Prediktiivinen analyysimenetelmä tilan kannattavuuden laskemiseksi Taloustohtorissa

Maria Yli-Heikkilä[1)], Jukka Tauriainen[2)], Mika Sulkava[3)]

[1)] *Luonnonvarakeskus, Tilastopalvelut, Tietotie 4, 31600 Jokioinen, Maria.Yli-Heikkila@luke.fi*

[2)] *Luonnonvarakeskus, Tilastopalvelut, Kampusranta 9 C, 60320 Seinäjoki, Jukka.Tauriainen@luke.fi*

[3)] *Luonnonvarakeskus, Tilastopalvelut, PL 2, 00791 Helsinki, Mika.Sulkava@luke.fi*

Luonnonvarakeskuksen (Luke) Taloustohtori-sivusto (www.luke.fi/taloustohtori) tarjoaa Suomen ja osin myös muiden EU-maiden biotalouden toimijoita koskevia tietoja yritystaloudesta, rakennekehityksestä ja ympäristökestävyydestä. Taloustohtorin Maa- ja puutarhatalous -verkkopalvelussa julkaistaan maatalouden kannattavuuskirjanpitoaineiston pohjalta lasketut maatilayritysten taloudellista asemaa ja kehitystä kuvaavat tunnusluvut alueittain, kokoluokittain ja tuotantosuunnittain ryhmä-keskiarvoina. Verkkopalvelun testausvaiheessa on toiminto, jolla maatilayrittäjä voi vertaiskehittää tuotantosuunnitelmaansa rinnastamalla oman tilansa tietoja vastaavan tilaryhmän tietoihin. Yrityksen tunnuslukujen laskemiseen tarvitaan tiedot kahdenkertaisesta kirjanpidosta. Tavanomainen yksityinen maatalouden harjoittaja ei ole kirjanpitovelvollinen. Tällöin maatalouden verotettava tulo lasketaan maksuperusteen mukaan, joten parhaan kuvan maatilayrityksen taloudellisesta tilasta saa verolomakkeelta. Tunnuslukujen laskemiseksi tarvittaisiin lisäksi tuote- ja panosvarastojen määrä- ja arvotiedot, käyttöomaisuuden arvo ja työtuntien määrät. Laskentatoimen näkökulmasta maksuperusteista kirjanpidosta saadut lähtöarvot ovat puutteelliset, joten kannattavuuden, maksuvalmiuden tai vakavaraisuuden tunnuslukuja ei voida niistä laskea. Olemme kehittäneet tekoälyyn pohjaavan sovelluksen, joka prediktiivistä analyysimenelmää käyttäen laskee tuotanto-, työtunti- ja verotietojen pohjalta yrityksen kannattavuuskertoimen. Kullekin tuotantosuuntatyypille on kehitetty oma mallinsa. Päätöspuumalleihin lukeutuva koneoppimisen menetelmä satunnaismetsä tuottaa suurimmalle osalle tuotantosuunnista tarkimman ennustemallin. Kannattavuuskertoimesta voidaan lisäksi johtaa muita yrittäjän kannattavuutta kuvaavia tunnuslukuja, kuten yrittäjänvoitto ja työn tuotto. Käyttäjä voi myös kokeilla, miten eri lähtötietojen muutos vaikuttaa kannattavuuteen. Siten sovellus tuo kaikille maa- ja puutarhatalouden yrityksille päätöksenteon tueksi mahdollisuuden arvioida kannattavuutta eri tuotantoskenaarioilla.

Asiasanat: kannattavuus, koneoppiminen, maatalous, maatila, satunnaismetsä

Johdanto

Luonnonvarakeskuksen (Luke) Taloustohtori-sivusto (www.luke.fi/taloustohtori) tarjoaa Suomen ja osin myös muiden EU-maiden biotalouden toimijoita koskevia tietoja yritystaloudesta, rakennekehityksestä ja ympäristökestävyydestä. Taloustohtorin Maa- ja puutarhatalous -verkkopalvelussa julkaistaan maatalouden kannattavuuskirjanpitoaineiston pohjalta lasketut maatilayritysten taloudellista asemaa ja kehitystä kuvaavat tunnusluvut alueittain, kokoluokittain ja tuotantosuunnittain ryhmäkeskiarvoina. Verkkopalvelun testausvaiheessa on toiminto, jolla maatilayrittäjä voi vertaiskehittää tuotantosuunnitelmaansa rinnastamalla oman tilansa tietoja vastaavan tilaryhmän tietoihin.

Yrityksen tunnuslukujen laskemiseen tarvitaan tiedot kahdenkertaisesta kirjanpidosta. Tavanomainen yksityinen maatalouden harjoittaja ei ole kirjanpitovelvollinen. Tällöin maatalouden verotettava tulo lasketaan maksuperusteen mukaan, joten parhaan kuvan maatilayrityksen taloudellisesta tilasta saa verolomakkeelta. Tunnuslukujen laskemiseksi tarvittaisiin lisäksi tuote- ja panosvarastojen määrä- ja arvotiedot, käyttöomaisuuden arvo ja työtuntien määrät. Laskentatoimen näkökulmasta maksuperusteista kirjanpidosta saadut lähtöarvot ovat puutteelliset, eikä kannattavuuden, maksuvalmiuden tai vakavaraisuuden tunnuslukuja voida niistä laskea.

Maatilatalouden kehittämisrahaston (Makera) rahoittamassa hankkeessa *Suomen maatalouden kannattavuus ja kilpailukyky (2013–2017)* olemme kehittäneet tekoälyyn pohjaavan sovelluksen, joka prediktiivistä analyysimenetelmää käyttäen laskee tuotanto-, työtunti- ja verotietojen pohjalta yrityksen kannattavuuskertoimen. Ennustettavasta kannattavuuskertoimesta voidaan lisäksi johtaa muita yrittäjän kannattavuutta kuvaavia tunnuslukuja, kuten yrittäjänvoitto ja työn tuotto. Taloustohtorissa käyttäjä voi myös kokeilla, miten eri lähtötietojen muutos vaikuttaa kannattavuuteen. Siten sovellus tuo kaikille maa- ja puutarhatalouden yrityksille päätöksenteon tueksi mahdollisuuden arvioida kannattavuutta eri tuotantoskenaarioilla.

Materiaali ja menetelmät

Tutkimuksessa hyödynnettiin kannattavuuskirjanpitoaineistoa vuosilta 2000–2015. Euroopan Unionin maatilojen tuloja ja taloutta koskevien kirjanpitotietojen yhdenmukaistamista varten kehitettyä standardituotosluokittelualgoritmia (engl. *Standard Output*, lyh. SO) käyttäen aineisto on luokiteltu 32:een tuotantosuuntaan. Aineistossa on eniten lypsykarjatiloja ($\sim$5400), sekä vilja-, öljy- ja valkuaiskasvien viljelytiloja ($\sim$2400).Yhteensä tiloja on aineistossa $\sim$14000. Esikäsittelyssä 1000:n muuttujan aineistosta valittiin vain maksuperusteiseen kirjanpitoon ja tuotantoon liittyvät muuttujat, sekä työtuntimäärät, yhteensä $\sim$300 muuttujaa.

Sovellusta varten pyrittiin löytämään malli, joka ennustaa vastemuuttujaa eli kannattavuuskerrointa tarkimmin. Vastemuuttujan ja selittävien muuttujien välisiä yhteyksiä ei tässä tarkasteltu. Kannattavuuskertoimen ennustamiseksi testattiin erilaisia koneoppimisen menetelmiä (ks. tarkemmin laajemmasta menetelmävertailusta lypsykarjatila-aineistolla Yli-Heikkilä ym. 2015). Aineisto jaettiin opetusaineistoon $(3/4)$ ja testiaineistoon $(1/4)$. Osa opetusaineistosta käytettiin esimerkkeinä mallin opettamisessa ja osa validointi- eli vahvistusesimerkkijoukkona, jonka avulla mallin toimivuutta ohjattiin tarkempaan suuntaan ja mahdollisia hyperparametreja säädettiin. Mallin tarkkuus testiaineistolla kertoo mallin yleistämiskyvystä, eli kuinka tarkkoja ovat mallin ennusteet todellisessa käyttötilanteessa. Taaksepäin askeltavassa muuttujien valinnassa ja mallin valinnassa käytettiin 5×10 ristiinvalidointimenetelmää. Mallin ennustuskykyä eli tarkkuutta mitattiin residuaalien keskineliövirheen neliöjuurella (RMSE). Mitä pienempi RMSE, sitä tarkempi ennuste.

Taulukossa 1 esitetään neljän eri menetelmän tulokset. Yksinkertaisimpana referenssimallina on yleistetty lineaarinen regressiomalli (engl. *generalized linear model*, lyh. GLM, Nelder ja Wedderburn 1972). Lasso (Tibshirani 1996) on niin ikään yleistetty lineaarinen malli, jossa käytetään ℓ_1-säännöstelyä rajoittamaan käytettävien selittäjien vaikutusta. *Rprop+* on vastavirta-algoritmilla (engl. *resilient backpropagation with weight backtracking*, Riedmiller 1994) opetettu neuroverkko, jossa on kolme neuronia yhdellä piilokerroksella. Päätöspuumalleihin lukeutuva satunnaismetsä (engl. *Random Forest*, Breiman 2001) perustuu bootstrap-aggregointiin ja ennustajamuuttujien satunnaisotokseen solmukohdissa. TensorFlow on kahden piilokerroksen neuroverkko, jossa on käytetty TensonFlow-ohjelmistokirjastoa

(Abadi ym. 2015). Yleismalli on koko aineistolla (kaikilla tuotantosuunnilla) opetettu malli.
Satunnaismetsä näytti tuottavan suurimmalle osalle tuotantosuunnista tarkimman ennustemallin. Sovellusta ajatellen satunnaismetsä on algoritmisesti tehokas ja suhteellisen vähän muistia käyttävä, joten jatkoimme ennustemallien kehittämistä satunnaismetsällä. Satunnaismetsämallista saa myös laskettua yksittäisen ennusteen (piste-estimaatin) luottamusvälin.
Sovelluksen kannalta yksinkertaisinta olisi käyttää mallia, joka toimisi kaikilla tuotantosuunnilla. Seuraavaksi kyseenalaistimme SO-typologialuokittelun. Sovelsimme akateemikko Teuvo Kohosen 1980-luvulla kehittämää neuroverkkoihin lukeutuvaa menetelmää itseorganisoituva kartta (engl. *self-organizing map*, lyh. SOM, Kohonen 2000), jolloin tavoitteena oli löytää uusia yhteyksiä tilojen välillä liittyen niiden tuotantorakenteisiin ja liiketoimintaan. SOM:iin perustuvalla klusterointimenetelmällä (Vesanto ja Sulkava 2002) saatu uusi luokkamuuttuja lisättiin yleismalliin selittäjäksi.
Kokeilimme myös ohjata mallin opetusta siten, että opetusjoukossa suurimpaan ryhmään eli pääryhmään kuuluvat tilat valitaan validointiaineistoon. Tällöin mallinvalinnassa malli säätyy toimimaan paremmin suurimmalle ryhmälle. Ajatuksena on, että pääryhmän ulkopuolelle jäävät ovat poikkeavia yksittäisiä tapauksia, mutta ne pidetään kuitenkin mukana mallin opetuksessa. Pääryhmään valittiin ensin havainnot, joiden havaintoresiduaali aiemmin opetetun mallin tuloksissa kuuluu residuaalijakauman välille 15...85%, eli ennustevirhe on ollut näillä maltillinen. Toisessa menetelmässä pääryhmään kuuluivat tilat, jotka olivat luokittuneet frekvenssiltään suurimpaan SOM-klusteriin. Taulukko 2 esittää, miten neljällä tuotantosuunnalla tilat ovat luokittuneet kahdeksaan eri SOM-klusteriin. Kullakin tuotantosuunnalla on yksi klusteri, johon suurin osa havainnoista selvästi sijoittuu.
Em. validointimenetelmät voidaan toteuttaa ristiinvalidoinnissa, joka on satunnaismetsän ulkopuolinen mallinvalintaprosessi. On huomattavaa, että satunnaismetsä-algoritmissa on itsessään sisäinen validointimenettely. Algoritmi ottaa syöteaineistosta bootstrap-otoksen $(2/3)$ yksittäisen päätöspuun rakentamiseksi. Ulkopuolelle jäävää osaa (engl. *out-of-bag* -aineisto) käytetään päätöspuun validointiaineistona. Alkuperäisessä random forest -ohjelmistokirjastossa on mahdollista vaikuttaa otantaan käyttämällä (painotettua) osittaista otantaa, mutta vain luokittelutehtävässä. Vaikka opetusta ei pystytty ohjaamaan regressiotehtävässä itse opetusalgoritmin sisällä, saadaan ristiinvalidoinnilla kuitenkin jonkin verran vaikutettua mallinvalintaan.

Tulokset

Kuvassa 1 nähdään, miten satunnaismetsämalli-menetelmällä tuotantosuuntakohtaisesti opetettu *SO* ja tuotantosuuntakohtaisesti testattu *Yleismalli* vertautuvat tarkkuudeltaan todellisessa käyttötilanteessa eli testiaineistolla. Useimmilla tuotantosuunnilla *Yleismalli* näyttääkin toimivan paremmin kuin tuotantosuuntakohtaisesti opetettu malli. Punainen katkoviiva lypsykarjatilojen *SO*-mallin tarkkuuden kohdalla toimii parhaan mallin referenssinä muita tuotantosuuntia vertailtaessa.
Kuvassa 2 on tarkemmassa tarkastelussa neljä tuotantosuuntaa. Mukana on tuotantosuuntakohtaisen perusmallin *SO* lisäksi maltillisten residuaalien validointiaineistolla säädetty *Valid*-malli, sekä suurimmalla SOM-klusterin validointiaineistolla säädetty *Valid SOM*. Kuvassa on myös *Yleismalli* ja tämän SOM-muuttujalla laajennettu malli. Porsastuotantotiloilla testattujen tuotantosuuntakohtaisten mallien osalta näemme, että *Valid* toimii huonommin kuin *SO*, mutta *Valid SOM* parantaa selvästi tarkkuutta. Kuitenkin yleismallit toimivat huomattavasti paremmin kuin tuotantosuuntakohtaiset mallit. SOM-muuttujan lisääminen ei tässä tuo parannusta *Yleismalliin*. Lypsykarjatiloilla kaikki mallit toimivat hyvin, paitsi *Valid SOM*, jonka tarkkuus on huomattavasti huonompi. Lihasiankasvatustiloilla kaikki mallien kehitysideat, *Valid*, *Valid SOM* ja *Yleis SOM* parantavat perusmalleja *SO* ja *Yleismalli*. Parhain tarkkuus saavutetaan *Valid SOM*:lla. Emolehmätuotantotiloilla tuotantosuuntakohtaiset mallit toimivat huonommin kuin yleismallit. Validointiaineistolla säädetyt *Valid* ja *Valid SOM* toimivat jopa selvästi huonommin kuin *SO*. Yleismalleista *Yleis* + *SOM* toimii hieman tarkemmin kuin *Yleismalli*.
Edellä olevat tulokset kertovat mallien ennustekyvystä todellisessa käyttötilanteessa testiaineistolla arvioituna. Koska testiaineisto on kuitenkin rajallinen, usein liiankin, on syytä tarkastella vielä opetusvaiheen hajontaa saadaksemme viitteitä opetusalgoritmin yleistämiskykyyn liittyvistä ongelmista. Taulukossa 3 on koottuna tietoja opetusaineiston koosta kullakin tuotantosuunnalla ja *Yleismallilla*. Opetuksessa ai-

Taulukko 1. Tuotantosuuntakohtaisesti opetettujen prediktiivisten mallien tarkkuudet eri menetelmillä keskineliövirheen neliöjuurella (RMSE) mitattuna testiaineistolla. Satunnaismetsällä opetettu malli on tarkin useimmissa tuotantosuunnissa. Lypsykarjatiloilla mallin tarkkuus on huomattavasti parempi, kuin esimerkiksi vilja-, öljy- ja valkuaiskasvien viljelytiloilla. Yleismalli on koko aineistolla opetettu malli.

	GLM	Lasso	*RPROP+*	Satunnaismetsä	TensorFlow
Emolehmätuotanto	0.86	0.80	0.66	0.54	0.83
Lypsykarjatilat	0.49	0.49	0.26	0.27	0.61
Muidenkasvienviljely	0.99	0.99	0.84	0.74	10.26
Vilja-, öljy- ja valkuaiskasvit	0.76	0.76	0.61	0.54	0.85
Yleismalli	0.82	0.82	0.62	0.56	0.62

Taulukko 2. Prosentuaalinen frekvenssijakauma näyttää, kuinka tilat kuuluvat SOM-menetelmällä laskettuihin kahdeksaan klusteriin tuotantosuuntakohtaisesti.

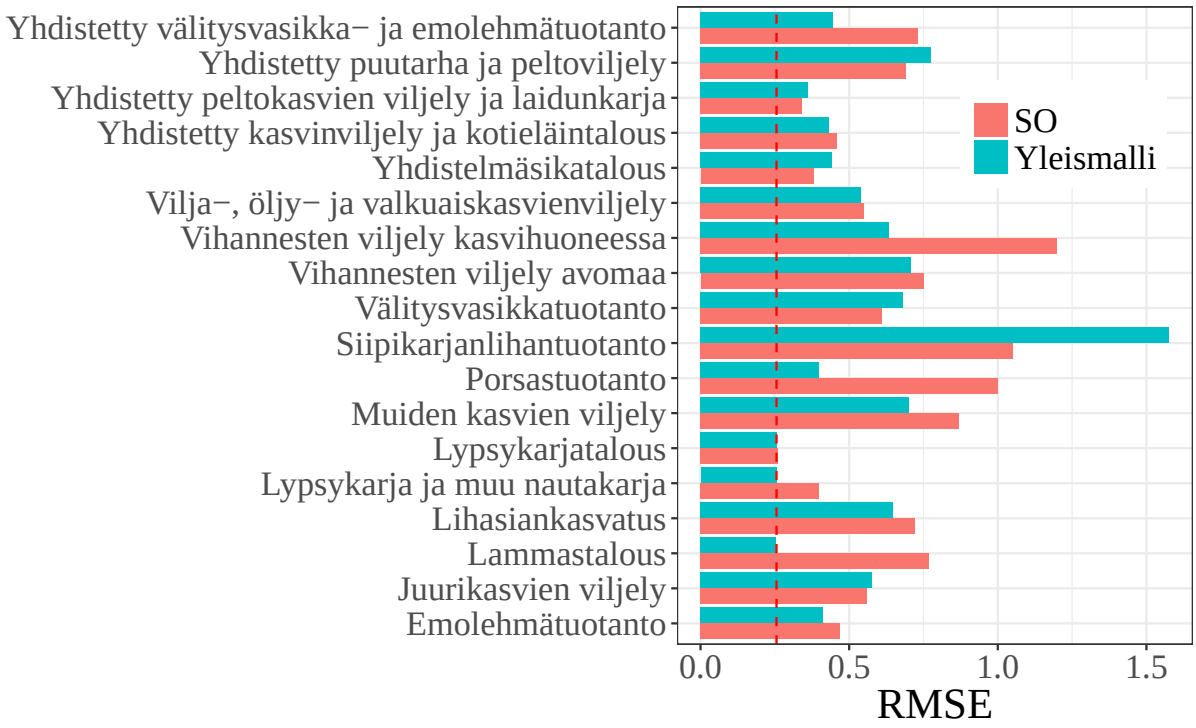
	1	2	3	4	5	6	7	8
Emolehmätuotanto	2	0	0	3	91	3	1	0
Lihasiankasvatus	77	0	22	1	0	0	0	0
Lypsykarjatalous	0	0	0	69	0	31	0	0
Porsastuotanto	12	0	83	1	0	4	1	0
Vilja-, öljy- ja valkuaiskasvienviljely	1	10	74	1	0	0	14	0

Taulukko 3. Eri mallinvalintatavoilla opetettujen mallien tuloksia opetusvaiheessa. Keskimääräinen keskihajonta (SD) ristiinvalidoinnissa (CV) kertoo, paljonko on vaihtelua mallin tarkkuudessa ristiinvalidointimenettelyn osioiden välillä. Satunnaismetsän opetus toistettiin viisi kertaa. Taulukossa mallien välinen keskimääräinen keskineliövirheen neliöjuuri (RMSE) ja tämän keskihajonta (SD).

	Mallinvalinta	Tiloja opetuksessa	Keskim. SD CV:n osioissa	Toistettujen mallien välinen keskim. RMSE $\pm$ SD
Emolehmätuotanto	SO	365	0.15	0.54$\pm$0.01
	Valid SOM		0.00	0.51$\pm$0.05
	Valid		0.00	0.51$\pm$0.03
Lihasiankasvatus	SO	204	0.12	0.63$\pm$0.03
	Valid SOM		0.00	0.63$\pm$0.02
	Valid		0.00	0.67$\pm$0.04
Lypsykarjatalous	SO	4066	0.02	0.27$\pm$0.00
	Valid SOM		0.00	0.27$\pm$0.00
	Valid		0.00	0.27$\pm$0.00
Porsastuotanto	SO	256	0.42	0.70$\pm$0.17
	Valid SOM		0.00	0.67$\pm$0.12
	Valid		0.00	0.64$\pm$0.13
Vilja-, öljy- ja valkuaiskasvit	SO	1846	0.08	0.54$\pm$0.01
	Valid SOM		0.00	0.55$\pm$0.02
	Valid		0.00	0.56$\pm$0.01
Kaikki tuotantosuunnat	Yleismalli	10464	0.13	0.64$\pm$0.07

neisto jaetaan ristiivalidointimenetelmällä viiteen osioon, joissa jätetään $1/4$ validointiaineistoksi. Kukin osio opetetaan 10 kertaa ja validointiaineiston ennusteista lasketaan mallin tarkkuus RMSE. Taulukossa 3 ristiinvalindoinnissa osioiden sisäinen keskimääräinen keskihajonta (SD) eri mallinnustavoilla osoittaa, että opetusalgoritmin vaihtelusta johtuva virhe on hyvin maltillinen tarpeeksi suurilla aineistoilla. *Valid* ja *Valid SOM* -malleilla keskihajonta pysyy hyvin pienenä. Tämä oli odotettavaa, sillä validointiaineistoa oli säädetty sisältämään samanlaisia tapauksia.

Ristiinvalidointia toistetaan viisi kertaa, jolloin aineiston jako opetus- ja testidataan muodostetaan uudelleen joka kerta. Näin saadaan kuva, miten otanta vaikuttaa mallin ennustuskykyyn. Opetus- ja testiaineiston jaossa käytettiin osittaisen otannan menetelmää, jotta havainnot jakautuisivat molempiin aineistoihin



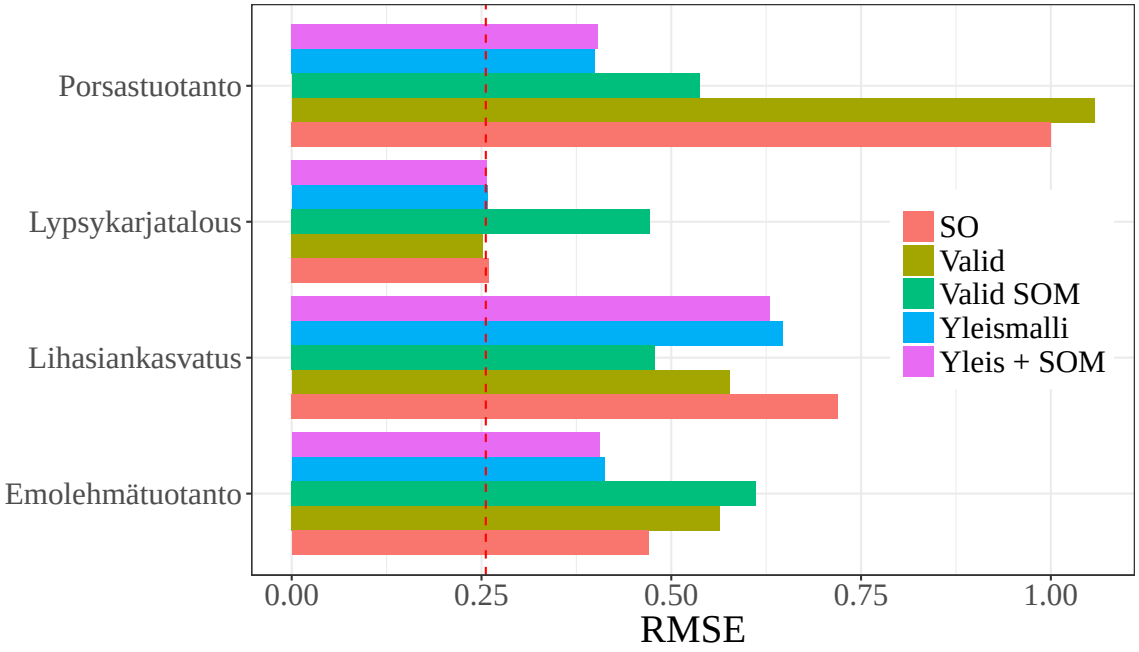
Kuva 1. Satunnaismetsä-menetelmällä opetettujen prediktiivisten mallien tarkkuus keskineliövirheen neliöjuurella (RMSE) mitattuna testiaineistolla. *SO* on malli, joka on opetettu tuotantosuuntaan kuuluvista tiloista. *Yleismalli* on opetettu koko kirjanpitotila-aineistolla, testidatassa on vain tuotantosuuntakohtaisia tiloja. Punainen katkoviiva toimii referenssinä parhaan tuotantosuuntakohtaisen mallin tarkkuudesta (lypsykarjatiloilla).

ositemuuttujan jakauman mukaisesti. *Valid SOM* aineisto jaettiin klusteri-muuttujan mukaan, muilla ositemuuttujana käytettiin vuosi-muuttujaa. Taulukosta 3 näemme, että toistettujen opetusten välillä mallin keskimääräinen tarkkuus huononee ja samalla sen keskihajonta kasvaa varsinkin, kun otoskoko on pieni. Esimerkiksi porsastuotannon mallissa nähdään, että mallin tarkkuus riippuu paljolti, millainen otos on opetukseen päätynyt. Tuotantosuunnilla, joilla vaihtelu on pientä, malli itsessään on tasapainoinen. Mallin tarkkuudessa eli harhaisuudessa on parantamisen varaa. Yllättäen *Yleismalli* kärsii suuresta vaihtelusta. Tämä viittaa aineistossa esiintyvään suureen vaihteluun joillakin tuotantosuunnilla. Myös opetusaineiston vähäinen määrä esimerkiksi porsastuotantotiloilla tuottaa harhaisuusongelmaa. Tulokset myös osoittavat, että validointimallien ennustuskyky ei vaihtele juurikaan enemmän kuin *SO*-mallilla. Tarkkuuskin on näillä toisinaan parempi. Tämä tulos kannustaa kokeilemaan validointimallien toimivuutta toisella opetusalgoritmilla, jonka validointiotantaa voidaan suoraan ohjata.

Tulosten tarkastelu

RMSE kertoo mallin keskimääräisen virheen vastemuuttujan asteikolla. Jos mallin RMSE on esimerkiksi 0.25, uuden tapauksen arvio kannattavuuskertoimesta saattaa heittää keskimäärin 0.25 yksikköä ylös- tai alaspäin. Esimerkiksi, jos kannattavuuskerroin on 1.00, viljelijä saa vuonna 2017 työtunnilleen 15.90€:n korvauksen tunnilta ja 4.35 prosentin korkotuoton omalle pääomalle. 0.25 yksikön heitto kannattavuuskertoimessa vastaa n. neljän euron virhettä tuntipalkkaan. Keskimääräisellä tilalla se tarkoittaisi vuonna 2017 yli 3000 euron virhettä yrittäjätuloennusteessa (13 000€±3 000€). Parhaimmillaan satunnaismetsä on saavuttanut tämän tarkkuuden vain lypsykarjatila-aineistolla.

Mallien tarkkuuden parantamiseksi olisi hyvä saada lisää aineistoa tai uusia, tietoa tiivistäviä muuttujia. Satunnaismetsän rajoitus on, että opetusaineiston ja testiaineiston muuttujat tulisivat olla samat ja suunnilleen samasta jakaumasta. Jatkossa pitäisi keskittyä hyödyntämään kirjanpitoaineiston tilinpäätös- ja tunnuslukumuuttujia, jotka ovat toistaiseksi jätetty opetuksen ulkopuolelle.



Kuva 2. Satunnaismetsä-menetelmällä opetettujen prediktiivisten mallien tarkkuus keskineliövirheen neliöjuurella (RMSE) mitattuna testiaineistolla. *SO* on malli, joka on opetettu tuotantosuuntaan kuuluvista tiloista. Tuotantosuuntakohtaisessa *Valid*-mallissa validointiaineistoon on valikoitu tiloja, jotka kuuluvat residuaalijakauman välille 15–85%. Tuotantosuuntakohtaisessa *Valid SOM* -mallissa validointiaineistoon on valikoitu tiloja, jotka ovat luokittuneet frekvenssiltään suurimpaan klusteriin. *Yleismalli* on opetettu koko kirjanpitotila-aineistolla. *Yleis + SOM* on yleismalli, johon on lisätty SOM-luokkamuuttuja. Yleismallien testidatassa on vain ko. tuotantosuuntaan kuuluvia tiloja. Punainen katkoviiva toimii referenssinä parhaan tuotantosuuntakohtaisen mallin tarkkuudesta (lypsykarjatiloilla).

Johtopäätökset

Kullakin tuotantosuuntatyypillä on testattu tuotantosuuntakohtaisesti opetettuja prediktiivisiä malleja, sekä koko aineistoa hyödyntävää *Yleismallia*. Päätöspuumalleihin lukeutuva koneoppimisen menetelmä satunnaismetsä tuottaa suurimmalle osalle tuotantosuunnista tarkimman ennustemallin. Satunnaismetsää on mallinvalintaan liittyen kehitetty vahvistusesimerkkijoukkoa valikoimalla, sekä lisäämällä SOM-muuttuja selittäjäksi yleismalliin. Johtopäätöksenä voidaan todeta, että mallin valinta on tehtävä tapauskohtaisesti. Mikään testatuista menetelmistä ei osoittautunut ylivertaiseksi. Sovellusta ajatellen yksinkertaisinta olisi valita *Yleismalli*, joka osoittautuikin useimmissa tapauksissa ennustuskyvyltään parhaimmaksi. Jatkossa kannattaisikin keskittyä kehittämään yleismallia esimerkiksi hyödyntämällä opetusaineistosta pois jätettyjä tilinpäätös- ja tunnuslukumuuttujia.

Kiitokset

Kiitokset aineistosta kannattavuuskirjanpitotiloille, hanketta rahoittaneelle Maatilatalouden kehittämisrahastolle, sekä R- ja Python-ohjelmistokirjastojen kehittäjille.

Kirjallisuus

Breiman, L. 2001. Random forests. *Machine learning* 45:5–32.

Kohonen, T. 2001. Self-Organizing Maps, 3rd Edition, Vol. 30 of Springer Series in Information Sciences.

Nelder, J.A. & Wedderburn, R.W.M. 1972. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)* 135:370–384.

Riedmiller, M. 1994. Advanced supervised learning in multi-layer perceptrons — From backpropagation to adaptive learning algorithms. *Computer Standards & Interfaces* 16:265–278.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Methodological* 58:267–288.
Vesanto, J. & Sulkava, M. 2002. Distance matrix based clustering of the self-organizing map. Teoksessa: Dorronsoro, J.R. (toim.). *Artificial Neural Networks — ICANN 2002 International Conference*. Vol. 2415 of Lecture Notes in Computer Science. Springer-Verlag. p. 951–956.
Yli-Heikkilä, M., Tauriainen, J. & Sulkava, M. 2015. Predicting the profitability of agricultural enterprises in dairy farming. Teoksessa: Verleysen, M. (toim.). *ESANN 2015, 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Univ. Cath. de Louvain and KULeuven, p. 155–160.

Käytetyt ohjelmistot:

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jozefowicz, R., Jia, Y., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Schuster, M., Monga, R., Moore, S., Murray, D., Olah, C., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, R., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., & Zheng, X. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.
Friedman, J., Trevor Hastie, T., Tibshirani, R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1–22. URL: http://www.jstatsoft.org/v33/i01/. (lasso R kirjastossa glmnet.)
Fritsch, S. ja Guenther, F. 2016. neuralnet: Training of Neural Networks. R package version 1.33.
Kuhn, M. 2017. caret: Classification and Regression Training. Kontribuutioineet: Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. ja Hunt, T. 2017. R package version 6.0–76. h
Liaw, A. ja Wiener, M. 2002. Classification and Regression by randomForest. R News 2(3), 18–22.
McKinney, W. 2010. Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51–56. (Python kirjasto pandas)
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É 2011. Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830. (Python kirjasto scikit-learn)
Python Software Foundation, Python versio 3.6.3, https://www.python.org/
R Core Team 2017. R: A language and environment for statistical computing. R versio 3.4.3. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.
van der Walt, S., Colbert C.S. ja Varoquaux, G. 2011. The NumPy Array: A Structure for Efficient Numerical Computation, Computing in Science & Engineering, 13, 22–30, DOI:10.1109/MCSE.2011.37. (Python kirjasto numpy)