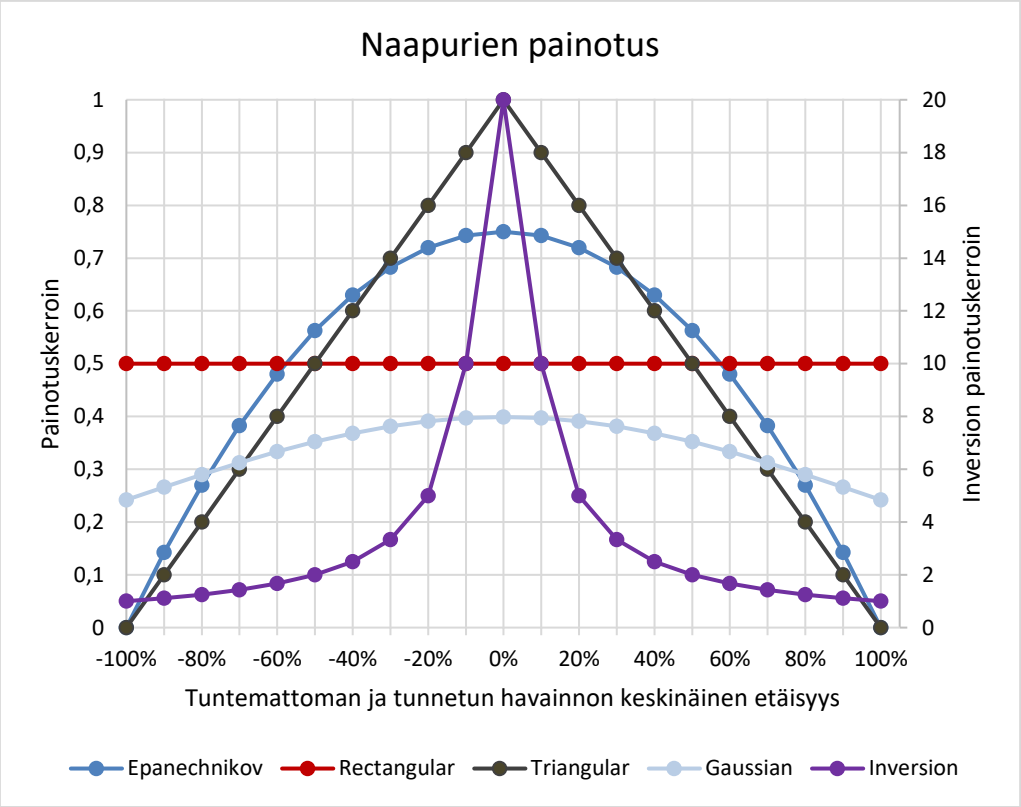
(13)

missä

d = Etäisyys havaintojen q ja p välillä.
$q_{1..n}$ = Tuntemattoman havainnon (q) muuttujassa (1…n) oleva arvo.
$p_{1..n}$ = Tunnetun havainnon (p) muuttujassa (1…n) oleva arvo.
m = Etäisyysfunktion ulottuvuus (0…n).

Naapureiden painotus on myös merkittävä tekijä luokittelussa. Painotus toimii siten, että tuntemattoman havainnon luokittelussa tunnetuille naapureille annetaan painotus sen mukaan, miten lähellä ne ovat piirreavaruudessa tuntematonta havaintoa. Mitä lähempänä tunnettu havainto on, sitä suuremman painoarvon sitä vastaava vastemuuttuja painotuksen kautta saa tuntemattoman havainnon luokittelussa. Erilaisia painotus-, eli kernelfunktioita on lukuisia erilaisia. Tässä tutkimuksessa soveltuvin painotusfunktio etsitään viiden vaihtoehtoisen painotusfunktion joukosta (kuvio 9). Validoinnissa testattiin painotusfunktion vaikutusta luokittelutarkkuuteen eri k:n lähimpien naapureiden määrällä.

Kuviossa 9 y-akselilla kuvataan painokertoimen suuruutta suhteessa x-akselilla esitettyyn tunnetun ja tuntemattoman havainnon etäisyyteen toisistaan. Kun etäisyys on alhaisimmillaan (0 %), antavat kernel-funktiot tätä etäisyyttä vastaavalle muuttujalle suurimman painotuksen (1). Painotuksista *rectangular* antaa kaikille yhtäläisen painotuksen (0,5) etäisyydestä riippumatta, joten kyseessä on painottamaton kernel. Voimakkain painotus sitä vastoin saadaan *inversion*-kernelillä, jossa painokerroin lasketaan etäisyyden käänteisluvun kautta. Painokerroin lähestyy ääretöntä silloin, kun tunnetun havainnon piirteet lähestyvät identtistä suhteessa tuntemattoman havainnon piirteisiin. Muut painokertoimet asettuvat voimakkuudeltaan näiden kahden välimaastoon. *Triangular*- painotus vastaa tilannetta, jossa painotus kasvaa lineaarisesti etäisyyden suhteen. *Gaussian*- painokerroin on sitä vastoin johdettu normaalijakaumasta. *Epanechnikov*-kernelin on sitä vastoin mainittu olevan paraboloidisen muotonsa vuoksi teoreettisesti optimaalisin vaihtoehto (Epanechnikov 1969). Optimaalisten parametrien valintaan liittyvä laskenta toteutettiin R-ohjelmalla (R Development Core Team 2018) käyttämällä *kknn*-, ja *caret*-kirjastoja (kts. Hechenbichler ja Schliep 2004; Kuhn 2008).



Kuvio 9. Naapureiden painottamisessa käytetyt kernel-funktiot. X-akselilla kuvataan suhteellista etäisyyttä tunnetun ja tuntemattoman havainnon piirteiden välillä. Y-akseleilla kuvataan painokertoimien suuruutta suhteessa etäisyyteen.

5.3.2. Tulosten laskenta ja mallin hyvyyden arviointi

Optimaalisten mallin parametrien valinnan jälkeen laskettiin k:n lähimmän naapurin menetelmällä luokittelutulokset käyttäen 30 % osuutta koko aineistosta. Luokittelutulokset laskettiin erikseen 50/50 % aineistoon ja 10/90 % aineistoon käyttämällä lähimpien naapureiden määränä 3, 5, 9 ja 15 naapuria. Luokitus laskettiin kahteen otteeseen, ensin ilman muuttujien keskinäiseen hyvyyteen perustuvaa ennakkotietoa, ja toisen kerran painottamalla muuttujia ennakkotiedolla. Muuttujien keskinäisessä painotuksessa hyödynnettiin ennakkotietona boruta-algoritmilla eri tunnusluvuille laskettuja z-arvoja. Malleihin valituille kolmelle muuttujalle laskettiin painotus vertaamalla tunnusluvun saamaa z-arvoa suhteessa parhaimmaksi validoidun muuttujan saamaan z-arvoon (taulukko 11). Konkurssi/terve yritys -luokittelun ennustamisessa painotusta käytettiin osana

etäisyysfunktiota kaavan 14 osoittamalla tavalla. Muuttujien keskinäisessä painotuksessa käytetyt painokertoimet ovat esitetty taulukossa 11.

$$d = ((w_1|q_1 - p_1|)^m + (w_2|q_2 - p_2|)^m + (w_3|q_3 - p_3|)^m)^{\frac{1}{m}} \quad (14)$$

missä

d = Etäisyys havaintojen q ja p välillä.
$q_{1..n}$ = Tuntemattoman havainnon (q) muuttujassa (1…n) oleva arvo.
$p_{1..n}$ = Tunnetun havainnon (p) muuttujassa (1…n) oleva arvo.
m = Etäisyysfunktion ulottuvuus (0…n).
$w_{1...n}$ = Muuttujaa (1…n) vastaava painokerroin

Taulukko 11. Selittävien muuttujien painokertoimet. Painokertoimet laskettu Boruta-algoritmilla muuttujille lasketuista tärkeysarvoista.

Selittävien muuttujien painotukset	2016	2015	2014
Omavaraisuusaste %	1	1	1,08
ROCE %	1,90	-	-
ROA %	-	1,30	-
Rahoitustulos %	1,53	-	1
Tilikauden tulos %	-	1,18	1,17

Tulosten hyvyyttä arvioidaan suhteessa siihen, onnistuuko luokittelu konkurssiyrityksen ja terveen yrityksen välillä. **Virhetyyppi 1** syntyy, jos malli luokittelee konkurssiyrityksen terveeksi yritykseksi (kaava 15; Laitinen ja Laitinen 2014:158). **Virhetyyppi 2** syntyy silloin, jos malli luokittelee terveen yrityksen konkurssiyritykseksi (kaava 16). Virhetyyppien vertailussa virhetyyppi 1 nähdään merkittävästi enemmän kustannuksia kerryttävänä kuin virhetyyppi 2. Kustannuksien on havaittu olevan jopa useita kymmeniä kertoja suuremmat virhetyyppi 1:ssä verrattuna virhetyyppi 2:een (Altman 1980). Kustannuksia aiheutuu erityisesti luotonantajan näkökulmasta silloin, jos konkurssin ennakointimalli ennustaa yrityksen olevan terve, mutta yritys meneekin konkurssiin, jolloin luotonantaja menettää lainaamansa pääoman ja sen korot (Laitinen ja Laitinen 2014:158).

Mallin **luokittelutarkkuus** (accuracy) kuvaa oikeinluokituksen suhteellista määrää ja se lasketaan oikein luokiteltujen konkurssiyritysten ja terveiden yritysten keskiarvon suhteena yritysten kokonaismäärään (kaava 17).

$$Virhetyyppi\ 1\ (\%) = \left(1 - \frac{A}{N}\right) * 100 \quad (15)$$

$$Virhetyyppi\ 2\ (\%) = \left(1 - \frac{B}{N}\right) * 100 \quad (16)$$

$$Luokittelutarkkuus - \%\ (ACC) = \frac{\left(\frac{A}{N} + \frac{B}{N}\right)}{2} * 100 \quad (17)$$

Missä:

A = Oikein luokitellut konkurssiyritykset (kpl)
B = Oikein luokitellut terveet yritykset (kpl)
N = Aineistossa olevien yritysten lukumäärä (kpl)

Uusimmissa tutkimuksissa mallien binäärisen luokitteluongelman luokittelutarkkuutta mitataan myös **ROC-käyrällä** (*Receiver Operating Characteristic curve*) ja siitä lasketulla **AUC -luvulla** (*Area Under the Curve*) (kts. Barboza ym. 2017). ROC-käyrä on graafinen kuvaaja, jossa esitetään yksinkertaistetulla tavalla binäärisen luokittelumallin luokittelun onnistuminen kaikilla mahdollisilla luokittelun kriittisillä pisteillä (kuvio 10). ROC-kuvaajassa y-akselilla **Sensivity-muuttuja** (kaava 18) ja x-akselilla on **Specificity-muuttuja** (kaava 19). **Sensitivity** kuvaa positiivista oikeinluokitusta, eli osuutta oikein luokitelluista konkurssiyrityksistä (Engelmann, Hayden ja Tasche 2003:6). **Specificity** taas kuvaa positiivista väärinluokitusta, eli yrityksiä, joiden ennakoitiin virheellisesti menevän konkurssiin. Mitä paremmin ennakointimalli kykenee luokittelemaan konkurssiyritykset ja terveet yritykset, sitä lähemmäksi ROC-käyrä siirtyy vasenta yläkulmaa ja koordinaatiston arvoa (0,1). Kuvaajan läpi kulmasta kulmaan kulkeva **satunnaisen arvauksen käyrä** kuvaa tilannetta, jossa luokittelumalli on pystynyt luokittelemaan vain 50

% yrityksistä oikeaan luokkaan. Tämä luokittelutarkkuus vastaa satunnaisen arvauksen, eli esimerkiksi kolikon heiton tarkkuutta.

AUC-luku lasketaan ROC-käyrän alle jäävän alueen pinta-alasta (Engelmann ym. 2003:6). Mitä paremmin luokittelumalli toimii, sitä lähemmäksi ROC-käyrä siirtyy kuvaajan vasenta yläkulmaa. Tällöin sinisen alueen pinta-ala kuviossa 10 kasvaa suhteessa kuvaajan kokonaispinta-alaan ja suhde kokonaispinta-alaan lähestyy 100 prosenttia. AUC-luvun arvolla voidaan vertailla kahden eri luokittelumallin hyvyyttä toisiinsa, mikä helpottaa paremman mallin löytämistä tilanteessa, jossa ROC-käyrät kulkevat malleilla lähellä toisiaan. AUC -arvon tulkinnasta ovat kirjoittaneet muun muassa Hosmer ja Lemeshow (2000:160-165). Jos AUC-luku on 0,5 niin luokittelumalli on yhtä hyvä kuin satunnainen arvaus. Luvun ollessa välillä 0,7 - 0,8 mallin luokittelukyky on hyväksyttävällä tasolla. Erinomaisesti luokittelevan mallin AUC-luku on välillä 0,8-0,9 ja luvun noustessa yli 0,9, mallin luokittelukyky on jo lähes täydellinen.

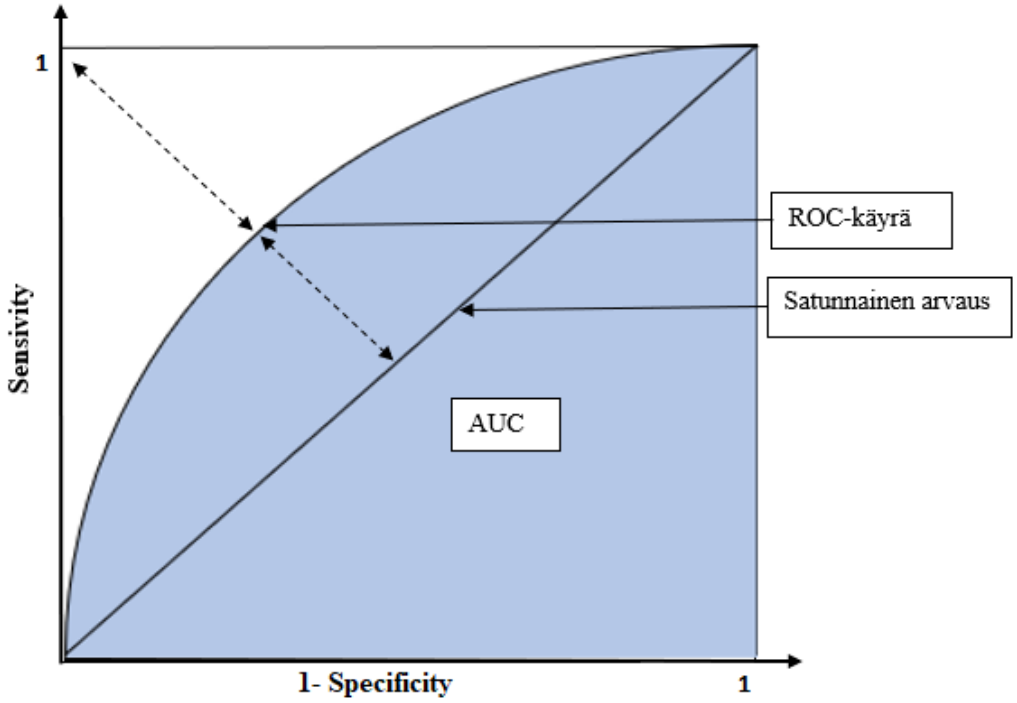
$$Sensivity = 1 - Virhetyyppi\ 1\ (\%) \quad (18)$$

$$1 - Specificity = \ 1 - Virhetyyppi\ 2\ (\%) \quad (19)$$

missä:

Virhetyyppi 1 = Konkurssiyritys luokiteltiin terveeksi yritykseksi

Virhetyyppi 2 = Terve yritys luokiteltiin konkurssiyritykseksi



Kuvio 10. ROC-käyrän ja AUC-arvon havainnollistus.

K:n lähimmän naapurin menetelmällä saatujen tulosten hyvyyttä vertailtiin eri k:n arvoilla, sekä painotuksilla, saatuihin tuloksiin virhetyyppi 1:en, virhetyyppi 2:en, luokittelutarkkuuden ja AUC-arvojen avulla. Luokittelutulosten eroja tarkasteltiin myös tilastotieteellisin menetelmin laskemalla k:n lähimmän naapurin menetelmällä saaduista luokittelutarkkuuksista z-testisuureen arvo ja sitä vastaava p-arvo. Testaamiseen käytettiin kaksisuuntaista z-testiä luokittelutarkkuuksien erotukselle (kaava 20).

$$z = \frac{p_1 - p_2}{\sqrt{p(100-p)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

missä:

p_1, p_2 = eri malleilla saadut luokittelutarkkuudet
n_1, n_2 = luokittelutarkkuuksia vastaavat yritysten määrät.
p = Painotettu keskiarvo p_1:stä ja p_2:sta

5.4. Laskenta logistisella regressiolla

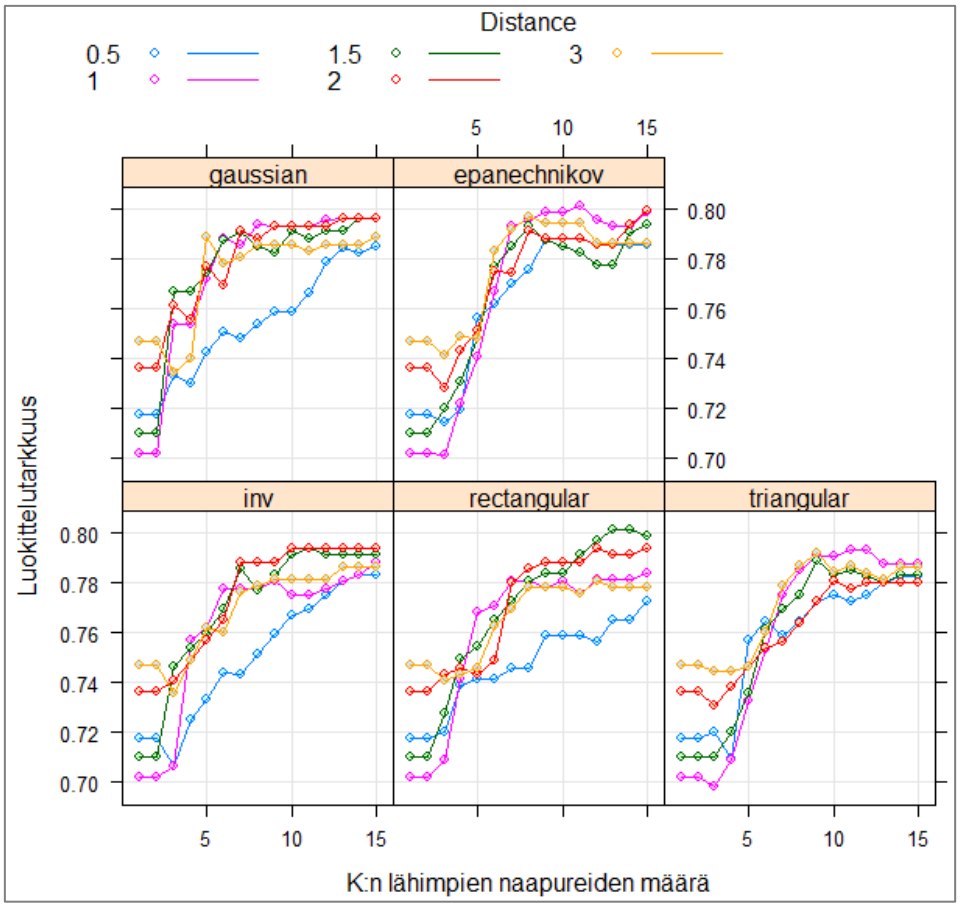
Logistista regressiota käytettiin tutkimuksessa referenssimenetelmänä, johon k:n lähimmän naapurin menetelmän luokittelutuloksia verrattiin. Logistisessa regressiossa käytettiin samoja selittäviä muuttujia kuin k:n lähimmän naapurin menetelmässä. Regressiomalli sovitettiin ensin mallin sovittamiseen tarkoitettuun aineistoon (70 %). Tämän jälkeen mallilla ennustettiin luokittelu käyttäen testiaineistoa (30 %).

Testiaineistoon ennustetuista luokitustuloksista laskettiin 50/50 %- ja 10/90 % aineiston osalta samat mallin luokittelutarkkuutta kuvaavat tunnusluvut kuin k:n lähimmän naapurin menetelmän osalta. Luokittelun hyvyyden arviointia varten tallennettiin virhetyyppi 1:stä ja 2:sta vastaavat luokittelutulokset, luokittelutarkkuus ja luokittelulle laskettu AUC-arvo. Mallissa käytettiin kriittisenä pisteenä, eli ehdollisena todennäköisyytenä luokkien välillä, arvoa 50 %. Käytännössä tämä tarkoittaa, että jos malli antaa konkurssiluokitukselle estimaatiksi yli 50 % todennäköisyyden, luokitellaan kyseinen yritys konkurssiyritykseksi.

6. TUTKIMUSTULOKSET

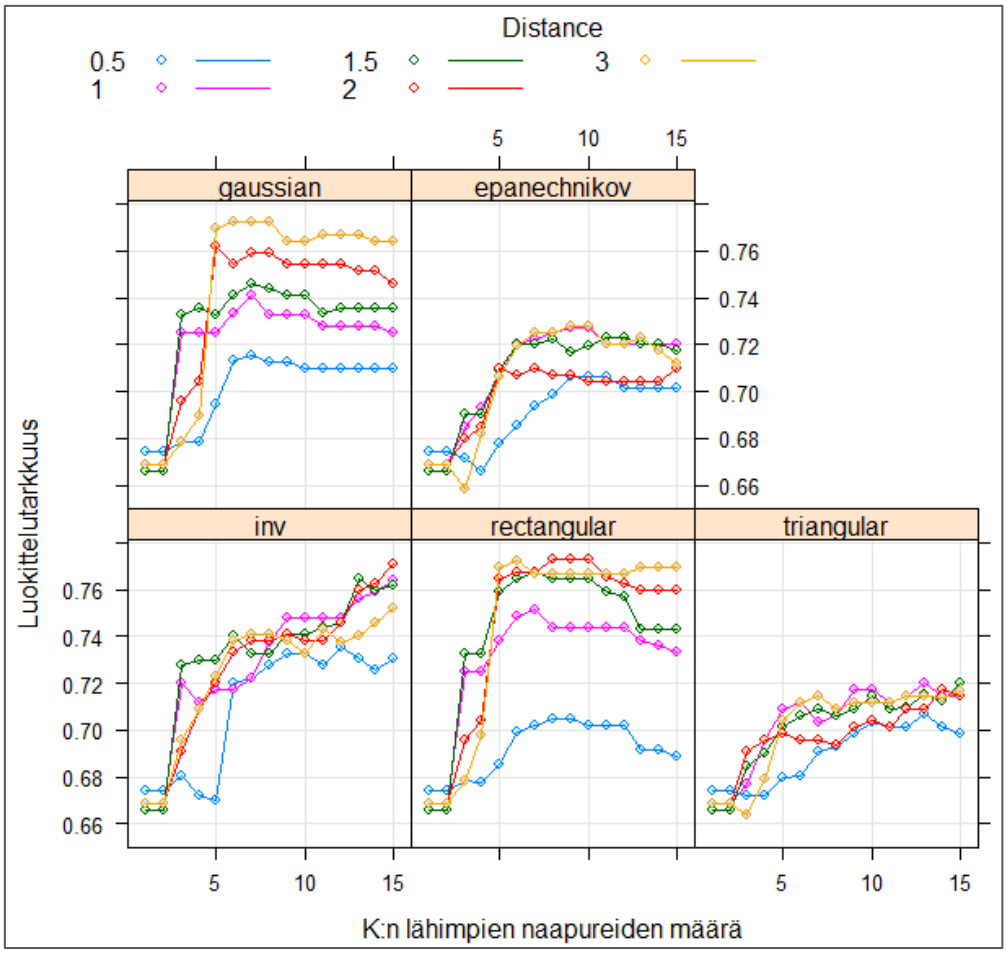
6.1. Parametrivalinnan tulokset

Parametrivalinnan tulokset vuodelle 2016 on esitetty kuviossa 11. Kuviossa y-akselilla on luokittelutarkkuus ja x-akselilla on lähimpien naapureiden lukumäärä. Eri väriset viivat kuvaavat etäisyysfunktion eri asteita ja eri painotukset on esitetty erillisinä osakuvaajina. Tulosten perusteella ei ole havaittavissa suurta eroa etäisyysfunktioiden ja painotusten välillä. Erot kernel-funktioiden ja etäisyysfunktioiden asteiden välillä ovat hyvin marginaaliset. Keskimääräisesti ehkä parhaimpana painotuksena on *gaussian*-kernel, jossa luokittelutarkkuus nousi melko nopeasti jo pienillä k:n arvoilla noin 80 % tuntumaan. Etäisyysfunktioiden asteiden keskinäinen hyvyys on sitä vastoin paljon hankalammin tulkittavissa, sillä asteilla 1-3 saadut tulokset ovat lähes identtisiä. Tästä syystä nähdään perusteltuna valita lopullisiksi parametreiksi k:n lähimmän naapurin menetelmään vuotta ennen konkurssia kernel-funktion osalta *gaussian*-kernel ja etäisyysfunktioksi euklidean etäisyys.



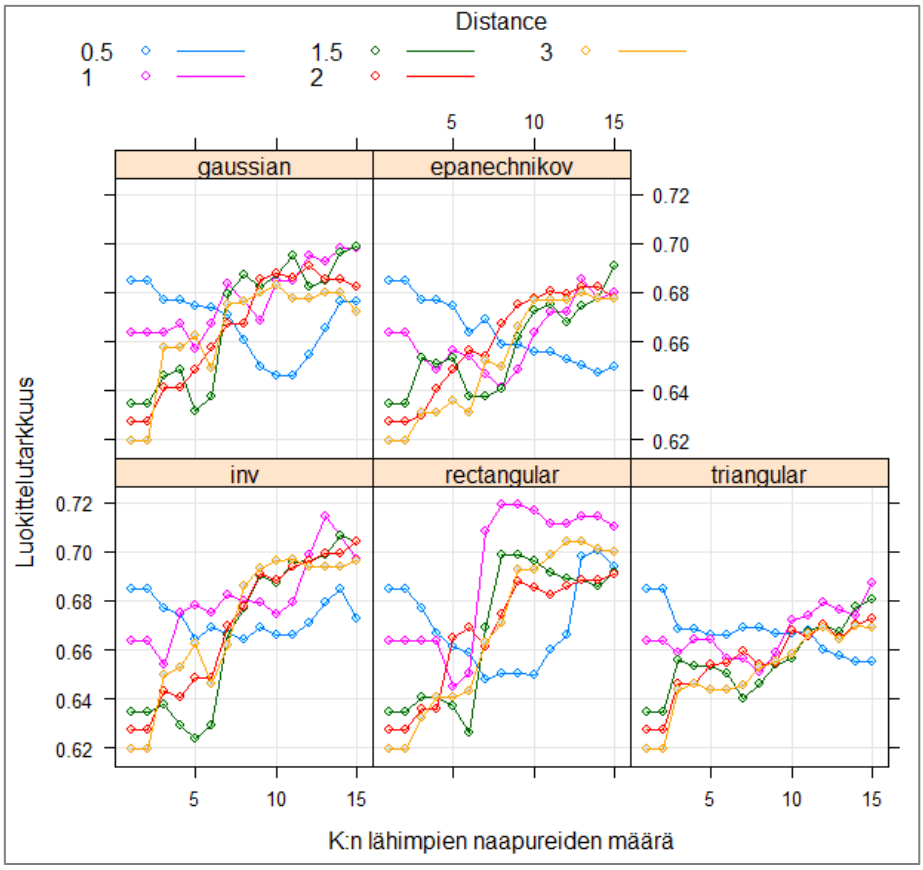
Kuvio 11. K:n lähimmän naapurin parametrivalinnan tulokset 2016

Kahta vuotta ennen konkurssihetkeä koskevan parametrivalinnan tulokset ovat nähtävissä kuviossa 12. Validoinnin perusteella silmämääräisesti arvioituna parhaat luokittelutulokset saatiin ilman painotusta (*rectangular*). Erot painotusten välillä olivat kuitenkin hyvin marginaaliset. Keskimäärin painottamattoman menetelmän luokittelutarkkuudet nousevat jo alhaisilla k:n arvoilla jyrkästi ja pysyttelevät lähellä maksimiarvoa (77 %) aina 15 naapurin tasolle saakka. *Gaussian*-kerneliä käyttämällä saadaan myös hyvät tulokset, mutta siinä etäisyysfunktion asteiden 1-3 välillä on havaittavissa luokittelutarkkuuden osalta enemmän hajontaa. Tästä syystä vuodelle 2015 valittiin lopullisten tulosten laskentaan etäisyysfunktion asteeksi 2 (euklidinen etäisyys) ja kernel- funktioksi painottamaton *rectangular*.



Kuvio 12. K:n lähimmän naapurin parametrivalinnan tulokset 2015.

Kolme vuotta ennen konkurssia aineistosta saaduista luokittelutuloksista voidaan selvästi nähdä, että eri parametrikombinaatioiden luokittelutarkkuuksissa on havaittavissa lisääntyvissä määrin kohinaa (kuvio 13). Kernel-funktioista korkeimmat yksittäiset luokittelutarkkuudet saadaan painottamattomalla (*rectangular*) kernelillä. Sitä käyttämällä luokittelutarkkuuksissa on kuitenkin suhteellisen paljon hajontaa eri lähimpien naapureiden arvoilla, joten sen vuoksi painottamatonta kerneliä ei valittu lopulliseksi painotukseksi. Lupaavia keskimääräisiä luokittelutarkkuuksia suhteellisen pienellä hajonnalla saadaan tulosten perusteella käytettäessä käänteislukuun perustuvaa (*inv*) painotusta. Etäisyysfunktiona toimi tulosten perusteella parhaiten manhattanin etäisyys. Tätä parametriyhdistelmää päätettiin käyttää lopullisten tulosten laskennassa.



Kuvio 13. K:n lähimmän naapurin parametrivalinnan tulokset 2014.

6.2. Tulokset testiaineiston luokittelusta

Taulukoissa 12, 13 ja 14 on esitetty lopulliset tulokset k:n lähimmän naapurin menetelmällä ja logistisella regressiolla lasketusta konkurssin ennakoinnista testiaineistoon. Tulostaulukot on koostettu yhteen vuosittain ja taulukoiden sisään on erotettu erikseen 50/50 % ja 10/90 % otossuhteita vastaavien aineistojen luokittelutulokset. Taulukoissa esitetyt tulokset sisältävät virhetyyppi 1 ja 2 suhteelliset osuudet, sekä luokittelutarkkuuden (ACC), AUC- arvon ja z-testin kautta lasketun p-arvon. Tulokset on järjestetty taulukoittain ja otossuhteittain luokittelutarkkuuden perusteella parhaimmasta huonoimpaan. Taulukkojen ”malli”-sarakkeessa on esitetty ensin mallin nimi, joka on joko k:n lähimmän naapurin menetelmä (knn) tai logistinen regressio (logit). Knn-menetelmän jälkeen mainittu lukuarvo viittaa kyseisessä sovituksessa käytössä olleeseen k:n lähimpien naapureiden määrään. Lisäksi suluissa olevalla merkinnällä ”w” viitataan boruta-algoritmilla

muuttujille laskettujen tärkeysarvojen kautta saatuihin muuttujien painotusten käyttämiseen.

Taulukossa 12 on konkurssin ennakointitulokset vuotta ennen konkurssihetkeä. Otossuhteella 50/50 % logistinen regressio toimii kaikilla tarkkuusmittareilla parhaiten. Luokittelutarkkuudeksi saatiin 87 %, joka on hyvin korkea. Tarkasteltaessa virhetyyppien määriä voidaan havaita osuuksien olevan alhaisia ja virhetyyppien määrien olevan toisiinsa nähden tasaväkisiä. Knn-algoritmilla 50/50 % aineiston tapauksessa parhaat luokittelutulokset saatiin käytettäessä naapureiden määränä 9 tai 15 naapuria. Luokittelutarkkuus jäi tällöin kuitenkin hieman alemmaksi (82,6 %), kuin logistisella regressiolla. Virhetyyppi 1:n määrä oli noin kaksinkertainen verrattuna virhetyyppi 2:n osuuteen. AUC-arvon perusteella logistisen regression ja knn-algoritmin suorituskykyä voidaan pitää lähes täydellisenä, kun AUC-arvo ylittää 90 % (Hosmer ja Lemeshow 2000:160-165).

Otossuhteella 10/90 % ovat luokittelutuloksissa havaittavissa selkeitä eroja (taulukko 12). Asetettaessa menetelmät luokittelutarkkuuden perusteella paremmuusjärjestykseen nousee k:n lähimmän naapurin menetelmä logistisen regression ohi. Erot menetelmien välillä ovat kuitenkin melko marginaaliset ja logistisella regressiolla saatu luokittelutarkkuus jää vain 0,8 % knn-algoritmilla saadusta parhaasta luokittelutarkkuudesta. Otossuhteella 10/90 % nousee virhetyyppi 1 määrä suhteessa virhetyyppi 2 määrään merkittävästi. Virhetyyppi 1 tarkoittaa osuutta konkurssiyrityksistä, jotka luokiteltiin virheellisesti terveiksi yrityksiksi. Tämä todennäköisesti johtuu suoraan siitä, että malli tyypillisesti ennustaa parhaiten populaation keskimääräisiä arvoja ja huonommin ääriarvoja. Logistisella regressiolla ja k:n lähimmän naapurin menetelmällä tulokset eri otossuhteilla osoittavat aukottomasti, että optimaalinen luokittelutilanne on silloin, kun populaatio sisältää konkurssiyrityksiä ja terveitä yrityksiä suhteessa 50/50%.

Virhetyyppi 1:stä pidetään yleisesti enemmän kustannuksia aiheuttavana luokitteluvirheenä, kuin virhetyyppi 2 lainoittajan näkökulmasta (Laitinen ja Laitinen 2014:158). Tulosten perusteella k:n lähimmän naapurin menetelmän muuttujien painotus (w) vähentää virhetyyppi 1 suhteellista määrää merkittävästi (taulukko 12). Virhetyyppi 1 määrä putoaa nollan tuntumaan aineistolla 50/50 % ja 10/90 % aineistollakin 8-12 % tasolle asti. Luokittelutarkkuus oli painotetulla menetelmällä kuitenkin painottamatonta alhaisempi.

Jos huomioidaan virhetyyppien väliset kustannukset, voi painotettu knn-algoritmi olla tosiasiassa optimaalisempi luokittelija

Taulukon 12 oikeanpuoleisimmassa sarakkeessa on esitetty jokaisesta menetelmästä tilastollista merkitsevyyttä kuvaava p-arvo. P-arvo laskettiin z-testin kautta vertaamalla k:n lähimmän naapurin luokittelutarkkuutta suhteessa logistisella regressiolla saatuun luokittelutarkkuuteen. Jotta menetelmien luokittelutarkkuuksien ero olisi tilastollisesti merkitsevä, tulee p-arvon olla alhaisempi kuin 0,05, joka on tilastollinen merkitsevyysraja 95 % luotettavuustasolla. Otossuhteella 50/50 % logistinen regressio luokittelee tilastollisesti merkitsevällä tasolla tarkemmin verrattuna painotettuun k:n lähimmän naapurin menetelmän tuloksiin. Painottamattomassa knn-menetelmässä erot eivät ole tilastollisesti merkitseviä. Sama tulos saatiin otossuhteella 10/90 %.

Taulukko 12. Luokittelutulokset vuotta ennen konkurssia. Tulokset laskettu testiaineistoon.

Otossuhde	Malli	Virhetyyppi 1	Virhetyyppi 2	ACC	AUC	P-arvo
50/50	Logit	12,0 %	14,3 %	87,0 %	94,9 %	-
	Knn9	24,0 %	9,5 %	82,6 %	90,5 %	0,281
	Knn15	24,0 %	9,5 %	82,6 %	90,5 %	0,281
	Knn5	20,0 %	23,8 %	78,3 %	90,5 %	0,136
	Knn3	16,0 %	33,3 %	76,1 %	85,1 %	0,090
	Knn5 (w)	0,0 %	76,2 %	65,2 %	61,2 %	0,007
	Knn9 (w)	0,0 %	76,2 %	65,2 %	68,4 %	0,007
	Knn15 (w)	0,0 %	76,2 %	65,2 %	68,4 %	0,007
	Knn3 (w)	0,0 %	81,0 %	63,0 %	63,0 %	0,004
10/90	Knn5	68,0 %	1,3 %	92,2 %	82,2 %	0,626
	Knn9	72,0 %	1,3 %	91,8 %	86,5 %	0,563
	Knn15	72,0 %	1,3 %	91,8 %	86,5 %	0,563
	Logit	80,0 %	0,9 %	91,4 %	93,0 %	-
	Knn3	68,0 %	2,6 %	91,0 %	80,3 %	0,438
	Knn9 (w)	12,0 %	31,6 %	70,3 %	86,0 %	6,55E-10
	Knn15 (w)	12,0 %	31,6 %	70,3 %	86,0 %	6,55E-10
	Knn5 (w)	8,0 %	32,9 %	69,5 %	86,9 %	2,15E-10
	Knn3 (w)	8,0 %	38,5 %	64,5 %	85,6 %	9,70E-14

Konkurssin ennakoinnissa kahta vuotta ennen konkurssiajanhetkeä, luokittelutarkkuus alenee suhteessa vuotta ennen konkurssia saatuun luokittelutarkkuuteen (taulukko 13). Keskimäärin luokittelutarkkuus näyttäisi alentuvan noin 10 %, kun verrataan 2016 aineistoa vuoden 2015 aineistoon otossuhteella 50/50 %. Logistinen regressio on luokittelutarkkuuden perusteella paras menetelmä kummallakin otossuhteella. Luokittelutarkkuus on noin 6-7 % parempi 50/50 % aineistossa, kuin parhaiten luokitelleella knn-menetelmällä. 10/90 % aineistossa menetelmien tarkkuuserot ovat marginaalisempia. AUC-arvolla mitattuna paras menetelmä vaikuttaisi olevan k:n lähimmän naapurin menetelmä. Tämä voi johtua siitä, että AUC-arvo mittaa luokittelutarkkuutta kaikilla mahdollisilla kriittisen pisteen arvoilla. Kriittinen piste on ehdollisen todennäköisyyden raja-arvo, jonka ylittymisen johdosta yritys luokitellaan konkurssiyritykseksi. Tässä tapauksessa, kun huomioidaan kaikki kriittisen pisteen vaihtoehtoiset arvot, toimii k:n lähimmän naapurin menetelmä luokittelussa parhaiten.

K:n lähimmän naapurin osalta tulosten perusteella voidaan nähdä, että muuttujien keskinäisellä hyvyydellä painotettu knn-menetelmä antoi parhaat luokittelutulokset sekä 50/50 % aineistolla, että 10/90% aineistolla verrattuna painottamattomaan knn-menetelmään. Lisäksi voidaan havaita, että 2016 aineistossa nähty trendi virhetyyppi 1 suhteellisesti pienemmälle määrälle on havaittavissa myös kahta vuotta ennen konkurssia saaduissa tuloksissa. Logistiseen regressioon verrattuna painotetussa k:n lähimmän naapurin menetelmässä virhetyyppi 1 osuus on kaksi kertaa alhaisempi, kuin logistisella regressiolla. (taulukko 13.)

Verrattaessa luokittelutarkkuutta (ACC) kahdella eri otossuhteella, voidaan havaita, että 10/90 aineiston osalta luokittelutarkkuudet ovat korkeampia, kuin 50/50 % aineistossa. Tämä johtuu siitä, että aineistojen keskinäinen epäsuhta konkurssiyritysten ja terveiden yritysten määrien suhteen vääristää osaltaan luokittelutarkkuuden tuloksia. 10/90 % aineisto sisältää suhteessa niin paljon terveitä yrityksiä, että vaikka malli luokittelisi kaikki yritykset terveiksi yrityksiksi, niin sen luokittelutarkkuus olisi tästä huolimatta 90 %. Tästä syystä otossuhteiden ollessa jotain muuta, kuin 50/50 %, ei luokittelutarkkuutta kuvaava tunnusluku ole aina luotettava. (taulukko 13.)

Logistisen regression ja k:n lähimmän naapurin menetelmän luokittelutarkkuudessa havaittiin tilastollisesti merkitsevä ero painottamattomien knn5, knn9 ja knn3 tapauksessa

otossuhteella 50/50 % (taulukko 13). Muiden osalta ero logistiseen regressioon ei ollut tilastollisesti merkitsevä. Otossuhteella 10/90 sitä vastoin menetelmien väliset erot olivat niin pieniä, että z-testin perusteella minkään knn-menetelmän luokittelutarkkuus ei ollut tilastollisesti merkitsevä verrattuna logistiseen regressioon.

Taulukko 13. Luokittelutulokset kahta vuotta ennen konkurssia. Tulokset laskettu testi-aineistoon.

Otossuhde	Malli	Virhetyyppi 1	Virhetyyppi 2	ACC	AUC	p-arvo
50/50	Logit	24,0 %	19,0 %	78,3 %	79,6 %	-
	Knn15 w	12,0 %	47,6 %	71,7 %	86,3 %	0,235
	Knn3 w	8,0 %	57,1 %	69,6 %	78,6 %	0,171
	Knn5 w	8,0 %	57,1 %	69,6 %	78,1 %	0,171
	Knn9 w	8,0 %	57,1 %	69,6 %	78,1 %	0,171
	Knn15	44,0 %	23,8 %	65,2 %	74,3 %	0,082
	Knn5	40,0 %	47,6 %	56,5 %	63,0 %	0,013
	Knn9	40,0 %	47,6 %	56,5 %	63,0 %	0,013
	Knn3	48,0 %	42,9 %	54,3 %	61,3 %	0,008
10/90	Logit	92,0 %	0,9 %	90,2 %	80,9 %	-
	Knn9 w	88,0 %	1,7 %	89,8 %	68,6 %	0,441
	Knn15 w	88,0 %	1,7 %	89,8 %	68,6 %	0,441
	Knn3 w	92,0 %	1,7 %	89,5 %	66,3 %	0,385
	Knn5 w	92,0 %	2,2 %	89,1 %	66,8 %	0,332
	Knn9	92,0 %	3,0 %	88,3 %	67,9 %	0,238
	Knn15	92,0 %	3,0 %	88,3 %	67,9 %	0,238
	Knn5	88,0 %	4,8 %	87,1 %	61,9 %	0,132
	Knn3	80,0 %	7,4 %	85,5 %	59,6 %	0,052

Taulukossa 14 on esitetty luokittelutulokset kolmea vuotta ennen konkurssihetkeä. Tulokset osoittavat selkeästi, että mitä aikaisemmin konkurssia ennakoidaan, niin sitä epätarkemmaksi ennuste muuttuu. Luokittelutarkkuudet laskevat menetelmästä riippumatta noin 56-65 % tasolle 50/50 % aineiston otossuhteella. 10/90 % aineiston tapauksessa luokittelutarkkuuksissa ei ole juurikaan muutosta suhteessa 2015 luokittelutuloksiin. 10/90 % otossuhteesta saadut luokittelutarkkuudet ovat korkeita, koska malli luokittelee lähes kaikki yritykset terveiksi, joka otossuhteen vinoutuneisuuden vuoksi näkyy korkeana luokittelutarkkuutena.

Kolme vuotta ennen konkurssihetkeä k:n lähimmän naapurin menetelmällä saatiin hieman korkeammat luokittelutarkkuudet, kuin logistisella regressiolla. Otossuhteella 50/50 % parhaan knn-menetelmän ja logistisen regression ero oli noin 5 % luokittelutarkkuuden osalta. Parhain luokittelutulos saatiin käytettäessä muuttujien keskinäistä hyvyyttä painotuksena ja naapureiden määränä 9:ää lähintä naapuria. Selkeitä eroja painotetun ja painottamattoman k:n lähimmän naapurin välillä ei ole kuitenkaan havaittavissa. Jos tarkastellaan virhetyyppi 1 ja 2 suhteellisia määriä, niin painotetussa menetelmässä virhetyyppi 1 määrä on alempi, kuin painottamattomassa menetelmässä. Tämä tarkoittaa sitä, että painotettuna k:n lähimmän naapurin menetelmä luokittelee herkemmin yrityksen konkurssiyritykseksi kuin terveeksi yritykseksi. Virhetyyppi 2 määrä sen sijaan on huomattavasti korkeampi, kuin painottamattomalla k:n lähimmän naapurin menetelmällä. Tällöin terveet yritykset on virheellisesti luokiteltu konkurssiyritykseksi, joka laskee menetelmän yhteenlasketun luokittelutarkkuuden kohtalaiselle tasolle. Painottamassa menetelmässä virhetyyppien suhteelliset määrät ovat lähes tasapainossa toistensa suhteen. (taulukko 14.)

Kolme vuotta ennen konkurssihetkeä menetelmien välisissä luokittelutarkkuuksissa ei havaittu olevan tilastollisesti merkitsevää eroa millään mallilla tai otossuhteella verrattaessa logistisella regressiolla saatuun luokittelutarkkuuteen (taulukko 14). Otossuhteella 50/50% ja 10/90 % logistisen regression ja k:n lähimmän naapurin menetelmän ero luokittelutarkkuudessa olisi tullut olla vähintään noin 20 %, jotta z-testin tulos olisi ollut tilastollisesti merkitsevä. Tulosten mukaan kuitenkin ero oli enimmillään vain noin 5 %.

Taulukko 14. Luokittelutulokset kolmea vuotta ennen konkurssia. Tulokset laskettu testiaineistoon.

Otossuhde	Malli	Virhetyyppi 1	Virhetyyppi 2	ACC	AUC	p-arvo
50/50	Knn9 w	12,0 %	61,9 %	65,2 %	67,2 %	0,667
	Knn3	36,0 %	38,1 %	63,0 %	58,7 %	0,585
	Knn5	36,0 %	38,1 %	63,0 %	58,7 %	0,585
	Knn15 w	8,0 %	71,4 %	63,0 %	69,1 %	0,585
	Logit	44,0 %	33,3 %	60,9 %	65,1 %	-
	Knn9	36,0 %	42,9 %	60,9 %	63,4 %	0,500
	Knn15	36,0 %	42,9 %	60,9 %	64,8 %	0,500
	Knn3 w	40,0 %	47,6 %	56,5 %	62,3 %	0,336
	Knn5 w	40,0 %	47,6 %	56,5 %	62,3 %	0,336
10/90	Knn15	96,0 %	1,3 %	89,5 %	67,7 %	0,711
	Knn9	92,0 %	2,2 %	89,1 %	64,8 %	0,661
	Knn9 w	100,0 %	1,3 %	89,1 %	68,4 %	0,661
	Knn15 w	100,0 %	1,3 %	89,1 %	70,1 %	0,661
	Knn3 w	92,0 %	3,0 %	88,3 %	67,3 %	0,554
	Knn5 w	96,0 %	2,6 %	88,3 %	67,6 %	0,554
	Logit	96,0 %	3,0 %	87,9 %	71,1 %	-
	Knn5	88,0 %	5,2 %	86,7 %	62,5 %	0,345
	Knn3	88,0 %	6,1 %	85,9 %	63,6 %	0,256

6.3. Tulosten suhde tutkimuksen hypoteeseihin ja aikaisempaan tutkimukseen

Tutkimuksen ensimmäinen hypoteesi on luonteeltaan nollahypoteesi, jossa käsiteltiin logistisen regression ja k:n lähimmän naapurin menetelmän välisen luokittelutarkkuuden eroa. Hypoteesiin voidaan vastata laskemalla tilastollinen merkitsevyys luokittelutarkkuuksissa logistisen regression ja k:n lähimmän naapurin välillä. Tulosten perusteella nollahypoteesi jää voimaan, koska selkeää yleistettävissä olevaa tilastollisesti merkitsevää eroa ei ollut havaittavissa logistisen regression ja k:n lähimmän naapurin menetelmän välillä (taulukot 12, 13, 14). Luokittelutarkkuuksissa oli yksittäisiä tilastollisesti merkitseviä eroja, mutta keskimäärin eroja voidaan kuitenkin pitää tilastollisesti merkitsemättöminä.

Tutkimuksen toinen hypoteesi otti kantaa mallinnus- ja testiaineiston sisältämien konkurssi- ja terveiden yritysten lukumäärän suhteeseen ja sen vaikutukseen

luokittelutarkkuuden kannalta. Hypoteesia testattiin oikaisemalla 10/90 % otossuhdetta vastaavat tulokset vertailukelpoiseksi 50/50% aineistossa laskettujen tulosten kanssa. Oikaisu tehtiin käyttämällä 50/50 % testiaineiston sisältämiä konkurssi- ja terveiden yritysten lukumääriä ja 10/90 % testiaineistoon ennustettuja luokittelutuloksia virhetyyppi 1-% ja virhetyyppi 2-% osalta. Oikaistut luokittelutarkkuudet on esitetty taulukossa 15. Taulukkoon on koottu otossuhteittain ja ajankohdittain logistisella regressiolla saadut luokittelutarkkuudet ja lisäksi korkeimmat luokittelutarkkuudet antanut painottamaton- ja painotettu k:n lähimmän naapurin malli. Z-testin kautta lasketut p-arvot ovat laskettu vertaamalla aina samalla rivillä taulukossa 15 esitettyjä luokittelutarkkuuksia keskenään. Tulosten perusteella mallin sovittamiseen käytetyllä aineiston otossuhteella on tilastollisesti merkitsevä vaikutus luokittelutarkkuuteen keskimäärin. P-arvoista noin 66 % oli suuruudeltaan alle 0,05.

Taulukko 15. Otossuhteen vaikutus luokittelutarkkuuteen. 10/90 %:in aineiston osalta luokittelutarkkuudet ovat oikaistuja (*) vastaamaan 50/50 % otossuhdetta.

Otossuhde	**50/50 %**		**10/90 %**		**p-arvo**
Vuosi	**Malli**	**ACC**	**Malli**	**ACC***	
2016	Logit	87,0 %	Logit	56,1 %	0,001
	Knn15	82,6 %	Knn5	62,5 %	0,015
	Knn9 w	65,2 %	Knn9 w	79,1 %	0,931
2015	Logit	78,3 %	Logit	49,6 %	0,002
	Knn15	65,2 %	Knn20	47,2 %	0,041
	Knn15 w	71,7 %	Knn20 w	49,6 %	0,015
2014	Logit	60,9 %	Logit	46,4 %	0,083
	Knn3	63,0 %	Knn15	47,2 %	0,064
	Knn9 w	65,2 %	Knn20 w	45,3 %	0,027

Kolmas hypoteesi käsitteli lähimpien naapureiden määrän vaikutusta menetelmän luokittelutarkkuuteen. Aikaisemmin taulukoissa 12-14 esitettyjen tulosten lisäksi laajennettiin luokittelutarkkuuksissa käytettyjä naapureiden lukumääriä ottamalla mukaan k:n arvot 1 ja 20. Kaiken kaikkiaan luokittelutarkkuus laskettiin viidellä eri k:n arvolla (1, 3, 5, 9, 15, 20). Taulukkoon 16 on koostettu yhteenveto luokittelutarkkuuksien ääriarvoista, eli korkein luokittelutarkkuus ja heikoin luokittelutarkkuus eri naapureiden määrillä. Näistä

laskettiin z-testillä tilastollista merkitsevyyttä kuvaava p-arvo eron suuruudelle. Luokittelutarkkuuden osalta huonoin tarkkuus saatiin yleisimmin k:n arvolla 1, kun taas parhaimman tarkkuuden tapauksessa k:n arvo oli yleisimmin 9 tai enemmän. Tilastollisesti merkitseviä eroja ei havaittu muissa kuin vuoden 2015 osalta otossuhteella 10/90 % ja vuoden 2014 osalta painottamattomalla knn-mallilla ja otossuhteella 10/90 %. Naapureiden määrällä on vaikututusta luokittelutarkkuuteen, mutta vaikutus ei ole kuitenkaan usein tilastollisesti merkittävällä tasolla.

Taulukko 16. Lähimpien naapureiden määrän vaikutus luokittelutarkkuuteen.

Otossuhde	**50/50 %**			**10/90 %**		
Vuosi	**Malli**	**ACC**	**p-arvo**	**Malli**	**ACC**	**p-arvo**
2016	Knn15	82,6 %	0,220	Knn5	92,2 %	0,068
	Knn3	76,1 %		Knn1	88,3 %	
	Knn9 w	65,2 %	0,333	Knn9 w	70,3 %	0,079
	Knn1 w	60,9 %		Knn3 w	64,5 %	
2015	Knn15	65,2 %	0,070	Knn20	89,5 %	0,020
	Knn1	50,0 %		Knn1	83,2 %	
	Knn15 w	71,7 %	0,187	Knn20 w	90,2 %	0,040
	Knn1 w	63,0 %		Knn1 w	85,2 %	
2014	Knn3	63,0 %	0,335	Knn15	89,5 %	0,008
	Knn1	58,7 %		Knn1	82,0 %	
	Knn9 w	65,2 %	0,196	Knn20 w	89,5 %	0,139
	Knn5 w	56,5 %		Knn1 w	86,3 %	

Neljäs hypoteesi käsitteli selittävien muuttujien painottamista ennakkotietoon perustuen ja sen merkitystä luokittelutarkkuuden kannalta. Hypoteesia testattiin poimimalla painotetuilla ja painottamattomilla malleilla lasketut parhaat luokittelutarkkuudet ja vertaamalla luokittelutarkkuuksien arvoja z-testin, ja siitä lasketun p-arvon avulla. Taulukkoon 17 on koostettu parhaimmat luokittelutarkkuudet antanut painottamaton ja painotettu

malli jokaista ajankohtaa ja otossuhdetta vastaavasti. Tulosten perusteella nähdään, että vuotta ennen konkurssia painottamaton k:n lähimmän naapurin menetelmä oli tilastollisesti merkitsevällä tasolla tarkempi, kuin painottamaton menetelmä kummallakin otossuhteella. Kaksi ja kolme vuotta ennen konkurssia painotuksella ei ollut tilastollisesti merkitsevää vaikutusta luokittelutarkkuuteen kummallakaan otossuhteella.

Taulukko 17. Muuttujien painotuksen vaikutus luokittelutarkkuuteen.

Otossuhde	**50/50 %**			**10/90 %**		
Vuosi	**Malli**	**ACC**	**p-arvo**	**Malli**	**ACC**	**p-arvo**
2016	Knn15	82,6 %	0,029	Knn5	92,2 %	1,14E-10
	Knn9 w	65,2 %		Knn9 w	70,3 %	
2015	Knn15	65,2 %	0,750	Knn20	89,5 %	0,615
	Knn15 w	71,7 %		Knn20 w	90,2 %	
2014	Knn3	63,0 %	0,586	Knn15	89,5 %	-
	Knn9 w	65,2 %		Knn1 w	89,5 %	

Aikaisempaan tutkimukseen verrattuna tässä tutkimuksessa saadut tulokset ovat hyvin samansuuntaisia kuin muissakin tutkimuksissa. Luokittelutarkkuudella mitattuna esimerkiksi Tam ja Kiang (1992), Park ja Han (2002), Bien ja Mazlack (2003), Pietruszkiewicz (2008), Chen ym. (2011) ja Zhao ym. (2016) ovat saaneet hyvin vastaavalla tasolla olevia tuloksia. Tässä tutkimuksessa havaittiin, että otossuhteella 50/50 % virhetyyppi 1 ja 2 suhteet ovat painottamattomassa menetelmässä lähes tasapainossa ja painotetusta menetelmästä saaduissa tuloksissa virhetyyppi 1 määrä on korostuneen alhainen verrattuna virhetyyppi 2 määrään. Osassa aikaisemmista tutkimuksista on havaittu suhteiden olevan päinvastaisia, eli virhetyyppi 1 suhteellinen määrä korostuu virhetyyppi 2 määrään verrattuna painottamattomassa k:n lähimmän naapurin menetelmässä (kts. Tam ja Kiang 1992; Serrano-Cinca ja Gutiérrez-Nieto 2013). Päinvastaisiakin tuloksia ja siten tässä tutkimuksessa saatuja tuloksia tukeviakin virhetyyppien suhteellisia määriä on esitetty (kts. Ribeiro ym. 2008; Pietruszkiewicz 2008; Chen ym. 2011; Zhao ym. 2016 ja Alrashed ja Che 2018).

Tässä tutkimuksessa saadut luokitteluerot eri otossuhteilla ovat myös suuruudeltaan vastaavia kuin aikaisemmassa tutkimuksessa. Esimerkiksi Kiviluodon (1998) tekemässä tutkimuksessa virhetyyppien keskinäiset suhteet olivat lähes identtiset tämän tutkimuksen

kanssa otossuhteella 10/90 % vuotta ennen konkurssia. Kiviluodon (1998) aineistossa konkurssiyritysten suhteellinen määrä oli noin 26 % koko aineiston koosta.

Aikaisemmista tutkimuksista ainoastaan Alrasheed ja Che (2018) ovat vertailleet k:n lähimmän naapurin luokittelutuloksia vaihtelevilla aineiston otossuhteilla. He havaitsivat selkeän trendin virhetyyppi 1 suhteellisen määrän kasvussa silloin, kun konkurssiyritysten määrä aineistossa on pieni. Tämä aikaisemmin raportoitu tulos on yhteneväinen tässä tutkimuksessa saatujen tulosten kanssa. Alrasheed ja Che (2018) havaitsivat k:n lähimmän naapurin menetelmän olevan logistista regressiota vähemmän herkempi otossuhteissa tapahtuville muutoksille. Tätä havaintoa ei kyetty tässä tutkimuksessa suoraan vahvistamaan.

6.4. Tulosten yleistettävyys

Jos tarkastellaan luokittelutuloksia vuodelta 2016 ja 2015 otossuhteella 50/50 %, niin on selkeästi havaittavissa logistisen regression olevan kahdesta menetelmästä suorituskyvyltään parempi menetelmä (taulukko 12 ja 13). Luokittelutarkkuuksien perusteella ero k:n lähimmän naapurin menetelmän ja logistisen regression välillä on ilmeinen. Menetelmien välillä ei kuitenkaan siitä huolimatta ole selkeästi yleistettävissä olevaa tilastollista eroa luokittelutarkkuuksien eroissa.

Logistinen regressio toimi luokittelutarkkuuden perusteella selkeästi parhaiten vuonna 2016 ja 2015. Vuodelta 2014, eli kolme vuotta ennen konkurssihetkeä tulokset kuitenkin osoittavat, että k:n lähimmän naapurin menetelmä kykeni ennakoimaan konkurssin tapahtumisen logistista regressiota tarkemmin. Vaikka ero luokittelutarkkuudessa kolme vuotta ennen konkurssia oli suhteellisen selkeä (n. 4 %) knn-menetelmän eduksi, niin tulee silti tulosten yleistettävyyden kannalta tiedostaa saatujen luokittelutarkkuuksien todellinen taso. Jos esimerkiksi tarkastellaan vuoden 2014 osalta saatuja AUC-arvoja otossuhteella 50/50 %, niin kummallakin menetelmällä AUC-arvolla mitattu tarkkuus jää alle 0,7. Esimerkiksi Hosmerin ja Lemeshowin (2000:160-165) mukaan luokittelumallin tarkkuus AUC-arvolla mitattuna tulisi ylittää vähintään 0,7 raja-arvo, jotta mallin luokittelutarkkuutta voitaisiin pitää hyvällä tasolla.

Yhtenä merkittävimpänä yleistettävyyttä vaarantavana asiana on tässä tutkimuksessa käytetty konkurssiyritysaineisto. Aineiston pohjaksi saatiin Asiakastieto Oyj:ltä käyttöön kaikki Suomessa vuonna 2017 konkurssiin menneet yritykset, joista oli ylipäätään olemassa tilinpäätöstietoja. Näitä yrityksiä oli lukumääräisesti 1402. Tunnusluvut ja niiden saatavuus osoittautuivat kuitenkin merkittäväksi rajoittavaksi tekijäksi tutkimuksen kannalta ja tämän vuoksi lopullinen konkurssiyritysaineisto supistui 86 yritykseen. Lopullinen aineisto oli siis vain noin 6 % alkuperäisestä populaatiosta. Koska alkuperäisestä aineistosta osoittautui käyttökelpoiseksi vain näin pieni osa yrityksiä, ei voida sulkea pois aineistossa olevaa systemaattista virhettä, eikä riskiä siitä, että tutkimuksen tulokset sisältävät saman tai kertautuneen systemaattisen virheen kuin aineisto. Koska mahdollisuutta tunnuslukujen hankkimiseen kaikilta yrityksiltä ei ole, niin ei voida myöskään todentaa mahdollisen virhelähteen mittaluokkaa tai laajuutta.

Asiakastiedolta saadulla aineistolla oli kuitenkin myös merkittäviä positiivisia vaikutuksia tulosten luotettavuuden näkökulmasta. Keskeisimpänä niistä oli luotettava tieto siitä, mitkä yritykset ovat todellisuudessa konkurssiyrityksiä. Valmista aineistoa voidaan pitää tutkimuksen yleistettävyyden kannalta keskeisenä. Jos aineistoa ei olisi saatu Asiakastiedolta, ei sataprosenttisesti luotettavia keinoja yritysten konkurssistatuksen validoimiseen olisi ollut käytössä ja tämä olisi voinut merkittävästi vääristää tutkimuksen tuloksia ja mallien keskinäistä vertailukelpoisuutta.

Aineiston käsittely ja osa tulosten laskennasta tehtiin manuaalisesti taulukkolaskenta ohjelmisto Microsoft Exceliä käyttämällä. Tämän vuoksi ei voida sulkea pois riskiä inhimillisestä virheestä, jonka seurauksesta matemaattisen oikeellisuuden kautta olisi tullut virheitä aineistoon tai tuloksiin. Lisäksi mahdollisia inhimillisiä kirjoitusvirheitä, pilkkuvirheitä tai merkintävirheitä ei voida täysin sulkea pois, vaikka niiden todennäköisyyttä voidaankin pitää suhteellisen alhaisena.

Yksi mahdollinen virhelähde voi lisäksi sisältyä terveiden yritysten hankintaan. Terveet yritykset etsittiin vastinparimenettelyä hyödyntäen Orbis-tietokannasta ja lopulliset aineistoon sisällytetyt yritykset valittiin satunnaisesti. Yritysten suodattamisessa käytettiin Orbiksessa kokorajoitteita ja kannattavuusrajoitetta. Hakuehdoksi haettaville terveille yrityksille asetettiin tilikauden tuloksen keskimääräinen positiivisuus viimeisen 4 vuoden aikana, jolla voitiin alentaa riskiä, että terveiden yritysten joukkoon tulisi mukaan

konkurssiyrityksiä. Jälkikäteen ajateltuna tämä rajoite on saattanut olla epäviisas, sillä tilikauden tuloksen negatiivisuus ei ole suoraan verrannollinen konkurssin joutumiseen. Terveilläkin yrityksillä voi olla välillä negatiivisia tilikauden tuloksia. Tilikauden tuloksen positiivisuusrajoite on voinut liiaksi eriyttää aineistossa olevien terveiden ja konkurssiyritysten rakennetta, jolla on voinut olla vaikutusta tutkimuksen tuloksiin. Esimerkiksi jos tarkastellaan tutkimuksessa käytettyjen tunnuslukujen tärkeyttä toistensa suhteen, niin kannattavuuden tunnusluvut ovat lähes kaikki, mukaan lukien erityisesti tilikauden tulos, hyvin konkurssiriskin kanssa korreloivia. Tästä syystä on mahdollista, että vastinparimenettelyn yhteydessä käytetty kannattavuusrajoite näkyy siten jollain tasolla myös saaduissa tuloksissa.

7. YHTEENVETO

Konkurssin ennakointia käsitteleviä tutkimuksia on tehty Suomessa ja maailmalla määrällisesti hyvin paljon. Tutkimuksellisesti aihe on ollut runsaan kiinnostuksen kohteena, koska konkurssin aiheuttamat taloudelliset ja sosiaaliset tappiot niin yrityksen omistajille kuin sidosryhmillekin ovat suuria. Aiheesta tehdyssä tutkimuksessa on havaittu jo varhain, että konkurssin todennäköisyyttä voidaan ennakoida erilaisten kannattavuuden, maksuvalmiuden ja vakavaraisuuden tunnuslukujen avulla yrityksen tilinpäätöstiedoista. Konkurssin ennakointiin on käytetty perinteisesti esimerkiksi lineaarista erotteluanalyysiä ja logistista regressiota. Tietotekninen kehitys ja tietokoneiden laskentatehossa tapahtunut kasvu on lisännyt erilaisten koneoppimisen algoritmien yleisyyttä luokitteluongelmien ratkaisussa ja myös konkurssin ennakoinnissa vastaavat menetelmät ovat olleet viime aikoina runsaan tutkimuksen kohteena.

K:n lähimmän naapurin menetelmä on luokittelualgoritmi, joka on kehitetty jo Coverin ja Hartin (1967) tutkimuksessa 60-luvun loppupuolella. Menetelmää on hyödynnetty lukuisissa erilaisissa käyttökohteissa aina metsäalalta laskentatoimen ja rahoituksen aihealueisiin. Vaikka menetelmä on suhteellisen vanha, ei sitä ole kansainvälisesti hyödynnetty kuin noin reilussa kymmenessä konkurssin ennakointitutkimuksessa. Suomessa tehtyjä tutkimuksia on tiettävästi vain yksi (Kiviluoto 1998). K:n lähimmän naapurin menetelmällä on tiettyjä vahvuuksia muihin koneoppimisen algoritmeihin verrattuna, kuten algoritmin intuitiivisuus, laskennallinen keveys ja yksinkertaisuus. Aikaisemman tutkimuksen vähyys Suomen olosuhteissa ja suomalaisella yritysaineistolla kannustaa tutkimaan menetelmän käyttökelpoisuutta ja suorituskykyä suhteessa runsaammin käytössä oleviin ennakointimenetelmiin.

Tässä tutkimuksessa koostettiin konkurssiyritysaineisto vuonna 2017 konkurssiin menneistä yrityksistä kolmen konkurssihetkeä edeltävän vuoden tilinpäätöstunnusluvuista. Lopullinen konkurssiyritysten määrä oli 86 yritystä. Konkurssiyrityksille etsittiin Beaverin (1966) esittämällä vastinparimenettelyllä niitä rakenteellisesti vastaavat terveet yritykset. Terveiden yritysten avulla koostettiin kaksi otossuhteiltaan erilaista aineistoa, joissa konkurssiyritysten määrä oli vakio, mutta terveiden yritysten määrä oli ensimmäisessä aineistossa 86 yritystä ja toisessa 744 yritystä. Tällöin saatiin otossuhteet

konkurssiyritysten ja terveiden yritysten välillä 50/50 % ja 10/90 %. Otossuhteiden valinnassa käytettiin esikuvana Alrasheedin ja Chen (2018) julkaisemaa tutkimusta.

K:n lähimmän naapurin menetelmän luokittelutuloksia verrattiin tässä tutkimuksessa logistisella regressiolla laskettuihin luokittelutuloksiin. Mallien sovittaminen tehtiin 70 % osuutta vastaaviin määriin aineistoa ja ennustamisessa käytettiin jäljelle jäänyttä 30 % osuutta aineistosta. K:n lähimmän naapurin menetelmän eri parametriyhdistelmiä, kuten naapureiden lukumäärän, etäisyysfunktion muodon, naapureiden etäisyysperusteisen painottamisen ja selittävien muuttujien keskinäisen painotuksen vaikutusta verrattiin suhteessa luokittelutarkkuuden kehittymiseen. Optimaalisten parametriyhdistelmän löydyttyä sillä laskettuja luokittelutuloksia verrattiin logistisen regression luokittelutuloksiin. Luokittelutuloksissa olevan eron tilastollista merkitsevyyttä tarkasteltiin z-testillä laskettujen p-arvojen perusteella.

Ensimmäisessä tutkimuksen hypoteesissa tutkittiin, että onko k:n lähimmän naapurin menetelmällä ja logistisella regressiolla tilastollisesti merkitsevää eroa luokittelutarkkuuksissa. Tulosten perusteella havaittiin, että otossuhteista riippumatta logistinen regressio toimii lähtökohtaisesti paremmin konkurssin- ja terveen yrityksen välisessä luokittelussa. Erot luokittelutarkkuudessa olivat sitä selkeämpiä, mitä lähempänä konkurssihetkeä luokitus tehtiin. Luokittelutarkkuudet k:n lähimmän naapurin menetelmällä olivat yhtä-, kahta- ja kolmea vuotta ennen konkurssia 82,6 %, 71,7 % ja 65,2 %. Logistisella regressiolla saatiin vastaaviksi luokittelutarkkuuksiksi 87,0 %, 78,3 % ja 60,9 %. Luokittelutarkkuuksien erot eivät kuitenkaan olleet z-testin perusteella tilastollisesti merkitseviä.

Toisen hypoteesin perusteella tutkittiin, onko konkurssiyritysten ja terveiden yritysten lukumäärien suhteella mallinsovitusaineistossa vaikutusta luokittelutarkkuuteen. Tulosten perusteella havaittiin, että jos konkurssiyritysten määrä on pieni (10 %) suhteessa koko aineiston yritysten lukumäärään, lisääntyvät mallilla ennustetussa luokituksessa virhetyyppi 1 suhteellisen osuus, jolloin konkurssiyritys luokitellaan terveeksi yritykseksi. Virhetyyppi 1 lisääntynyt osuus vaikutti mallien luokittelutarkkuuteen tilastollisesti merkitsevällä tasolla verrattaessa aineistoon, jossa suhteet olivat 50/50 %.

Kolmannessa hypoteesissa tutkittiin lähimpien naapureiden lukumäärän vaikutusta luokittelutarkkuuteen. Tulosten joukosta poimittiin heikoiten suoriutuneet mallivaihtoehdot ja vastaavasti parhaiten luokittelutarkkuuden perusteella toimineet mallit. Kaikki muut

tekijät vakioitiin, paitsi lähimpien naapureiden lukumäärä. Vuotta ennen konkurssia luokittelutarkkuuden vaihteluväli oli 4-6 %, kahta vuotta ennen konkurssia 5-15 % ja kolme vuotta ennen konkurssia 4-7 %. Erot eivät olleet tilastollisesti merkitsevällä tasolla, joten naapureiden määrällä ei havaittu olevan tilastollisesti merkitsevää vaikutusta luokittelutarkkuuteen.

Neljännessä hypoteesissa tutkittiin, onko selittävien muuttujien keskinäisellä painotuksella vaikutusta luokittelutarkkuuden hyvyyteen. Esikuvatutkimuksena tässä käytettiin Pakin ja Hanin (2002) tutkimusta, jossa on aikaisemmin tutkittu vaihtoehtoisia painotusmenetelmiä. Tässä tutkimuksessa painotukset määritettiin tunnuslukujen selityskyvyn perusteella. Tulosten perusteella havaittiin, että muuttujien keskinäinen painotus lisää todennäköisyyttä, että yritys luokitellaan konkurssiyritykseksi. Vuotta ennen konkurssia painotus huononsi luokittelutarkkuutta kokonaisuudessaan tilastollisesti merkitsevällä tasolla. Kahta vuotta ennen konkurssia vaikutus oli luokittelutarkkuutta parantava, mutta suuruudeltaan tilastollisesti merkitsemätön. Kolme vuotta ennen konkurssia painotuksen vaikutus luokittelutarkkuuteen oli hyvin marginaalinen ja siksi tilastollisesti ei-merkitsevä.

LÄHDELUETTELO

Aghaie, A., & A. Saeedi (2009). Using bayesian networks for bankruptcy prediction: Empirical evidence from Iranian companies. *Information Management and Engineering*. 450-455

Alrasheed, D., & D. Che. (2018). Improving Bankruptcy Prediction Using Oversampling and Feature Selection Techniques. *The 2018 World congress in computer science, computer engineering & applied computing*. 440-446. SBN: 1-60132-480-4

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*. 23:4. 589-609.

Altman, E. I. (1980). Commercial bank lending: process, credit scoring, and costs of errors in lending. *Journal of Financial and Quantitative Analysis*. 15:4. 813-832.

Amani, F. A., & A. M. Fadlalla. (2017). Data mining applications in accounting: A review of the literature and organizing framework. *International Journal of Accounting Information Systems*. 24. 32-58.

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on neural networks*. 12:4. 929-935.

Back, B., T. Laitinen & K. Sere (1996). Neural networks and genetic algorithms for bankruptcy predictions. *Expert Systems with Applications*. 11:4. 407-413.

Barboza, F., H. Kimura & E. Altman (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications.* 83. 405-417.

Beaver, W., H. (1966). Financial ratios as Predictors of Failure. *Journal of accounting research.* 71-111.

Bian, H., & L. Mazlack (2003). Fuzzy-rough nearest-neighbor classification approach. *In Fuzzy Information Processing Society. 22nd International Conference of the North American* 500-505.

Blum, M. (1974). Failing company discriminant analysis. *Journal of Accounting Research* 12:1. 1-25.

Breiman, L. (1996). Bagging predictors. *Machine learning*. 24:2. 123-140.

Breiman, L. (2001). Random forests. *Machine learning*. 45:1. 5-32.

Chen, H. L., B. Yang, G. Wang, J. Liu, X. Xu, S. J. Wang & D. Y. Liu (2011). A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method. *Knowledge-Based Systems.* 24:8. 1348-1359.

Cover, T., & P. Hart (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory.* 13:1. 21-27.

Deakin, E. (1972). A discriminant analysis of predictors of business failure. *Journal of Accounting Research* 1:5. 167-179.

Diamond, H. (1976). Pattern recognition and the detection of corporate failure. *Ph.D. Dissertation. New York University.*

Edminister, R. (1972). An empirical test of financial ratio analysis for small business failure prediction. *Journal of Financial and Quantitive analysis* 2:3 1477-1493.

Engelmann, B., E. Hayden & D. Tasche (2003). Measuring the discriminative power of rating systems. *Banking and Financial Supervision*. 2003:1. ISBN 3–935821–67–0

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*. 14:1. 153-158.

Gao, Z., M. Cui & L. M. Po (2008). Enterprise bankruptcy prediction using noisy-tolerant support vector machine. *In Future Information Technology and Management Engineering, 2008. FITME'08. International Seminar*. 153-156.

Gilbert, L. R., K. Menon & K. B. Schwartz (1990). Predicting bankruptcy for firms in financial distress. *Journal of Business Finance & Accounting*. 17:1. 161-171.

Haara, A. (2002). Kasvuennusteiden luotettavuuden selvittäminen knn-menetelmällä ja monitavoiteoptimoinnilla. *Metsätieteen aikakauskirja* 2002:3. 391–406.

Hamer, M. (1983). Failure prediction: Sensitivity of classification accuracy to alternative statistical methods and variable sets. *Journal of accounting and public policy* 2. 289-307

Hechenbichler, K., & Schliep, K. (2004). Weighted k-nearest-neighbor techniques and ordinal classification. Collaborative Research Center 386. In University of Munich discussion paper 399. 16 s.

Imandoust, S. B., & M. Bolandraftar (2013). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. *International Journal of Engineering Research and Applications*. 3:5. 605-610.

Kalela, J., J. Kiander, U. Kivikuru, H.A. Loikkanen. & J. Simpura (2001). Down from the heavens, Up from the ashes, the economic crisis of the 1990s in the light of economic and social research. *Vatt puclications 27:6, Valtion taloudellinen tutkimuslaitos*, Gummerus Kirjapaino Oy: Helsinki. 543. ISBN 951-561-381-7.

Kaski, S., J. Sinkkonen & J. Peltonen (2001). Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*. 12:4. 936-947.

Keller, J. M., M. R. Gray & J. A. Givens (1985). A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics.* 4. 580-585.

Kiviluoto, K. (1998). Predicting bankruptcies with the self-organizing map. *Neurocomputing*. 21:1-3. 191-201.

Kolari, J., D. Glennon, H. Shin & M. Caputo (2002). Predicting large US commercial bank failures. *Journal of Economics and Business.* 54:4. 361-387.

Konkurssilaki 20.2.2004/120.

Koulu, R (2009). Konkurssioikeus. Juva: WS Bookwell Oy. 449 s. ISBN: 978-951-0 35247-2.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software.* 28:5. 1-26.

Kumar, P. R., & V. Ravi (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. *European journal of operational research*. 180:1. 1-28.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software.* 36:11. 1-13.

Laakso, T., E. K. Laitinen & H. Vento (2010). Uhkaava maksukyvyttömyys ja onnistunut yrityssaneeraus. Helsinki: Talentum Media Oy. 501 s. ISBN 978-952-14-1448-0.

Laitinen, E. (1990) Konkurssin ennustaminen. Sundom: Vaasan yritysinformaatio Oy. 302 s. ISBN 952-90-2236-0.

Laitinen, E. K. (1991). Financial ratios and different failure processes. *Journal of Business Finance & Accounting.* 18:5. 649-673.

Laitinen, E. K. & T. Laitinen (2004). Yrityksen rahoituskriisin ennustaminen. Helsinki: Talentum Media Oy. 405 s. ISBN 952-14-0771-9.

Laitinen E. K. & T. Laitinen (2014). Yrityksen maksukyky. Arviointi ja ennakointi. Helsinki: KHT Media Oy. 287 s. ISBN: 978-952-218-209-8.

Laitinen, T., & M. Kankaanpää (1997). Comparative analysis of failure prediction methods. Vaasan yliopiston julkaisuja 216. 78s. ISBN 951-683-678-x.

Laitinen, T., & M. Kankaanpää (1999). Comparative analysis of failure prediction methods: the Finnish case. *European Accounting Review*. 8:1. 67-92.

Laki yrityksen saneerauksesta 25.1.1993/47.

Lampela-Kivistö, L., H. Sorri & J. Kiiski (2001). Individual Survival. Down from the heavens, Up from the ashes, the economic crisis of the 1990s in the light of economic and social research. Edited by: Kalela, J., J. Kiander, U. Kivikuru, H. A. Loikkanen & J. Simpura (2001). *Vatt puclications 27:6, Valtion taloudellinen tutkimuslaitos*, Gummerus Kirjapaino Oy: Helsinki. 462-478s. ISBN 951-561-381-7

Lantz, B. (2013). Machine learning with R. *Packt Publishing Ltd.* ISBN 978-1-78216-214-8

Martin, D. (1977). Early warning of bank failure, a logit regression approach. *Journal of banking finance* 1:3. 249-276.

Meyer, P. A., & H. W. Pifer (1970). Prediction of bank failures. *The Journal of Finance.* 25:4. 853-868.

Miron B. Kursa, Witold R. Rudnicki (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, 36:11. 1-13.

Odom, M. D., & R. Sharda (1990). A neural network model for bankruptcy prediction. *In Neural Networks IJCNN International Joint Conference*. 163-168.

Park, C. S., & I. Han (2002). A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. *Expert Systems with Applications.* 23:3. 255-264.

Pietruszkiewicz, W. (2008). Dynamical systems and nonlinear Kalman filtering applied in classification. *Cybernetic Intelligent Systems. 7th IEEE International Conference.* 1-6.

Prihti, A. (1975). Konkurssin ennustaminen taseinformaation avulla. Helsingin kauppakorkeakoulu. *Acta Academiae Oeconimicae Helsingiensis*. A:13.

Pukkala, T. (2007). Metsäsuunnittelun menetelmät. Gummerus Kirjapaino Oy. Vaajakoski. 208s. ISBN 978-952-92-1731-1

Ranta, E., H. Rita & J. Kouki 2012. Biometria – Tilastotiedettä ekologeille. Yliopistopaino, Helsinki. 569 s.

R Development Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Ribeiro, B., A. Vieira, J. Duarte, C. Silva, J. C. Das Neves, Q. Liu & A. H. Sung (2008). Learning manifolds for bankruptcy analysis. Edited by: Köppen, M., N. Kasabov, & G. Coghill. (2008). Advances in neuro-information processing. *In International conference on neural information processing.* 723-730. Springer, Berlin, Heidelberg. ISSN 0302-9743

Serrano-Cinca, C., & B. Gutiérrez-Nieto (2013). Partial least square discriminant analysis for bankruptcy prediction. *Decision Support Systems.* 54:3. 1245-1255.

Shmueli, G., P. C. Bruce, M. L. Stephens & N. R. Patel (2017). Data mining for business analytics: concepts, techniques, and applications with JMP Pro. *John Wiley & Sons*. ISBN 978-111-887-752-4

Stokes, D., & R. Blackburn (2002). Learning the hard way: the lessons of owner-managers who have closed their businesses. *Journal of small business and enterprise development.* 9:1. 17-27.

Strang, L. (2000). Yritystoiminnan uhkatekijät – tunnista, ennakoi, selviydy. Enterprise Adviser -kirjasarja nro. 14. Gummerus Kirjapaino Oy Jyväskylä. 195s. ISBN 952-14-0233-4.

Sundgren, S. (1995). Bankruptcy costs and the bankruptcy code: A case study of the finnish code. *Publications of the Swedish School of Economics.* nro.61. 177. ISBN 951-555-462-4

Taffler, R. J. (1982). Forecasting company failure in the UK using discriminant analysis and financial ratio data. *Journal of the Royal Statistical Society*. Series A. 342-358.

Tam, K. Y., & M. Y. Kiang (1992). Managerial applications of neural networks: the case of bank failure predictions. *Management science*. 38:7. 926-947.

Tilastokeskus (2012) Konkursseja eniten rakentamisen ja muiden palveluiden päätoimialoilla . [verkkojulkaisu]. Helsinki: Tilastokeskus [viitattu: 30.10.2018]. Saatavana World Wide Webistä: <URL:http://www.stat.fi/til/konk/2012/12/konk_2012_12_2013-02-08_kat_001_fi.html> ISSN=1798-4424.

Tilastokeskus (2016). Konkurssiin haetut yritykset ja niiden henkilökunnan määrä maakunnan mukaan tammi-joulukuussa 2016 ja 2015. [verkkojulkaisu]. Helsinki: Tilastokeskus [viitattu: 11.11.2018]. Saatavana World Wide Webistä: <URL:http://www.stat.fi/til/konk/2016/12/konk_2016_12_2017-01-25_tau_001_fi.html> ISSN=1798- 4424

Tilastokeskus (2017). Konkurssiin haetut yritykset ja niiden henkilökunnan määrä maakunnan mukaan tammi-.syyskuussa 2018 ja 2017. [verkkojulkaisu] Saatavissa: https://www.stat.fi/til/konk/2018/09/konk_2018_09_2018-10-17_tau_001_fi.html. [Viitattu: 26.10.2018.]

Tilastokeskus (2019) Yritysten rakenne- ja tilinpäätöstilasto. [verkkojulkaisu]. Helsinki: Tilastokeskus [viitattu: 11.2.2019]. Saatavana World Wide Webistä: <URL:http://www.stat.fi/til/yrti/index.html> ISSN=2342-6217

Tsai, C. F. (2009). Feature selection in bankruptcy prediction. *Knowledge-Based Systems*, 22:2. 120-127.

Wu, X., V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinback, D. J. Hand & D. Steinberg (2008). Top 10 algorithms in data mining. *Knowledge and information systems*. 14:1. 1-37.

Yeh, C. C., D. J. Chi & Y. R. Lin (2014). Going-concern prediction using hybrid random forests and rough set approach. *Information Sciences*. 254. 98-110.

Yip, A. Y. (2004). Predicting business failure with a case-based reasoning approach. *In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems.* 665-671.

Yritystutkimus (2011). Yritystutkimuksen tilinpäätösanalyysi. *Gaudeamus Helsinki University Press 2011. 105 s.* ISBN *978-952-495-204-0.*

Yu, Q., A. Lendasse & E. Séverin (2009). Ensemble KNNs for bankruptcy prediction. In Int. *Conf. on Computing in Economics and Finance*. 1-9.

Zhao, D., C. Huang, Y. Wei, F. Yu, M. Wang & H. Chen (2017). An effective computational model for bankruptcy prediction using kernel extreme learning machine approach. *Computational Economics*. 49:2. 325-341.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting research.* 59-82.