Lineaarisen regression, Random Forestin ja XGBoostin vertailu PISA-aineistolla

LuK Eetu Tammi

Pro gradu -tutkielma
Toukokuu 2026

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

Tarkastajat:
Prof. Ion Petre
Prof. Henri Nyberg

Turun yliopiston laatujärjestelmän mukaisesti tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck-järjestelmällä

TURUN YLIOPISTO, Matematiikan ja tilastotieteen laitos

Pro gradu -tutkielma
Pääaine: Tilastotiede
Tekijä: Eetu Tammi
Otsikko: Lineaarisen regression, Random Forestin ja XGBoostin vertailu PISA-aineistolla
Ohjaaja: Prof. Ion Petre
Sivumäärä: 51 sivua + liitteet 7 sivua
Aika: Toukokuu 2026

Tässä pro gradu -tutkielmassa tarkastellaan oppimistuloksiin yhteydessä olevia tekijöitä PISA 2022 -aineiston avulla sekä vertaillaan kolmen mallinnusmenetelmän, Elastic Netin, Random Forestin ja XGBoostin, suorituskykyä. Tavoitteena on selvittää keskeiset taustatekijät matematiikan, lukutaidon ja luonnontieteiden osaamisessa sekä arvioida mallien tuottamien tulosten yhteneväisyyttä. Lisäksi tarkastellaan tulosten eroja Suomen, Ruotsin ja Yhdysvaltojen välillä ja sitä, missä määrin havaitut ilmiöt ovat samankaltaisia eri konteksteissa.

Mallien suorituskykyä arvioitiin sisäkkäisellä ristiinvalidoinnilla käyttäen useaa satunnaissiementä sekä PISA-tuloksia kuvaavaa PV-arvoa, jotta tulokset olisivat mahdollisimman vakaita. Tulosten perusteella Elastic Net suoriutui parhaiten testiaineistossa, kun taas puupohjaisissa malleissa havaittiin ylisovittamista, erityisesti Random Forest -mallissa. XGBoost tuotti kuitenkin kilpailukykyisiä tuloksia, vaikka sen yleistettävyys jäi hieman heikommaksi kuin Elastic Netillä.

Keskeisiksi oppimistuloksiin liittyviksi tekijöiksi kaikilla osa-alueilla nousivat erityisesti motivaatio ja vaivannäkö sekä sosioekonomista taustaa ja oppimisympäristöä kuvaavat muuttujat. Mallien välillä havaittiin eroja siinä, miten nämä tekijät painottuvat, mikä viittaa siihen, että eri menetelmät tunnistavat aineistosta osin erilaisia rakenteita. Tulokset korostavat mallivalinnan merkitystä ja sitä, että oppimistuloksia tulee tulkita yhteyksinä eikä kausaalisina vaikutuksina.

Asiasanat: PISA, koneoppiminen, Elastic Net, Random Forest, XGBoost, oppimistulokset

Sisällys

Liitteet

A Muuttujien selitykset (Excel-tiedosto)

B Koodit (Github)

C Kuvat

1 Johdanto

Tarkastellessa oppimistuloksia ei riitä pelkkä keskiarvojen vertailu, vaan olennaista on ymmärtää, mitkä tekijät ovat yhteydessä oppilaiden osaamiseen. PISA-aineisto tarjoaa tähän hyvän lähtökohdan, sillä se sisältää osaamistulosten lisäksi runsaasti tietoa oppilaiden taustasta, asenteista ja oppimisympäristöstä.

Näiden tekijöiden yhteyttä oppimistuloksiin on usein tarkasteltu lineaaristen mallien avulla. Ne ovat selkeitä ja helposti tulkittavia, mutta perustuvat oletukseen suoraviivaisista yhteyksistä muuttujien välillä. Käytännössä yhteydet voivat kuitenkin olla monimutkaisempia, jolloin lineaarinen malli ei välttämättä kuvaa ilmiötä riittävän hyvin.

Tämän vuoksi on hyödyllistä hyödyntää enemmän koneoppimismenetelmiä, kuten Random Forestia ja XGBoostia, jotka mahdollistavat myös ei-lineaaristen yhteyksien löytämisen. Ne mahdollistavat joustavamman mallinnuksen, mutta samalla niiden tulkinta on vaikeampaa. Tästä syntyy tarve vertailla eri menetelmiä keskenään ja arvioida, tuottavatko ne saman kuvan ilmiöstä vai korostavatko ne eri asioita.

Tässä työssä tarkastellaan PISA 2022 -aineiston avulla oppilaiden matematiikan, lukutaidon ja luonnontieteiden osaamiseen liittyviä tekijöitä. Analyysissä käytetään kolmea eri menetelmää: Elastic Net -regularisoitua lineaarista regressiota sekä Random Forest- ja XGBoost -malleja. Tarkastelu perustuu ennustavaan näkökulmaan, jossa hyödynnetään laajaa joukkoa oppilaiden taustamuuttujia ja kyselyvastauksia.

Tässä työssä pyritään vastaamaan kahteen keskeiseen kysymykseen. Ensinnäkin tarkastellaan, mitkä käytetyistä ennustemalleista ennustavat PISA-pistemääriä parhaiten. Tätä arvioidaan vertailemalla mallien ennustetarkkuutta yhtenäisillä arviointikriteereillä. Toiseksi tarkastellaan, mitkä taustatekijät ovat keskeisimpiä ennusteiden kannalta eri malleissa. Tällöin huomio kohdistuu siihen, mitkä muuttujat nousevat toistuvasti tärkeiksi ja onko niiden merkitys yhtenevä eri menetelmien välillä.

Tavoitteena on saada kokonaiskuva siitä, miten eri mallinnusmenetelmät kuvaavat oppimistuloksiin liittyviä tekijöitä ja miten niiden tuottamat tulokset ovat keskenään yhteneviä.

Tutkielmassa on hyödynnetty avustavaa kielimallia (ChatGPT) kielenhuollon tukena sekä yksittäisten ohjelmointiongelmien ratkaisemisessa. Tutkielman tutkimuksellinen sisältö, analyysit ja johtopäätökset on tuotettu itsenäisesti.

2 PISA-tutkimus ja suomalaisnuorten oppimistulokset

2.1 PISA-tutkimuksen tavoitteet ja rakenne

Tässä alaluvussa käsitellään PISA-tutkimuksen tarkoitusta, rakennetta ja keskeisiä sisältöjä. Tarkastelu perustuu teokseen *PISA 2022 Results (Volume I): The State of Learning and Equity in Education* [1].

PISA-tutkimus (Programme for International Student Assessment) on OECD:n koordinoima maailmanlaajuinen oppimistulosten arviointitutkimus, joka mittaa 15-vuotiaiden nuorten osaamista matematiikassa, lukutaidossa ja luonnontieteissä. PISA-tutkimuksen päätavoitteena on tarjota tietoa koulutusjärjestelmien suorituskyvystä sekä tukea koulutuspolitiikkaa, joka edistää oppimistuloksia ja koulutuksen tasa-arvoa eri maissa. Tutkimuksen avulla voidaan arvioida, kuinka hyvin oppilaat pystyvät soveltamaan koulussa oppimaansa tietoa reaalimaailman tilanteisiin. PISA-tutkimus ei keskity pelkästään opetussuunnitelmien mukaisten oppisisältöjen hallintaan, vaan se mittaa laajemmin oppilaiden ajattelutaitoja ja ongelmanratkaisukykyä [1, s. 38–39].

PISA-tutkimus toteutetaan kolmen vuoden välein, ja jokaisella arviointikierroksella yksi kolmesta pääaineesta (matematiikka, lukutaito, luonnontieteet) on erityisessä painopisteessä. Vuoden 2021 tutkimus siirrettiin koronapandemian vuoksi vuodelle 2022 [1, s. 38]. Esimerkiksi vuonna 2022 pääpaino oli matematiikassa, kun taas lukutaito ja luonnontieteet arvioitiin suppeammin [1, s. 40]. Tutkimus koostuu kahdesta pääosiosta: kognitiivisista testeistä, jotka mittaavat oppilaiden osaamista ja kykyä soveltaa oppimaansa käytännön ongelmanratkaisutilanteisiin, sekä taustakyselyistä, joissa kerätään tietoa oppilaiden sosiaalisesta taustasta, oppimisympäristöstä, asenteista ja koulujärjestelmistä. Taustamuuttujien avulla voidaan tarkastella, miten erilaiset tekijät, kuten sukupuoli, sosioekonominen tausta ja koulun resurssit, vaikuttavat oppimistuloksiin.

PISA-tutkimuksen rakenne on suunniteltu varmistamaan, että tulokset ovat kansainvälisesti vertailukelpoisia. Mittausmenetelmät ja arviointikriteerit kehitetään huolellisesti, jotta voidaan tehdä luotettavia johtopäätöksiä koulutusjärjestelmien eroista ja kehityssuunnista. Tulokset tarjoavat kattavan kuvan siitä, kuinka hyvin oppilaat eri maissa selviytyvät tulevaisuuden työelämän ja yhteiskunnan vaatimuksista. PISA-tutkimuksen avulla voidaan myös seurata oppimistulosten muutoksia pitkällä aikavälillä ja tunnistaa rakenteellisia muutoksia koulutusjärjestelmissä sekä

oppilaiden osaamisessa.

Tässä tutkimuksessa analysoidaan suomalaisten nuorten matematiikan osaamisen muutoksia eri PISA-arviointikierroksilla. PISA-tutkimus on järjestetty vuosina 2000, 2003, 2006, 2009, 2012, 2015, 2018 ja 2022. Jokaisella arviointikierroksella on ollut eri pääainepainotus, mikä vaikuttaa tulosten tarkasteluun (taulukko 1). Esimerkiksi vuonna 2003 matematiikka oli pääpainopisteenä, mikä tekee siitä suoraan vertailukelpoisen vuoden 2012 ja 2022 tulosten kanssa [1, s. 40]. Tässä työssä keskitytään vuoden 2022 tuloksiin.

Vuosi	Pääkategoria
2000	Lukutaito
2003	Matematiikka
2006	Luonnontieteet
2009	Lukutaito
2012	Matematiikka
2015	Luonnontieteet
2018	Lukutaito
2022	Matematiikka

Taulukko 1: PISA-arviointivuodet ja niiden pääkategoriat

2.2 PISA-aineiston keruu ja käsittely

Tässä alaluvussa tarkastellaan, miten PISA-aineisto on kerätty ja käsitelty vuoden 2022 arvioinnissa. Tämä perustuu erityisesti teokseen *PISA 2022 Technical Report* [2] sekä raporttiin *PISA 2022 Results (Volume I)* [1].

Vuonna 2022 PISA-tutkimukseen osallistui arviolta 690 000 oppilasta 81:stä OECD:n jäsen- ja kumppanimaasta. He edustavat noin 29 miljoonaa 15-vuotiasta maailmanlaajuisesti. PISA-tutkimukseen osallistuvat oppilaat ovat iältään 15 v 3 kk–16 v 2 kk arviointihetkellä ja ovat suorittaneet vähintään kuusi vuotta perusopetusta. [1, s. 3, 40–41]

Aineisto kerättiin kaksivaiheisen otannan avulla. Ensin valittiin satunnaisesti kouluja kansallisesta koululuettelosta siten, että suuremmat koulut valikoituivat todennäköisemmin. Tätä kutsutaan otannaksi suhteessa koulun kokoon (probability proportional to size, PPS). Koulut ryhmiteltiin ennen valintaa esimerkiksi sijainnin tai koulutyypin mukaan, jotta eri alueet ja koulumuodot tulisivat varmasti eduste-

tuiksi. Toisessa vaiheessa valittiin oppilaat valituista kouluista. Sähköisessä arvioinnissa tavoitteena oli saada mukaan 42 oppilasta jokaisesta koulusta. Jos koulussa oli vähemmän 15-vuotiaita, kaikki otettiin mukaan. Jos oppilaita oli enemmän, heistä valittiin satunnaisesti 42. Tätä määrää kutsutaan tavoiteklusterikooksi. PISAn vähimmäisnäytestandardi on yleensä vähintään 150 koulua ja vähintään 6 300 arvioitua oppilasta (CBA-maat) tai 5 250 (PBA-maat). Jos maassa on alle 150 koulua, kaikki otetaan mukaan. CBA tarkoittaa tietokonepohjaista arviointia ja PBA paperipohjaista arviointia. [2, luku 6]

Koulutason ja koulun sisäisten poissulkujen yhteenlaskettu osuus tulee pitää alle 5 %:ssa kohdepopulaatiosta. Ohjeellinen raja koulutasolle on $< 0{,}5$ % ja koulun sisäisille poissuluille $< 2{,}5$ %. Lisäksi PISA 2022:ssa sallittiin poikkeuksellisesti poissulku oppilaille, jotka saivat kaiken opetuksen etänä eivätkä voineet osallistua koululla tehtävään arviointiin. [2, s. 105–106]

Koulu- ja oppilastason vastauskatoa korjataan painotuksilla, mutta laadun varmistamiseksi koulu, jonka oppilasvastausaste jäi alle 33 %:n, käsiteltiin PISA 2022:ssa ei-vastanneena ja sen oppilastiedot poistettiin analyysiaineistosta. [2, s. 107–108]

Oppilastason analyysit perustuvat lopullisiin opiskelijapainoihin, jotka koostuvat koulun ja oppilaan valintatodennäköisyyksistä, ei-vastauskorjauksista ja mahdollisesta painojen säädöstä. Otantavirhe arvioidaan tasapainotetulla toistoreplikaatiolla (BRR) ja sen Fay'n muunnelmalla. BRR:ssä muodostetaan useita tasapainoisia replikoita ja kullekin niistä lasketaan arvio. Varianssi saadaan replika-arvioiden vaihtelusta. Fay'n muunnelma pienentää painojen vaihtelua ja vakauttaa arviot, erityisesti pienissä alaryhmissä. PISA 2022:ssa replikat rakennetaan Hadamardin taulukon pohjalta, ja aineiston mukana toimitetaan 80 replikointipainoa. Tavoitteena on varmistaa, että tulokset edustavat koko maan 15-vuotiaita. [2, luku 10]

PISA-tutkimuksessa oppimistuloksia kuvaamaan käytetään plausible value -arvoja (PV). PV-arvot ovat oppilaskohtaisia satunnaisia pistelukuja, jotka kuvaavat osaamista mittausepävarmuus huomioiden. Niitä ei tulkita yksittäisen oppilaan "todellisina" pisteinä [2, s. 226]. PV-menetelmä liittyy siihen, että kaikki oppilaat eivät vastaa samoihin tehtäviin. PISA-tutkimuksessa kukin oppilas suorittaa vain osan kaikista mahdollisista tehtävistä. Tämän vuoksi yksittäisen oppilaan osaamistasoa ei voida mitata täydellisesti yhdellä pistemäärällä. Sen sijaan tilastollisen mallin avulla tuotetaan useita vaihtoehtoisia arvioita oppilaan osaamisesta. Yhdelle oppilaalle tuotetaan kullekin osa-alueelle 10 PV:tä, ja analyysit lasketaan aina kaikkien 10 PV:n yli [2, s. 226, 234]. Raportoinnissa PV-arvot on esitetty PISA-asteikolla,

joka on alun perin määritelty siten, että OECD-maiden keskiarvo on ollut 500 pistettä ja keskihajonta 100 pistettä osa-alueen ensimmäisellä mittauskerralla. Tämän jälkeen pistemäärät on raportoitu samalla asteikolla, minkä vuoksi OECD-keskiarvo voi myöhemmillä kierroksilla poiketa tästä viitearvosta. Pistemäärät sijoittuvat tyypillisesti noin välille 200–800 [2, s. 299]. Myös tässä analyysissä oppimistuloksia kuvataan PV-arvojen avulla.

Aineisto tarkistetaan ensin kansallisesti ja sen jälkeen kansainvälisesti (koodaus, validoinnit, tarkistukset ja yhdistäminen). Ennen julkaisua se käy läpi aineiston arviointi -prosessin, jossa maiden aineistot hyväksytään lopullisesti. Volume I -raportissa julkaistaan maakohtaiset kattavuus- ja otantaindikaattorit. [2, luku 12, luku 16][1, liite I.A2]

2.3 Suomen sijoitus PISA-tutkimuksessa eri vuosina

Suomi on ollut PISA-tutkimuksessa pitkään OECD-maiden kärkijoukossa, erityisesti 2000-luvun alkupuolella, jolloin Suomen osaamistaso herätti laajaa kansainvälistä huomiota. OECD:n maaraportoinnin perusteella Suomen tulokset ovat kuitenkin heikentyneet useilla peräkkäisillä PISA-kierroksilla. PISA 2022 -tuloksissa suomalaiset oppilaat saavuttivat keskimäärin OECD-keskiarvoa paremman tuloksen matematiikassa, lukutaidossa sekä luonnontieteissä, mutta samalla raportti korostaa pitkän aikavälin laskevaa kehitystä. Kokonaisuutena Suomen sijoitus on siirtynyt 2000-luvun alun aivan kärkipaikoilta kohti hyvää mutta aiempaa heikompaa tasoa kansainvälisessä vertailussa, vaikka Suomi säilyy edelleen OECD-maiden keskiarvon yläpuolella keskeisillä osa-alueilla. [3]

3 Aineisto

3.1 Käytetty PISA-aineisto

Tässä työssä käytetään OECD:n julkaisemaa PISA 2022 -tutkimuksen aineistoa [4]. PISA 2022 valittiin tarkasteltavaksi tutkimusvuodeksi, koska se on uusin saatavilla oleva PISA-kierros ja se tarjoaa ajantasaisimman kuvan 15-vuotiaiden oppilaiden osaamisesta sekä siihen liittyvistä taustatekijöistä.

OECD tarjoaa PISA 2022 -aineiston avoimena tutkimusaineistona useissa eri tiedostomuodoissa ja erillisinä tiedostoina eri kyselyille. Tässä työssä käytettiin SPSS-muotoista (`.SAV`) opiskelijakyselyaineistoa (Student Questionnaire Data File), joka sisältää oppilaskohtaisia taustamuuttujia, kyselyvastauksia sekä PISA-osaamista kuvaavat PV-arvot ja otospainot. Aineisto muodostaa lähtökohdan tässä tutkimuksessa toteutetuille ennustemalleille ja mahdollistaa oppilaiden taustatekijöiden ja PISA-osaamisen välisen yhteyden tarkastelun.

Muuttujien kuvaukset, mahdolliset arvot ja tiedot puuttuvuudesta on esitetty erillisessä Excel-tiedostossa, joka löytyy liitteistä (liite A).

3.2 Aineiston esikäsittely ja puuttuvien arvojen käsittely

Aineisto luettiin SPSS-muodossa (`.SAV`) R-ohjelmistoon `haven`-paketin `read_sav()`-funktiolla. Analyysiin rajattiin ainoastaan Suomen havainnot valitsemalla maamuuttujan perusteella (`CNT == "FIN"`). Tämän jälkeen aineistossa tehtiin kohdenettuja uudelleenkoodauksia, joilla korjattiin analyysin kannalta muutettavia tai yhdenmukaistamista vaativia arvoja. Muuttujissa `ST330D10WA`, `ST250D06JA` ja `ST250D07JA` koodi `9999999` tulkittiin puuttuvaksi arvoksi ja muunnettiin arvoon `NA`. Kodin kieltä kuvaavassa muuttujassa `LANGN` yhdistettiin kaksi suomen kieltä kuvaavaa koodia siten, niin että arvo `815` muunnettiin arvoon `420`, ja arvo `999` tulkittiin puuttuvaksi (`NA`). Lisäksi syntymämaata kuvaavissa muuttujissa `COBN_S` (oppilas), `COBN_M` (äiti) ja `COBN_F` (isä) yhdistettiin kaksi Suomea kuvaavaa koodia, muuttamalla arvo `924600` arvoon `024600`, ja koodi `999999` tulkittiin puuttuvaksi arvoksi (`NA`). Edellä mainitut puuttuvien arvojen koodit on määritelty PISA 2022 -koodikirjassa puuttuviksi arvoiksi (*missing*) [5].

Puuttuvia arvoja tarkasteltiin laskemalla sekä kokonaispuuttuvuus että muuttujakohtaiset puuttuvien arvojen määrät ja osuudet. Puuttuvuuden perusteella aineistoa karsittiin vaiheittain. Ensin poistettiin muuttujat, joissa puuttuvia arvoja oli vähintään 50 % havainnoista. Tämän jälkeen poistettiin vielä havainnot, joissa

puuttuvien arvojen osuus oli 30 % tai enemmän suhteessa jäljelle jääneiden muuttujien määrään. Karsinnan jälkeen aineistosta poistettiin lisäksi muuttujat, joilla ei ollut varianssia (eli joissa arvojen joukossa oli vain yksi uniikki arvo), koska ne eivät sisällä selittävää vaihtelua.

Seuraavaksi aineistosta poistettiin analyysin kannalta tarpeettomia valmiita indeksejä ja taustasummamuuttujia (WLE/indeksimuuttujat) sekä tunniste- ja otantamuuttujia (esim. oppilastunnisteet sekä otantaan ja aineiston hallintaan liittyvät muuttujat), jotta jatkoanalyysi perustuisi valittuihin mittareihin ja kyselyaineiston muuuttujiin.

Sukupuolimuuttuja muodostettiin muuttujasta `ST004D01T` siten, että arvo `0` tulkittiin puuttuvaksi (`NA`) ja jäljelle jääneet arvot luokiteltiin kahteen luokkaan (`female`/`male`). Lopuksi muuttujatyyppien yhdenmukaistamiseksi aineistossa muunnettiin suurin osa muuttujista faktoreiksi, mutta mallinnuksessa numeerisina käsiteltävät muuttujat `ST016Q01NA`, `ST059Q01TA`, `ST059Q02JA`, `AGE`, `GRADE`, `UNIT`, `WVARSTRR` ja `SENWT` säilytettiin tai muunnettiin numeerisiksi. Lisäksi kaikki `PV`-alkuiset (plausible value) ja `W_FSTU`-alkuiset painomuuttujat käsiteltiin numeerisina. Muodostettu analyysiaineisto tallennettiin (`.rds`-tiedostona) toistettavuutta ja muuttujatyyppien säilymistä varten.

3.3 Aineiston kuvaus

Alkuperäisessä koko PISA-aineistossa oli 613 744 havaintoa ja 1 278 muuttujaa. Tämän jälkeen aineisto rajattiin Suomea koskeviin havaintoihin ja tehtiin luvussa 3.2 kuvatut rajaukset ja muutokset, minkä seurauksena analyysiaineistoon jäi 9 124 havaintoa ja 558 muuttujaa.

Puuttuvuutta käsiteltiin luvussa 3.2 kuvatulla tavalla karsimalla sekä muuttujia että havaintoja ennalta määriteltyjen raja-arvojen perusteella. Puuttuvien arvojen osuus vaihteli muuttujittain. Rivikohtaisesti puuttuvien arvojen osuus oli keskimäärin noin 7,6 % (mediaani 4,3 %). Muuttujatasolla puuttuvuus painottui erityisesti osaan oppilaskyselyn väittämistä.

3.3.1 Otantapainot ja replikaatiopainot

Tämä alaluku perustuu erityisesti teokseen *PISA 2022 Technical Report* [2, luku 10]. PISA-aineisto perustuu monivaiheiseen otantaan, kuten luvussa 2.2 todetaan. Tämän vuoksi havaintojen valintatodennäköisyydet eivät ole yhtäsuuria. Tämän korjaamiseksi analyyseissa hyödynnettiin OECD:n määrittelemää oppilastason lopul-

lista otantapainoa `W_FSTUWT`. Painotuksen tavoitteena on, että estimaatit edustavat koko Suomen 15-vuotiaiden perusjoukkoa vuonna 2022.

Kuvailevissa tarkasteluissa tunnusluvut laskettiin painotettuina siten, että jokainen havainto sai painokseen $w_i = \texttt{W\_FSTUWT}_i$. Painotettu keskiarvo (1) estimoitiin muodossa

$$\bar{y}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}, \tag{1}$$

jossa y_i on oppilaan pistemäärä ja w_i vastaava oppilaspaino.

Painotettu varianssi (2) voidaan määritellä muodossa

$$s_w^2 = \frac{\sum_{i=1}^{n} w_i \left(y_i - \bar{y}_w\right)^2}{\sum_{i=1}^{n} w_i}, \tag{2}$$

ja painotettu keskihajonta $s_w = \sqrt{s_w^2}$. Kvantiilit laskettiin painotetusta jakaumasta siten, että havaintojen painot huomioitiin jakauman kertymää muodostettaessa. Taulukoissa raportoitu havaintojen määrä (n) kuvaa otoksen todellista havaintomäärää, jotta otoksen koko on lukijalle selkeä.

3.3.2 Plausible values ja niiden käyttö kuvailevassa tarkastelussa

PISA-osaamista mitataan plausible value (PV) -pistemäärillä, joita on 10 (PV1–PV10) kaikkia osa-aluita kohden. PV-arvot ovat useita vaihtoehtoisia, mallipohjaisia realisaatioita oppilaan osaamisesta, ja niiden tarkoitus on heijastaa mittaamiseen liittyvää epävarmuutta, kuten luvussa 2.2 todetaan. PV-menetelmän vuoksi varsinaisissa analyyseissä tulisi estimoida malli erikseen jokaisella PV-arvolla ja yhdistää tulokset [6, luku 6].

Tässä luvussa PV-arvoja hyödynnettiin ensisijaisesti kuvailevaan analyysiin muodostamalla oppilaskohtainen osa-aluekohtainen PV-keskiarvo

$$\mathrm{PVMEAN}_{i,d} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{PV}_{i,k,d},$$

jossa $d \in \{\mathrm{MATH}, \mathrm{READ}, \mathrm{SCIE}\}$ on osa-alue, $K = 10$ on käytettyjen PV-arvojen lukumäärä ja $\mathrm{PV}_{i,k,d}$ on oppilaan i k:s PV-arvo osa-alueella d. PV-keskiarvoa käytetään tässä havainnollistavana tiivistyksenä osaamisen tasosta, eikä sitä tule tulkita "tarkkana" pistemääränä.

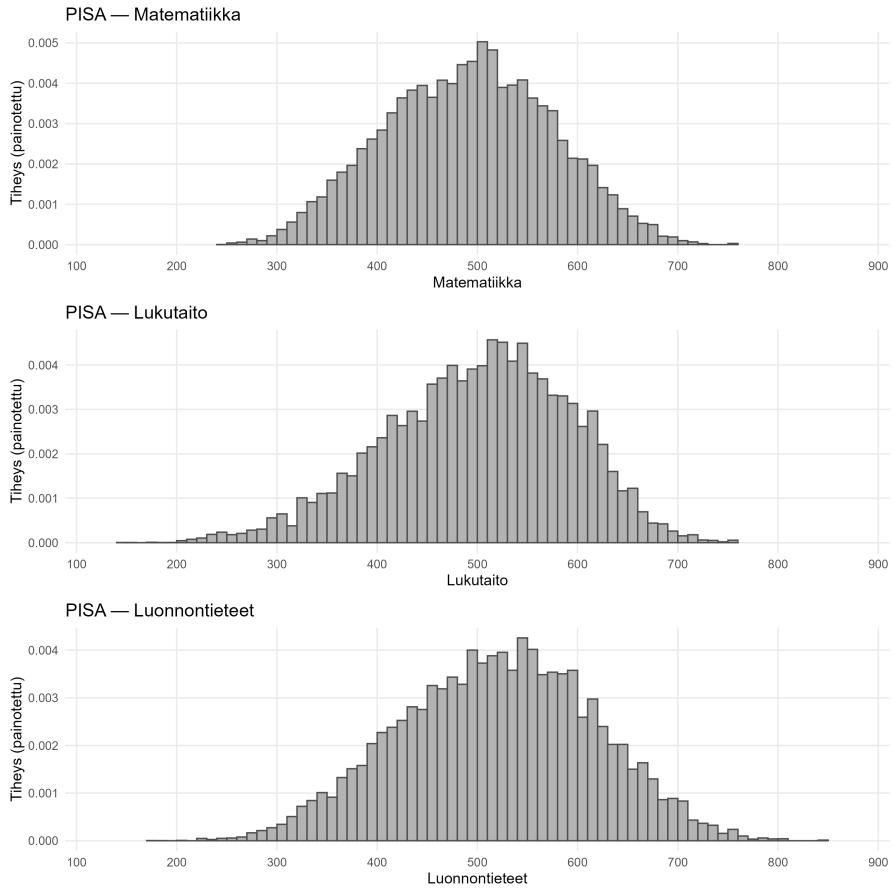
3.3.3 Kuvailevat tulokset

Taulukossa 2 esitetään matematiikan, lukutaidon ja luonnontieteiden PISA-tulosten painotetut kuvailevat tunnusluvut koko Suomen aineistossa. Tunnusluvut (keskiar-

vo, keskihajonta, kvartiilit sekä minimi- ja maksimiarvot) kuvaavat osaamisen keskimääräistä tasoa, jakaumien hajontaa sekä havaintojen vaihteluväliä eri osa-alueilla.

Muuttuja	N	Keskiarvo	Keskihaj.	Min	Q1	Mediaani	Q3	Max
Matematiikka	9124	491.04	82.89	249.10	431.04	493.09	550.24	756.96
Lukutaito	9124	500.21	91.89	140.50	438.60	507.52	567.29	758.64
Luonnontieteet	9124	519.42	96.61	170.78	451.39	522.13	588.39	841.08

Taulukko 2: PISA-tulosten (PV-keskiarvot) painotetut kuvailevat tunnusluvut



Kuva 1: PISA-osaamisjakaumat (matematiikka, lukutaito ja luonnontieteet) koko aineistossa. Histogrammit esittävät painotetun tiheyden.

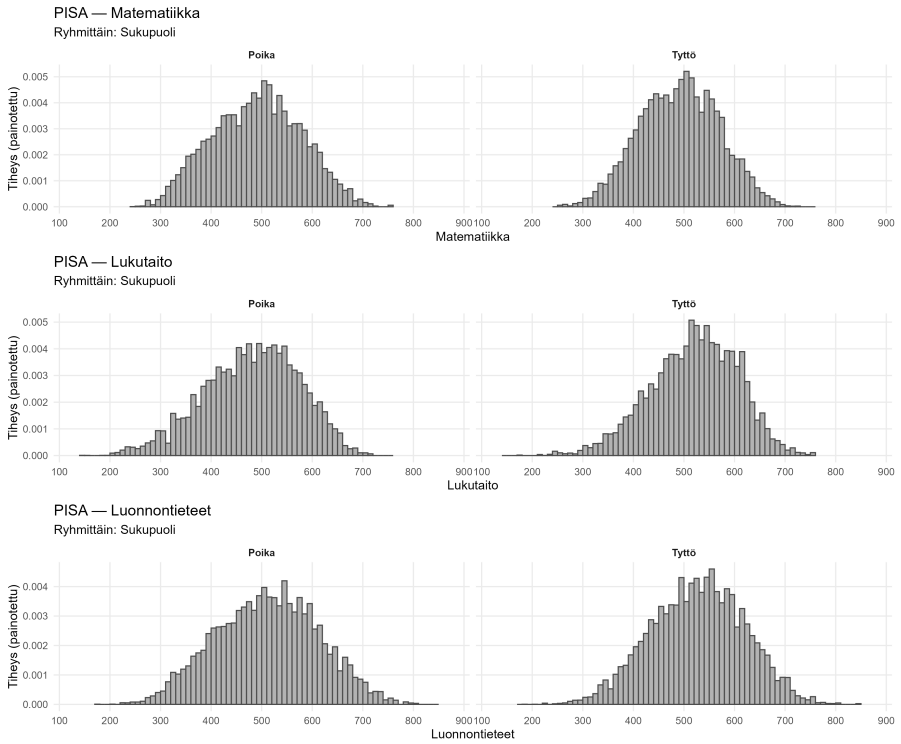
Kuvassa 1 esitetyt histogrammit ja taulukon 2 kuvailevat tunnusluvut antavat yhteneväisen kokonaiskuvan osaamisen jakautumisesta. Kaikilla kolmella osa-alueella pistemäärien jakaumat ovat yksihuippuisia ja muodoltaan likimain symmetrisiä, mikä viittaa normaalijakaumaa muistuttavaan rakenteeseen ilman selviä

poikkeavia piirteitä. Taulukon perusteella keskimääräinen osaamistaso on korkein luonnontieteissä ja matalin matematiikassa, joskin erot keskiarvoissa ovat suhteellisen maltillisia. Keskihajonnat, kvartiilivälit sekä laaja minimi- ja maksimiarvojen vaihteluväli osoittavat, että oppilaiden välinen vaihtelu on huomattavaa kaikilla osaalueilla, mikä on yhdenmukaista kuvassa havaittavan jakaumien leveyden kanssa.

Koska oppimistuloksissa voi esiintyä systemaattisia eroja taustatekijöiden mukaan, osaamista tarkasteltiin myös sukupuolittain. Taulukossa 3 esitetään PV-keskiarvojen painotetut kuvailevat tunnusluvut sukupuolen mukaan, ja kuvassa 2 esitetään vastaavat jakaumat histogrammeina.

Muuttuja	Ryhmä	N	Keskiarvo	Keskihaj.	Q1	Mediaani	Q3
Matematiikka	Poika	4527	491.04	87.44	427.35	493.24	553.36
	Tyttö	4597	491.03	78.17	435.06	493.01	548.16
Lukutaito	Poika	4527	480.27	93.78	415.54	486.19	548.69
	Tyttö	4597	519.76	85.61	463.88	524.22	583.17
Luonnontieteet	Poika	4527	511.25	100.63	438.80	512.59	584.14
	Tyttö	4597	527.44	91.79	462.60	529.73	591.82

Taulukko 3: PISA-tulokset sukupuolen mukaan (painotettu; PV-keskiarvot)



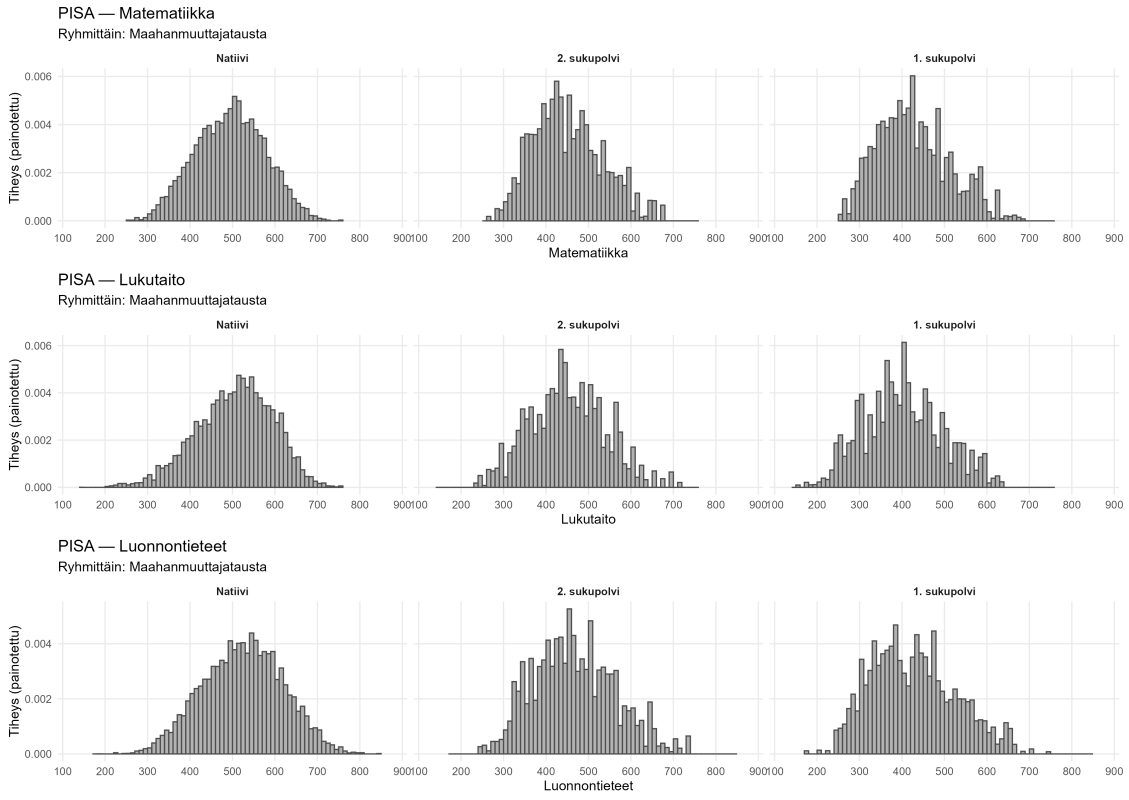
Kuva 2: PISA-osaamisjakaumat sukupuolen mukaan (tytöt ja pojat) osa-alueittain. Histogrammit esittävät painotetun tiheyden.

Taulukon 3 ja kuvan 2 perusteella lukutaidossa tyttöjen keskimääräinen pistemäärä on jonkin verran poikia korkeampi, ja myös jakauma sijoittuu kokonaisuudessaan korkeammalle tasolle. Matematiikassa sukupuolten keskiarvot ovat käytännössä samansuuruiset, ja luonnontieteissä erot ovat suhteellisen pieniä. Keskihajonnat ja kvartiilit osoittavat, että osaamisen hajonta on molemmissa sukupuoliryhmissä melko samankaltainen. Histogrammien perusteella jakaumat ovat kummassakin ryhmässä yksihuippuisia ja muodoltaan likimain symmetrisiä, eikä niissä havaita selviä poikkeavia rakenteita, mikä viittaa normaalijakaumaa muistuttavaan muotoon.

Osaamista tarkasteltiin lisäksi maahanmuuttajataustan mukaan, sillä koulutukselliset erot voivat liittyä sekä oppilaan omaan että perhetaustaan. Taulukossa 4 esitetään PV-keskiarvojen painotetut kuvailevat tunnusluvut kolmessa ryhmässä (natiivi, toisen sukupolven maahanmuuttajataustainen ja ensimmäisen sukupolven maahanmuuttajataustainen), ja kuvassa 3 esitetään vastaavat jakaumat histogrammeina. Muuttujasta puuttui tieto 89 havainnon osalta.

Muuttuja	Ryhmä	N	Keskiarvo	Keskihaj.	Q1	Mediaani	Q3
Matematiikka	Natiivi	7556	494.79	81.44	436.27	496.90	552.86
	2. sukupolvi	674	451.84	82.22	391.48	441.41	507.09
	1. sukupolvi	805	427.48	85.51	362.99	418.88	484.23
Lukutaito	Natiivi	7556	505.61	89.25	446.93	512.72	570.27
	2. sukupolvi	674	450.90	89.66	385.89	446.76	510.45
	1. sukupolvi	805	402.45	92.93	334.65	400.49	464.37
Luonnontieteet	Natiivi	7556	524.81	94.04	458.12	526.59	591.40
	2. sukupolvi	674	466.04	95.42	396.21	458.56	530.74
	1. sukupolvi	805	426.85	99.08	350.25	420.90	491.99

Taulukko 4: PISA-tulokset maahanmuuttajataustan mukaan (painotettu; PV-keskiarvot)



Kuva 3: PISA-osaamisjakaumat maahanmuuttajataustan mukaan (natiivi, 2. sukupolvi, 1. sukupolvi) osa-alueittain. Histogrammit esittävät painotetun tiheyden.

Taulukon 4 ja kuvan 3 perusteella natiivitaustaisilla oppilailla keskimääräinen pistemäärä on kaikilla osa-alueilla korkeampi kuin maahanmuuttajataustaisilla ryhmillä. Erot ovat erityisen selviä lukutaidossa. Ensimmäisen sukupolven maahanmuuttajataustaisilla oppilailla keskiarvot ovat johdonmukaisesti alhaisimmat, kun taas toisen sukupolven oppilaat sijoittuvat keskimäärin näiden ryhmien väliin. Keskihajonnat ja kvartiilit osoittavat, että osaamisen hajonta on kaikissa ryhmissä huomattavaa. Histogrammien perusteella jakaumat ovat pääosin yksihuippuisia ja muodoltaan melko symmetrisiä, joskin maahanmuuttajataustaisissa ryhmissä voidaan havaita lievää vasemmalle vinoutta. Kokonaisuutena jakaumat muistuttavat kuitenkin normaalijakaumaa ilman selviä poikkeavia rakenteita.

Kokonaisuutena ryhmittäinen tarkastelu viittaa siihen, että maahanmuuttajatausta on keskeinen osaamiseen yhteydessä oleva taustatekijä, ja se huomioidaan myöhemmässä mallinnuksessa muiden selittävien muuttujien rinnalla.

4 Käytettävien analyysimenetelmien teoria

4.1 Lineaarinen regressio

Tämä alaluku perustuu ISLR-kirjaan *An Introduction to Statistical Learning* [7, luku 3]. Lineaarinen regressio on klassinen, helposti tulkittava malli. Se kuuluu koneoppimisen ohjatun oppimisen menetelmiin ja ratkaisee regressio-ongelman, jossa selitettävä muuttuja on jatkuva. Menetelmä on parametrinen malli, jossa vasteen odotusarvo kuvataan selittäjien lineaarisena yhdistelmänä ja satunnaisvirheellä. Lineaarisen regression kiinnostuksen kohteena ovat parametrien kertoimet. Tällöin selvitetään, miten selittäjät ovat yhteydessä vasteen keskimääräiseen tasoon, kun muut selittäjät pysyvät vakioina. Menetelmän avulla voidaan myös luoda ennusteita. Tällöin tavoitteena on tuottaa mahdollisimman tarkkoja arvoja uusille havainnoille.

Yleinen monimuuttujainen malliyhtälö kirjoitetaan muotoon

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{3}$$

jossa n on havaintojen määrä ja p selittäjien lukumäärä, β_0 on vakiotermi, β_j ovat kertoimet, x_{ij} on havainnon i arvo selittäjälle X_j, ja ε_i satunnaisvirhe. Malliyhtälön (3) kertoimien tulkinta on suoraviivainen, missä β_j kuvaa vasteen odotusarvon keskimääräistä muutosta, kun X_j kasvaa yhden yksikön muiden selittäjien pysyessä samoina. Kategoriset selittäjät sisällytetään indikaattorimuuttujina suhteessa valittuun viitekategoriaan [7, luku 3.3.1]. Mallin tulosteeseen kuuluu yleensä pisteestimaattien lisäksi keskivirheet, t-arvot ja luottamusvälit sekä mallin sovitusmittarit, kuten jäännösvirheen keskihajonta ja selitysaste R^2.

Parametrit estimoidaan pienimmän neliösumman (PNS/OLS) periaatteella minimoimalla jäännössumma

$$\mathrm{RSS} \;=\; \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 . \tag{4}$$

PNS minimoi (4):n. Matriisimuodossa ratkaisu on

$$\hat{\beta} \;=\; (X^\top X)^{-1} X^\top y,$$

kun $X^\top X$ on kääntyvä. Sovitteen laatua kuvataan selitysasteella

$$R^2 \;=\; 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}, \qquad \mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

(ja tarvittaessa oikaistulla $\bar{R}^2$, joka ottaa huomioon mallin monimutkaisuuden), sekä jäännösvirheen keskihajonnalla

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-p-1}}.$$

Lineaarisen mallin luotettava käyttö edellyttää perusoletusten kohtuullista täyttymistä: yhteyden likimain lineaarisuutta, virheiden nollakeskiarvoa ja homoskedastisuutta (virheiden likimain vakioista varianssia), havaintojen riittävää riippumattomuutta sekä sitä, ettei selittäjissä esiinny täydellistä monikollineaarisuutta. Nämä tarkistetaan jäännöskuvioilla (lineaarisuus ja varianssin vakioisuus), QQ-kuvalla (normaalisuus, erityisesti pienen otoksen päättelyssä) sekä vaikutusdiagnostiikalla, kuten vipuvoimalla ja Cookin etäisyydellä. [7, luku 3.3.3]

Ei-lineaarisuutta ja muuttujien keskinäisiä riippuvuuksia voidaan kuvata laajentamalla mallia polynomitermeillä tai vuorovaikutuksilla. Tällöin vuorovaikutustermi kuvaa sitä, miten toisen selittäjän vaikutus riippuu toisen selittäjän tasosta. Muuttujien keskistäminen ennen vuorovaikutusten lisäämistä helpottaa kertoimien tulkintaa ja voi pienentää monikollineaarisuutta. [7, luku 3.3.3]

Mallin yleistettävyyttä voidaan arvioida erillisellä testijoukolla tai k-kertaisella ristiinvalidoinnilla, jotta pelkät sovitusmittarit (kuten R^2) eivät johda harhaan. Vertailu tehdään ensisijaisesti ennustemittareilla (esim. RMSE tai MAE), ja tulosten vakaus voidaan varmistaa toistamalla vertailu. Mallispesifikaatioita (perusmalli, polynomit, vuorovaikutukset, tarvittaessa harju- tai lassoregressio) verrataan samoilla menetelmillä. [7, luku 5]

Lineaarisen regression vahvuuksia ovat selkeä tulkittavuus (kertoimet kuvaavat vaikutuksia muiden muuttujien pysyessä vakiona), tehokas estimointi ja hyvin vakiintunut päättelykehikko (keskivirheet, testit, luottamusvälit). Menetelmä toimii usein pienilläkin otoskoilla ja antaa suoraan epävarmuusarviot ennusteille. Rajoitteita ovat melko tarkat oletukset aineistolle. Jos oletukset eivät toteudu, kertoimien epävarmuus kasvaa ja tulkinta heikkenee. Lisäksi kategoriset muuttujat on koodattava indikaattoreiksi. [7, luku 3]

4.2 Elastic Net -regularisointi

Tavallinen pienimmän neliösumman estimaattori voi olla epävakaa, jos selittäjien lukumäärä p on suuri suhteessa havaintojen määrään n tai jos selittäjissä esiintyy voimakasta multikollineaarisuutta, kuten tässä aineistossa. Tällöin kertoimien varianssi kasvaa ja ennustetarkkuus heikkenee. Regularisointimenetelmät korjaavat

ongelmaa lisäämällä rangaistustermin, joka pienentää kertoimien suuruuksia ja parantaa mallin yleistettävyyttä.

Harjuregressio (ridge) lisää neliöllisen L_2-rangaistuksen [8]. Parametrit estimoidaan minimoimalla

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{5}$$

jossa $\lambda \geq 0$ on regularisaatioparametri. Harjuregressio kutistaa kertoimia kohti nollaa, mutta tyypillisesti ei aseta niitä täsmälleen nollaksi, joten se parantaa etenkin monikollineaarisuustilanteissa mallin stabiiliutta.

Lassoregressio käyttää absoluuttista L_1-rangaistusta [9]:

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{6}$$

Lasson keskeinen etu on, että osa kertoimista voi saada arvon 0, jolloin malli tekee muuttujavalintaa. Toisaalta vahvasti korreloivien selittäjien tapauksessa lasso voi valita mielivaltaisesti yhden muuttujan ryhmästä ja jättää muut pois, mikä voi heikentää valinnan vakautta.

Elastic Net yhdistää harju- ja lassoregression rangaistukset [10]. Yleinen muoto voidaan esittää kahden hyperparametrin avulla:

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \left((1-\alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right), \tag{7}$$

jossa $\lambda \geq 0$ säätelee kokonaisregularisoinnin voimakkuutta ja $\alpha \in [0, 1]$ painottaa L_1- ja L_2-osien suhdetta. Erikoistapauksina $\alpha = 1$ vastaa tavallista lassoregressiota ja $\alpha = 0$ harjuregressiota. Elastic Netin käytännöllinen etu on, että se voi tuottaa harvan mallin (kuten lasso) mutta samalla käyttäytyä stabiilimmin korreloivilla selittäjillä (harjuregression vaikutus). Menetelmä pyrkii myös valitsemaan korreloivia muuttujia ”ryhminä” yhden muuttujan sijaan [10].

Koska rangaistustermit riippuvat kertoimien mittakaavasta, selittäjät standardoidaan tyypillisesti ennen mallin sovitusta. Vakiotermiä β_0 ei yleensä penalisoida. Hyperparametrit λ ja Elastic Net -mallissa myös α valitaan ristiinvalidoinnilla ennustemittarin, kuten RMSE, perusteella. Näin saadaan malli, joka tasapainottaa sovituskyvyn ja yleistettävyyden erityisesti tilanteissa, joissa perinteinen PNS-estimointi on epävakaa. Tässä työssä tullaan käyttämään Elastic Net -regularisointia.

4.3 Random Forest

Random Forest esiteltiin alun perin Breimanin artikkelissa [11], ja menetelmän käytännön tulkinnat sekä menetelmät on koottu ytimekkäästi ISLR-kirjan lukuun 8 [7]. Random Forest on puupohjainen menetelmä, jossa yksittäinen päätöspuu jakaa selittäjäavaruuden rekursiivisella binäärisellä jakamisella alueisiin ja ennustaa jokaisessa lehdessä alueen havaintojen keskiarvolla (regressio) tai enemmistöluokalla (luokittelu). Varianssia pienennetään opettamalla useita puita eri bootstrap-näytteisiin (bagging) ja ottamalla niiden ennusteiden keskiarvo. Random Forest tehostaa tätä satunnaistamalla piirrevalintaa jokaisessa jaossa (mtry), mikä vähentää puiden keskinäistä korrelaatiota ja parantaa keskiarvoistamisen vaikutusta. Menetelmä soveltuu hyvin ei-lineaarisiin riippuvuuksiin ja vuorovaikutuksiin. [7, 11, luku 8.2]

Menetelmän perusidea on seuraava. Jokaiselle puulle arvotaan näyte alkuperäisestä aineistosta, ja puu kasvatetaan ilman karsintaa. Otantaosuutta voidaan säätää. Täysi bootstrap-otanta vastaa osuutta $1-e^{-1} \approx 0{,}632$, mutta pienempiä osuuksia voidaan käyttää laskennallisen tehokkuuden tai yleistettävyyden parantamiseksi. Jokaisessa solmussa valitaan jakoon paras piirre vain satunnaisesti valitusta m piirteen osajoukosta (tyypillisesti regressiossa $m \approx p/3$, luokittelussa $m \approx \sqrt{p}$). Ennuste muodostetaan yksittäisten puiden keskiarvona:

$$\hat{f}_{\mathrm{RF}}(x) \;=\; \frac{1}{B}\sum_{b=1}^{B} h_b(x), \tag{8}$$

jossa h_b on b:nnen näytteen perusteella opetettu puu ja B puiden lukumäärä. Yhtälö (8) havainnollistaa, että Random Forestin ennuste on yksittäisten puiden ennusteiden keskiarvo, mikä pienentää varianssia verrattuna yksittäiseen puuhun. [7, luku 8.2]

Random Forestilla on luonteva sisäinen arvio yleistettävyydestä. Opetuksen ulkopuolelle jäävät havainnot (*out-of-bag*, OOB) muodostavat luontevan validointijoukon. Näiden puiden ennusteita vertaamalla havaittuun arvoon saadaan OOB-virhe, joka toimii hyvänä approksimaationa testivirheelle ilman erillistä validointijoukkoa. Samalla periaatteella lasketaan piirteiden tärkeydet. Esimerkiksi permutaatiotärkeydessä sekoitetaan yhden piirteen arvot OOB-aineistossa ja mitataan virheen kasvu: suurempi kasvu viittaa informatiivisempaan piirteeseen. [7, 11, luku 8.2]

Käyttöön liittyvät keskeiset asetukset ovat puiden lukumäärä B, jaossa tarkasteltavien piirteiden määrä m, minimilehtikoko ja otantaosuus. B:n kasvattaminen parantaa yleensä vakautta eikä aiheuta ylisovittamista, kun taas m:n valinta tasapainottaa puiden korrelaatiota ja yksittäisen puun laatua. Skaalausta tai kategoristen

muuttujien erityiskäsittelyä ei yleensä tarvita, toisin kuin lineaarisessa regressiossa, mikä tekee menetelmästä käytännössä helppokäyttöisen. Menetelmän vahvuuksia ovat hyvä ennustetarkkuus monimutkaisissa, ei-lineaarisissa tilanteissa, vähäinen esikäsittelyn tarve sekä OOB-arvio, joka nopeuttaa mallinvalintaa. Rajoitteita ovat tulkittavuuden rajallisuus yksittäisten kertoimien tarkasteluun verrattuna. [7, luku 8]

4.4 Extreme Gradient Boosting (XGBoost)

Tämä alaluku perustuu Chenin ja Guestrinin esitykseen skaalautuvasta puutehostusmenetelmästä XGBoost [12]. XGBoost on gradienttitehostukseen perustuva puutehostusmenetelmä, jossa malli rakennetaan lisäämällä yksi heikko oppija kerrallaan ja jokainen uusi puu kohdistetaan selittämään edellisten mallien jättämiä virheitä. Tätä kutsutaan tehostukseksi (boosting). Toisin kuin Random Forestissa, jossa puut opetetaan rinnakkain ja yhdistetään keskiarvona, tehostuksessa puut opetetaan järjestyksessä ja niiden ennusteet summataan. Menetelmä hyödyntää puiden joustavuutta ei-lineaaristen riippuvuuksien ja vuorovaikutusten mallintamisessa sekä sisäistä säännöllistystä ylisovittamisen hillitsemiseksi. Tehostuksen yleinen kehys esitellään myös ISLR:ssa [7, luku 8.2.3].

XGBoostin perusmalli on puukokonaisuus

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in \mathcal{F}, \tag{9}$$

jossa puuluokka on

$$\mathcal{F} = \{\, f(x) = w_{q(x)} \mid q : \mathbb{R}^p \to \{1, \ldots, T\},\ w \in \mathbb{R}^T \,\}.$$

Yhtälö (9) havainnollistaa, että ennuste muodostuu K:n puun summana, jolloin kukin lisätty puu täydentää aiempien mallien jättämiä virheitä. Näissä n on havaintojen määrä ja p selittäjien lukumäärä. Symboli K tarkoittaa puiden lukumäärää eli tehostusiteraatioiden määrää, kun taas T on kunkin yksittäisen puun lehtien lukumäärä. Funktio $q(x)$ liittää havainnon x johonkin lehteen ja w_j on lehden j paino. Oppiminen määritellään säännöllistetyn tavoitefunktion minimointina

$$\mathcal{L}(\phi) \;=\; \sum_{i=1}^{n} \ell\big(y_i, \hat{y}_i\big) \;+\; \sum_{k=1}^{K} \Omega(f_k), \qquad \Omega(f) \;=\; \gamma T \;+\; \tfrac{1}{2}\lambda \|w\|^2,$$

jolloin menetelmä suosii yksinkertaisia ja hyvin yleistyviä puita, mikä hillitsee ylisovittamista. Tässä $\ell(y_i, \hat{y}_i)$ on valittu häviöfunktio (esim. neliö- tai logistinen häviö)

ja $\Omega(f)$ on säännöllistystermi, joka rankaisee puun lehtien määrää T ja lehtipainojen suuruutta w. Optimoinnissa iteraatiossa t lisätään uusi puu f_t parantamaan edellisen vaiheen ennustetta $\hat{y}_i^{(t-1)}$, ja tavoitetta lähennetään toisen kertaluvun approksimaatiolla, jossa hyödynnetään sekä ensimmäisen että toisen derivaatan gradienttitilastoja. Käytännössä XGBoost tukee kahta jakotapaa: tarkkaa hakua sekä suurille aineistoille soveltuvia histogrammi- tai kvantiilipohjaisia likimääräisiä menetelmiä. Ylisovittamisen ehkäisemiseksi käytetään oppimisnopeutta η (shrinkage) sekä aliotantaa, jossa osa havainnoista (`subsample`) tai osa selittävistä muuttujista (`colsample_bytree`) valitaan kunkin puun rakentamiseen. Algoritmi käsittelee harvaa aineistoa tehokkaasti, sillä se oppii puuttuville arvoille oletussuunnan ja muodostaa jakoehdokkaat painotetun kvantiilimenetelmän avulla suurissa tai painotetuissa aineistoissa, jolloin jakokohdat voidaan valita tiiviistä edustajajoukosta ilman koko jakauman läpikäyntiä. [12]

XGBoostin käytön keskeisiä parametreja ovat puiden lukumäärä K, yksittäisen puun maksimisyvyys (`max_depth`), oppimisnopeus η, solmun jakamiseen vaadittavaa minimipainoa säätelevä `min_child_weight`, havaintojen aliotantaa kuvaava `subsample`, piirrekohtainen aliotanta (`colsample_bytree`) sekä säännöllistysparametrit λ ja γ. Parametri `min_child_weight` määrittää pienimmän sallitun havaintopainon, joka solmussa täytyy olla ennen uuden jaon tekemistä, mikä toimii rakenteellisena säännöllistyksenä. Parametri `subsample` kontrolloi kuinka suuri osa havainnoista käytetään yksittäisen puun rakentamiseen, kun taas `colsample_bytree` määrittää kuinka suuri osa selittävistä muuttujista valitaan puun rakentamiseen. Parametri γ määrittää vähimmäisparannuksen häviöfunktiossa, joka vaaditaan uuden splitin tekemiseen. Pienempi η ja riittävän suuri K parantavat usein yleistettävyyttä, kun taas λ ja γ säätelevät lehtipainojen suuruutta ja puun rakenteellista monimutkaisuutta. Menetelmän vahvuuksia ovat korkea ennustetarkkuus erilaisissa tehtävissä ja aineistoissa, toimivat keinot ylisovittamisen hallintaan (säännöllistys, oppimisnopeus, aliotanta) sekä tehokas toteutus myös suurille ja harvoille aineistoille. Rajoitteita ovat tulkittavuuden rajallisuus verrattuna parametristen mallien kertoimiin sekä hyperparametrien herkkyys, minkä vuoksi huolellinen mallinvalinta, optimointi ja virheen seuranta ovat keskeisiä käytännön soveltamisessa. [12, 13]

5 Analyysi

5.1 Analyysissa käytetyt ohjelmat ja paketit

Analyysi toteutettiin R-ohjelmointikielellä (versio 4.5.2), joka tarjoaa laajat mahdollisuudet tilastolliseen analyysiin ja koneoppimiseen. R valittiin erityisesti sen monipuolisten data-analyysiin soveltuvien työkalujen vuoksi.

Aineiston käsittelyssä ja muokkauksessa hyödynnettiin erityisesti `dplyr`- ja `tidyr`-paketteja, jotka mahdollistavat tehokkaan datan suodatuksen, muuntamisen ja yhdistelyn. PISA-aineiston lukemiseen käytettiin `haven`-pakettia, jonka avulla SPSS-muotoiset tiedostot voitiin käsitellä suoraan R-ympäristössä. Mallinnuksessa käytettiin kolmea eri koneoppimismenetelmää. Elastic Net -mallit toteutettiin `glmnet`-paketin avulla, Random Forest -mallit `ranger`-paketilla ja XGBoost -mallit `xgboost`-paketilla. Mallien arvioinnissa hyödynnettiin `rsample`-pakettia, jonka avulla toteutettiin sisäkkäinen ristiinvalidointi. Tulosten analysoinnissa ja visualisoinnissa käytettiin `ggplot2`-pakettia, joka mahdollistaa joustavan ja selkeän kuvien tuottamisen. Lisäksi `ggrepel`-pakettia hyödynnettiin kuvien selkeyttämisessä esimerkiksi muuttujien nimien sijoittelussa. Muita työssä käytettyjä paketteja olivat `tibble`, `patchwork` ja `Matrix`.

Analyysissa käytetyt koodit sekä mallien toteutukset on koottu ja dokumentoitu GitHub-repositorioon, joka löytyy liitteistä (liite B).

5.2 Mallien arviointimittarit

Mallien suorituskykyä arvioitiin regressioanalyysissä yleisesti käytetyillä virhemittareilla: jäännösvirhehajonnalla (Root Mean Squared Error, RMSE), keskiitseisvirheellä (Mean Absolute Error, MAE) sekä selitysasteella (R^2). Koska PISA-aineistossa jokaisella oppilaalla on otospaino `W_FSTUWT`, joka kuvaa oppilaan edustamien oppilaiden lukumäärää perusjoukossa [2, luku 10], kaikki metriikat laskettiin painotettuina versioina. Olkoon y_i havaittu arvo, $\hat{y}_i$ mallin ennustama arvo, $w_i > 0$ otospaino ja $\bar{y}_w$ painotettu keskiarvo kuten luvussa 3.2 määriteltiin. Regressiomallien virhettä mitataan usein ensin keskineliövirheen (Mean Squared Error, MSE) avulla

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

josta painotettu RMSE saadaan laajentamalla otospainoilla ja ottamalla neliöjuuri:

$$\mathrm{RMSE}_w = \sqrt{\frac{\sum_i w_i (y_i - \hat{y}_i)^2}{\sum_i w_i}}.$$

RMSE on samassa yksikössä kuin selitettävä muuttuja ja antaa suuremman painon suurille virheille neliöinnin takia [7, luku 2.2]. Painotettu MAE määritellään

$$\mathrm{MAE}_w = \frac{\sum_i w_i |y_i - \hat{y}_i|}{\sum_i w_i}.$$

Toisin kuin RMSE, MAE ei korota virheitä toiseen potenssiin, joten se on vähemmän herkkä yksittäisille suurille virheille ja toimii RMSE:n täydentävänä mittarina [14]. Painotettu R^2 määritellään

$$R^2_w = 1 - \frac{\sum_i w_i (y_i - \hat{y}_i)^2}{\sum_i w_i (y_i - \bar{y}_w)^2},$$

ja se kuvaa, kuinka suuren osan selitettävän muuttujan painotetusta vaihtelusta malli pystyy selittämään [7, luku 3.1]. Vertailutasona käytettiin jokaisessa testitaitoksessa ulomman opetusaineiston painotettua keskiarvoa, joka kuvaa yksinkertaisimman mahdollisen ennustimen suorituskykyä.

Mallien tulkitsemisen tueksi laskettiin muuttujien tärkeysarvot kullekin mallille. Tärkeysarvot kuvaavat, kuinka paljon kukin selittävä muuttuja vaikuttaa mallin ennusteisiin, mutta niiden laskutapa eroaa mallien välillä. Elastic Netissä tärkeys määritellään sovitetun mallin kertoimen itseisarvona $|\hat{\beta}_j|$, Random Forestissa permutaatiotärkeydellä ja XGBoostissa Gain-mittarilla. Koska tärkeysarvot eivät ole suoraan vertailukelpoisia keskenään mallien välillä, kunkin mallin tärkeys tulkittiin erikseen ja mallien välistä vertailua tehtiin muuttujien suhteellisen tärkeysjärjestyksen perusteella.

Koska kaikki kategoriset muuttujat enkoodattiin one-hot-muotoon, yksittäinen alkuperäinen muuttuja voi tuottaa useita binäärisiä sarakkeita. Tärkeysarvot aggregoitiin takaisin alkuperäisen muuttujan tasolle summaamalla binääristen sarakkeiden tärkeysarvot, jotta muuttujien vertailu olisi mielekästä alkuperäisten PISA-muuttujien tasolla.

Yksittäisten ajojen vakauden arvioimiseksi laskettiin jokaiselle muuttujalle kolme yhteenvetomittaria kaikista 250 ajosta (5 ulompaa taitosta $\times$ 10 PV-arvoa $\times$ 5 siementä). Kokonaistärkeys kuvaa binääristen sarakkeiden tärkeysarvojen summaa kaikissa ajoissa. Se on pääasiallinen tärkeysmittari. Paras keskimääräinen sijoitus kuvaa muuttujan tärkeimmän binääristen sarakkeen tyypillistä sijaa tärkeysjärjestyksessä kaikissa ajoissa, pienempi arvo tarkoittaa tärkeämpää muuttujaa. Top-

k-osuus kuvaa, kuinka suuressa osassa ajoja muuttujan tärkein binäärinen sarake kuuluu 20 tärkeimmän muuttujan joukkoon. Muuttujat, joilla on suuri kokonaistärkeys, pieni paras keskimääräinen sijoitus ja korkea top-k-osuus yhtäaikaisesti, ovat mallinnuksen kannalta johdonmukaisimmin merkittäviä. Tärkeysarvojen visualisoimiseksi käytettiin kahta kuvatyyppiä: pylväsdiagrammia, joka esittää 20 tärkeimmän muuttujan kokonaistärkeyden, sekä kuplakuvaajaa, jossa vaaka-akseli kuvaa keskimääräistä parhaan sijoituksen keskiarvoa, pystyakseli top-k-osuutta ja pisteiden koko kokonaistärkeyttä. Näin yhdessä kuvassa näkyvät muuttujan tärkeys, johdonmukaisuus ja vakaus.

5.3 Hyperparametrien optimointi ja sisäkkäinen ristiinvalidointi

Kaikkien kolmen mallin Elastic Netin (luku 4.2), Random Forestin (luku 4.3) ja XGBoostin (luku 4.4) hyperparametrit optimoitiin sisäkkäisellä ristiinvalidoinnilla (nested cross-validation). Menetelmässä mallin valinta ja suorituskyvyn arviointi erotetaan toisistaan kahdella erillisellä ristiinvalidointikerroksella. Ulompi ristiinvalidointi tuottaa harhattoman testiarvion mallin yleistettävyyskyvystä ja sisempää ristiinvalidointia käytetään vain hyperparametrien valintaan opetusaineiston sisällä. Jos hyperparametreja viritettäisiin ja mallia arvioitaisiin samalla aineistolla (flat cross-validation), suorituskykyarvio voisi olla optimistisesti harhainen, koska testiaineisto on epäsuorasti vaikuttanut hyperparametrien valintaan [15]. Vaikka Wainer & Cawley [16] toteavat, että perinteinen ristiinvalidointi voi käytännön sovelluksissa antaa samankaltaisia tuloksia kuin sisäkkäinen ristiinvalidointi. Valittiin tässä työssä turvallisempana vaihtoehtona käyttää sisäkkäistä ristiinvalidointia erityisesti siksi, että useita malleja vertaillaan toisiinsa.

Ulompi ristiinvalidointi toteutettiin $k = 5$-taittojaolla. Jokaisessa ulommassa kierroksessa $\frac{4}{5}$ aineistosta toimi ulompana opetusaineistona ja $\frac{1}{5}$ testiaineistona. Hyperparametrien optimointi tehtiin sisemmässä ristiinvalidoinnissa $k = 3$-taittojaolla, jolloin ulompi opetusaineisto jaettiin siten, että $\frac{2}{3}$ toimi opetusaineistona ja $\frac{1}{3}$ validointiaineistona. Koko aineistoon suhteutettuna yhden ulomman kierroksen aikana aineisto jakautuu likimäärin $\frac{8}{15}$ opetusaineistoon, $\frac{4}{15}$ validointiaineistoon ja $\frac{1}{5}$ testiaineistoon. Testiaineistoa ei käytetty missään vaiheessa hyperparametrien valintaan tietovuotojen välttämiseksi [15]. Sisemmässä ristiinvalidoinnissa hyperparametrit valittiin minimoimalla painotettu RMSE (5.2) validointiaineistolla. Tämän jälkeen mallien suorituskykyä arvioitiin kaikilla kolmella mittarilla (luku 5.2) ulom-

massa testiarvioinnissa. Toistettavuuden varmistamiseksi ajot toistettiin viidellä eri satunnaissiemenellä ($seed \in \{101, 202, 303, 404, 505\}$), jotta taittojakojen ja algoritmien satunnaisuuden vaikutus tuloksiin ei perustuisi yksittäiseen aineiston jakoon. Lopulliset tulokset raportoitiin kaikkien siementen ja kaikkien kymmenen plausible value -muuttujan yli laskettuina keskiarvoina 95 %:n luottamusväleineen (t-jakauma, vapausasteet $n - 1$).

Kuten luvussa 3.3.2 kuvataan, PISA raportoi osaamisen kymmenellä PV-muuttujalla (PV1–PV10) mittausepävarmuuden huomioimiseksi. OECD:n suosituksen mukaisesti [6, luku 6] jokainen PV-muuttuja toimi tässä työssä vastemuuttujana erikseen, ja lopullinen suorituskykyarvio muodostettiin kaikkien kymmenen PV-muuttujan tulosten painotettuna keskiarvona. Ennen mallinnusta kaikki PV-muuttujat sekä niistä johdetut aiemmin lasketut aggregaatit (esim. PVMEAN-muuttujat) poistettiin selittäjistä tietovuodon estämiseksi, sillä ne sisältävät suoraan vastemuuttujan informaatiota.

Elastic Net -regressiossa (luku 4.2) optimoitiin kaksi hyperparametria. Regularisaation muoto $\alpha \in [0, 1]$ ja sen voimakkuus $\lambda > 0$ [10]. Optimointi toteutettiin siten, että α valittiin ennalta määritellystä diskreetistä arvojoukosta {0,0, 0,1, ..., 1,0}, joka kattaa 11 arvoa puhtaasta harjuregressiosta ($\alpha = 0$) puhtaaseen lassoon ($\alpha = 1$). Jokaiselle α-arvolle sovitettiin `glmnet`-malli, joka tuottaa automaattisesti λ-polun ($n_\lambda = 100$, `lambda.min.ratio` $= 10^{-4}$) yhdellä mallinsovituksella, mikä tekee menetelmästä laskennallisesti tehokkaan [17]. Sisemmässä ristiinvalidoinnissa mallilla ennustettiin validointiaineiston arvot koko λ-polulla, ja parhaaksi λ-arvoksi valittiin se, joka tuotti pienimmän painotetun RMSE:n. Parametri α valittiin sen perusteella, mikä arvo tuotti pienimmän keskimääräisen validaatio-RMSE:n sisemmissä taitoksissa. Ulomman taitoksen lopullinen $\lambda^\star$ määritettiin ottamalla sisemmissä taitoksissa valittujen parhaiden λ-arvojen mediaani, minkä jälkeen malli sovitettiin ulomman taitoksen koko opetusaineistoon ja arvioitiin testiaineistolla. Mallissa käytetyt selittävät muuttujat standardoitiin.

Random Forest -mallissa (luku 4.3) optimoitiin kolme hyperparametria ruudukkohaulla. Parametrit `mtry` (selittäjien lukumäärä kutakin jakoa kohden), `min.node.size` (pienin sallittu solmukoko) ja `sample.fraction` (osuus havainnoista yksittäisen puun kasvattamiseen), jotka ovat random forestin suorituskykyyn eniten vaikuttavat hyperparametrit [18]. Parametrin `mtry`-hakutila muodostettiin automaattisesti muuttujien lukumäärän p perusteella: $\{\lfloor\sqrt{p}/2\rfloor, \lfloor\sqrt{p}\rfloor, 2\lfloor\sqrt{p}\rfloor\}$. Hakutila rajattiin pieniin arvoihin, koska esitesteissä suuremmat `mtry`-arvot (kuten $\lfloor p/3\rfloor$) eivät parantaneet ennustetarkkuutta mutta kasvattivat laskenta-aikaa jopa viisin-

kertaiseksi. Probst *et al.* [18] mukaan pieni `mtry` on suositeltava silloin kun relevantteja selittäjiä on paljon, koska se estää vahvoja muuttujia peittämästä heikompia, mikä vastaa tämän aineiston rakennetta. One-hot-enkoodauksen jälkeen $p = 1815$, joten hakutila vastaa arvoja $\{21, 42, 84\}$. Parametrin `min.node.size`-arvoina käytettiin $\{5, 20, 40\}$ ja `sample.fraction`-arvoina käytettiin {0,5, 0,632, 0,8}, missä $0{,}632 \approx 1-e^{-1}$ vastaa tavallisen bootstrap-otannan odotettua peittoa. Yhteensä ruudukko sisälsi $3 \times 3 \times 3 = 27$ parametrikombinaatiota. Kullekin parametrikombinaatiolle opetettiin malli sisemmällä koulutusaineistolla ja laskettiin painotettu RMSE sisemmällä validointiaineistolla. Paras yhdistelmä valittiin sisempien taittojen keskimääräisen validaatio-RMSE:n perusteella. Otospainot syötettiin `case.weights`-parametrin kautta. Viritysvaiheessa käytettiin 200 puuta laskennallisen tehokkuuden vuoksi ja lopullisessa ulommassa sovituksessa 500 puuta, jotta lopullinen malli on vakaampi. Permutaatiotärkeys laskettiin ainoastaan lopulliselle ulommalle mallille. [18, 19].

XGBoostissa (luku 4.4) hyperparametreja on enemmän ja ne voivat vaikuttaa yhdessä, joten optimointi tehtiin satunnaishaulla (random search). Satunnaishaku on ruudukkohaun veroinen tai sitä parempi korkeaulotteisessa hakutilassa, koska se kohdistaa kokeet tasaisemmin koko hakuavaruuteen [20]. Kussakin ulommassa taitoksessa arvottiin $m = 30$ parametrikombinaatiota ja vertailtiin ne sisemmässä ristiinvalidoinnissa painotetulla RMSE:llä. Hakutila oli seuraava: `eta` log-uniformisti [0,01; 0,20], `max_depth` kokonaislukuna [2; 10], `min_child_weight` log-uniformisti [0,5; 20], `subsample` ja `colsample_bytree` tasaisesti [0,5; 1], `gamma` niin että noin $\frac{1}{4}$ kandidaateista käytti arvoa 0 ja loput log-uniformisti $[10^{-4}; 10]$, sekä L_2-regularisaatio `lambda` log-uniformisti $[10^{-3}; 100]$ ja L_1-regularisaatio `alpha` log-uniformisti $[10^{-3}; 10]$. Hakuvälit määriteltiin suositusten ja [13] ohjeiden perusteella siten, että ne kattavat käytännössä relevantin parametriavaruuden. Mallin opetus tehtiin painotetulla `xgb.DMatrix`-datalla ja laskentaa nopeutettiin `tree_method=hist`-asetuksella. Puiden lukumäärä (`nrounds`) valittiin sisemmän ristiinvalidoinnin sisällä varhaisen pysäyttämisen (`early stopping`) menettelelmällä. Kaikissa sisemmissä taitoksissa mallia opetettiin sisemmällä opetusaineistolla ja sen suoriutumista seurattiin sisemmällä validointiaineistolla, jos validointi-RMSE ei parantunut 50 peräkkäiseen iteraatioon (`early_stopping_rounds = 50`), opetus pysäytettiin ja talteen otettiin paras iteraatiokierros `best_iteration`. Enimmäisiteraatiomäärä oli 3 000. Parametrikombinaation lopulliseksi kierrosmääräksi asetettiin sisempien taitteiden `best_iteration`-arvojen mediaani, ja ulomman taitoksen lopullinen malli opetettiin koko ulommalla opetusaineistolla kiinnitetyllä kierros-

määrällä ilman varhaistaa pysäyttämistä. Näin `nrounds` valittiin ilman, että ulomman taitoksen testiaineisto vaikuttaa siihen. Muuttujien tärkeys laskettiin Gain-mittarilla, joka kuvaa kunkin muuttujan keskimääräistä parannusta kohdefunktiossa sen jakoja sisältävissä puissa. [12, 13]

Kaikille kolmelle mallille käytettiin identtistä esikäsittelyä vertailukelpoisuuden varmistamiseksi. Jatkuvien muuttujien puuttuvat arvot korvattiin mediaanilla ja kategoristen muuttujien puuttuvat arvot enkoodattiin erilliseksi indikaattoriluokaksi (missing-dummy). Jälkimmäinen on PISA-kontekstissa perusteltu valinta, sillä vastaamattomuus ei ole satunnaista vaan voi olla yhteydessä oppilaan osaamiseen [21]. Indikaattorimuuttuja mahdollistaa mallin itse arvioida puuttuvuuden informatiivisuuden. PISA-kyselymuuttujat koodattiin faktoreiksi, koska niiden asteikkovälit eivät ole yhtäsuuria eikä lineaarisuusoletus ole perusteltu. Kaikki kategoriset muuttujat enkoodattiin tämän jälkeen one-hot-muotoon (`sparse.model.matrix`), jolloin mallit operoivat identtisellä piirreavaruudella. Ennen mallinnusta selittäjistä poistettiin PV-muuttujien lisäksi kaikki niistä johdetut aggregaatit, OCOD-koodausmuuttujat ja vakiomuuttujat. Tietovuodon estämiseksi esikäsittelyparametrit mediaanit ja faktoritasot laskettiin aina kulloisenkin opetusaineiston perusteella eikä koko aineistosta [15].

6 Tulokset

6.1 Menetelmien vertailu

Taulukossa 6 esitetään kolmen mallin suorituskyky matematiikan, lukutaidon ja luonnontieteiden ennustamisessa opetusaineistossa, ja vastaavat tulokset testiaineistossa on esitetty taulukossa 5. Testiaineistossa sovitteet on korvattu ennusteilla. Tulokset on raportoitu kaikkien 250 ajon keskiarvoina ja 95 %:n luottamusväleinä, joissa 250 ajoa koostuu viidestä ulommasta taitoksesta, kymmenestä PV-muuttujasta ja viidestä satunnaissiemenestä. Luottamusvälit kuvaavat taittojakojen, PV-muuttujien ja siementen aiheuttamaa satunnaisvaihtelua suorituskykyarviossa. Kapeat luottamusvälit viittaavat siihen, että tulokset ovat vakaita eivätkä riipu yksittäisestä aineistojaosta tai satunnaissiemenestä. Kuvissa 4–6 havainnollistetaan lisäksi mallien välistä vaihtelua yksittäisten ajojen tasolla RMSE-, MAE- ja R^2-mittareilla, jolloin myös yksittäisten taitosten ja PV-arvojen välinen vaihtelu on nähtävissä.

Malli	RMSE	MAE	R^2
Matematiikka (opetus)			
Elastic Net	46.0 [45.9, 46.1]	36.5 [36.4, 36.6]	0.719 [0.718, 0.720]
Random Forest	29.2 [28.9, 29.6]	21.6 [21.2, 21.9]	0.885 [0.882, 0.889]
XGBoost	36.2 [35.5, 36.9]	28.4 [27.8, 29.0]	0.822 [0.815, 0.828]
Lukutaito (opetus)			
Elastic Net	52.5 [52.4, 52.5]	41.6 [41.5, 41.7]	0.720 [0.719, 0.720]
Random Forest	34.3 [33.8, 34.8]	25.2 [24.7, 25.6]	0.879 [0.875, 0.883]
XGBoost	41.5 [40.6, 42.3]	32.5 [31.8, 33.2]	0.820 [0.813, 0.827]
Luonnontieteet (opetus)			
Elastic Net	54.3 [54.2, 54.4]	43.3 [43.2, 43.3]	0.723 [0.722, 0.724]
Random Forest	36.2 [35.7, 36.8]	26.9 [26.4, 27.4]	0.875 [0.871, 0.879]
XGBoost	42.7 [41.8, 43.5]	33.5 [32.8, 34.3]	0.825 [0.818, 0.831]

Taulukko 5: Mallien suorituskyky opetusaineistossa (keskiarvo [95 % luottamusväli])

Malli	RMSE	MAE	R^2
Matematiikka (testi)			
Elastic Net	51.8 [51.7, 51.9]	41.2 [41.1, 41.3]	0.644 [0.642, 0.645]
Random Forest	62.2 [62.1, 62.4]	49.8 [49.7, 49.9]	0.486 [0.485, 0.487]
XGBoost	52.5 [52.4, 52.6]	41.7 [41.6, 41.8]	0.634 [0.632, 0.636]
Lukutaito (testi)			
Elastic Net	59.4 [59.3, 59.6]	47.1 [47.0, 47.2]	0.639 [0.638, 0.641]
Random Forest	72.3 [72.2, 72.5]	57.3 [57.2, 57.4]	0.466 [0.465, 0.467]
XGBoost	61.6 [61.5, 61.8]	48.8 [48.7, 48.9]	0.612 [0.610, 0.614]
Luonnontieteet (testi)			
Elastic Net	61.7 [61.6, 61.8]	49.2 [49.1, 49.3]	0.642 [0.640, 0.644]
Random Forest	75.6 [75.5, 75.8]	60.5 [60.4, 60.6]	0.462 [0.461, 0.464]
XGBoost	63.5 [63.4, 63.7]	50.7 [50.6, 50.8]	0.620 [0.618, 0.622]

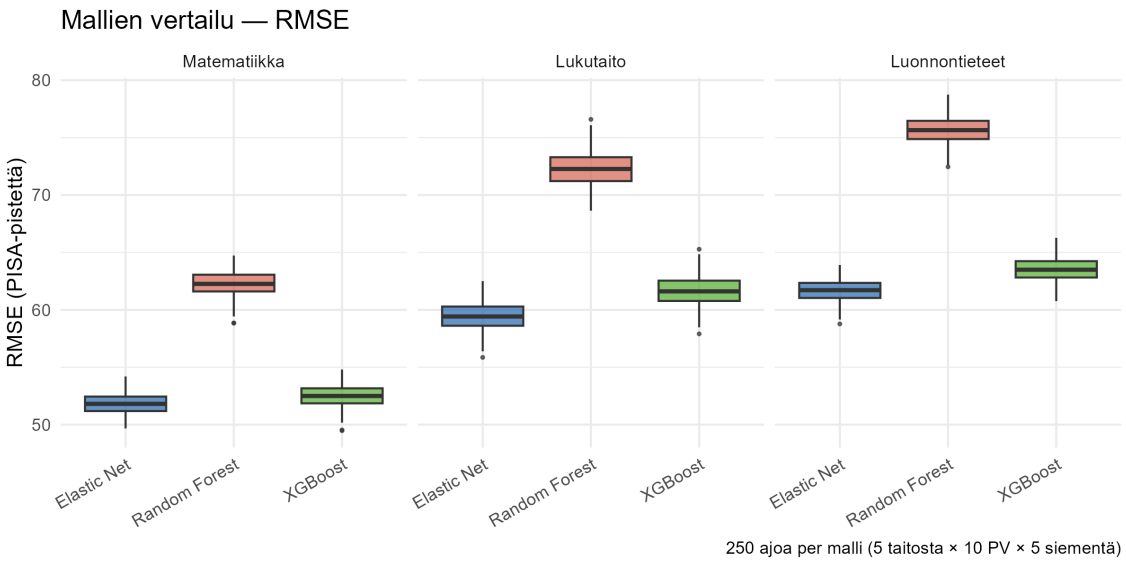
Taulukko 6: Mallien suorituskyky testiaineistossa osa-alueittain (keskiarvo [95 % luottamusväli])

Opetusaineiston ja testiaineiston tulosten vertailu osoittaa selviä eroja mallien yleistettävyyskyvyssä. Erityisesti Random Forest ja XGBoost saavuttavat erittäin korkean selitysasteen opetusaineistossa, mutta niiden suorituskyky heikkenee merkittävästi testiaineistossa. Tämä viittaa ylisovittamiseen, jossa malli oppii aineistosta paitsi todellisen signaalin myös satunnaisvaihtelua ja kohinaa, jotka eivät yleisty uusiin havaintoihin. Ilmiö näkyy erityisen selvästi Random Forest -mallissa, jossa R^2 laskee opetusaineiston noin 0,88:sta testiaineiston noin 0,46:een kaikilla osa-alueilla.

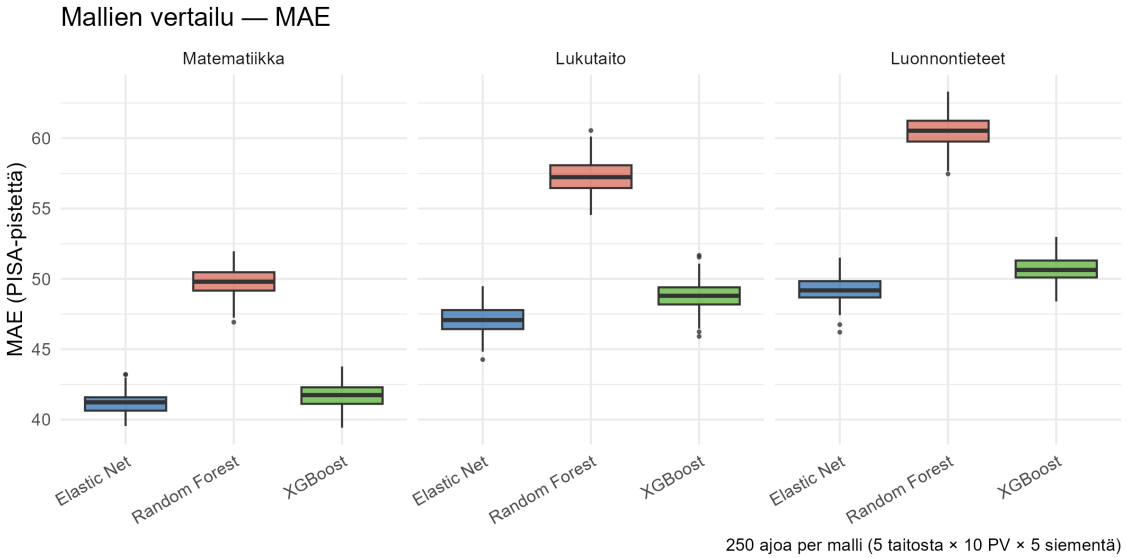
Ylisovittaminen voi johtua useista tekijöistä. Puupohjaiset mallit ovat rakenteeltaan hyvin joustavia, jolloin ne kykenevät muodostamaan monimutkaisia ei-lineaarisia riippuvuuksia ja vuorovaikutuksia muuttujien välillä. Mikäli mallin hyperparametrit ovat suhteellisen väljät (esimerkiksi suuri puiden määrä, syvät puut, pieni minimisolmukoko tai korkea oppimisnopeus), malli voi sovittaa opetusaineiston havaintoja liiankin tarkasti. Tällöin malli oppii myös aineistokohtaisia sattumanvaraisia piirteitä, jotka eivät toistu testiaineistossa, mikä heikentää ennustetarkkuutta.

Elastic Net -malli puolestaan on regularisoitu lineaarinen malli, joka rajoittaa mallin kompleksisuutta rankaisemalla suuria kertoimia. Tämä vähentää ylisovittamisen riskiä ja johtaa pienempään eroon opetusaineiston ja testiaineiston suorituskyvyn välillä. Tästä syystä Elastic Net ei saavuta yhtä korkeaa selitysastetta opetusaineistossa kuin joustavammat mallit, mutta sen suorituskyky säilyy paremmin

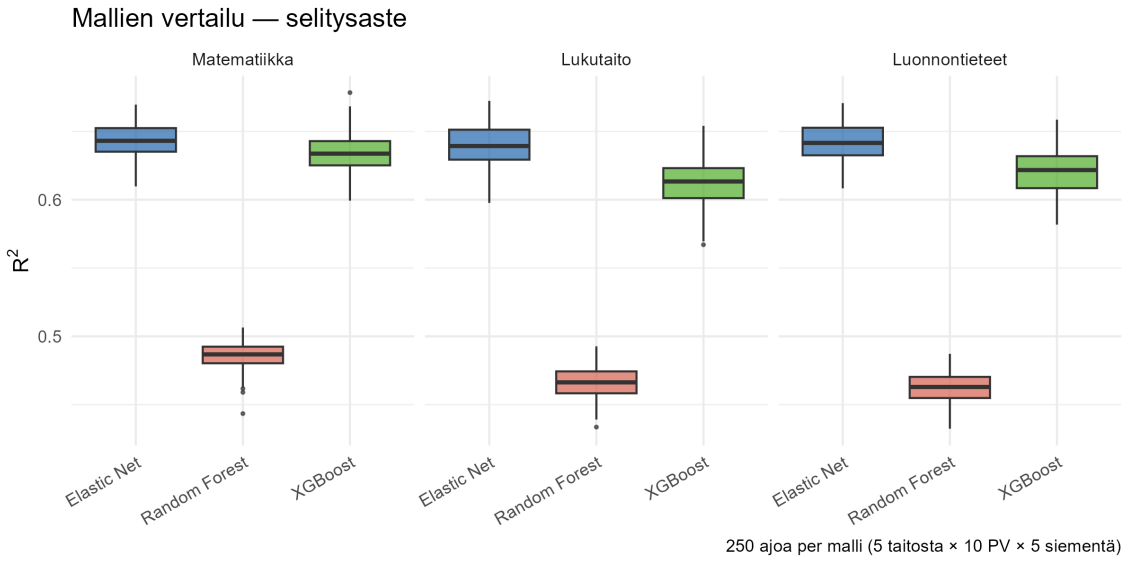
testiaineistossa, mikä tekee siitä luotettavamman ennustajan.



Kuva 4: Mallien RMSE-jakaumat testiaineistossa osa-alueittain (PISA-pistettä).



Kuva 5: Mallien MAE-jakaumat testiaineistossa osa-alueittain (PISA-pistettä).



Kuva 6: Mallien selitysasteen (R^2) jakaumat testiaineitossa osa-alueittain.

Kaikilla kolmella osa-alueella parhaat tulokset saavutettiin pääosin Elastic Net -mallilla. Matematiikassa Elastic Net tuotti pienimmän virheen (RMSE 51,8, MAE 41,2) ja korkeimman selitysasteen ($R^2 = 0{,}644$). XGBoost -mallin suorituskyky oli hyvin lähellä Elastic Net -mallia (RMSE 52,5, MAE 41,7, $R^2 = 0{,}634$). Random Forest -malli jäi selvästi jälkeen molemmista muista malleista (RMSE 62,2, MAE 49,8, $R^2 = 0{,}486$). Näin ollen matematiikan osa-alueella lineaarinen regularisoitu malli oli hieman ei-lineaarista tehostusmallia parempi, eikä puupohjainen lähestymistapa saavuttanut vastaavaa ennustetarkkuutta.

Lukutaidossa Elastic Net oli niin ikään paras kaikilla mittareilla (RMSE 59,4, MAE 47,1, $R^2 = 0{,}639$). XGBoost sijoittui toiseksi (RMSE 61,6, MAE 48,8, $R^2 = 0{,}612$), ja Random Forest jäi jälleen selvästi heikoimmaksi (RMSE 72,3, MAE 57,3, $R^2 = 0{,}466$). Absoluuttiset virheet olivat lukutaidossa suurempia kuin matematiikassa kaikilla malleilla — Elastic Netin RMSE kasvoi 51,8:sta 59,4:ään, mikä heijastaa lukutaidon suurempaa hajontaa aineistossa (keskihajonta 91,9 vs. 82,9 pistettä, ks. taulukko 2). Suhteellinen selitysaste R^2 pysyi kuitenkin lähes samana (0,639 vs. 0,644), joten mallien suhteellinen kyky selittää vaihtelua ei heikentynyt lukutaidossa.

Luonnontieteissä Elastic Net saavutti myös parhaat tulokset (RMSE 61,7, MAE 49,2, $R^2 = 0{,}642$), XGBoost sijoittui toiseksi (RMSE 63,5, MAE 50,7, $R^2 = 0{,}620$), ja Random Forest jäi jälleen selvästi heikoimmaksi (RMSE 75,6, MAE 60,5, $R^2 = 0{,}462$). Absoluuttiset virheet olivat luonnontieteissä suurimmat kaikista kolmesta

osa-alueesta, mikä on odotuksenmukaista aineiston suurimman hajonnan (96,6 pistettä, ks. taulukko 2) vuoksi. Suhteellinen selitysaste oli kuitenkin käytännössä identtinen muiden osa-alueiden kanssa, joten mallien kyky selittää vaihtelua oli yhtä hyvä kaikilla osa-alueilla.

Kaikkien kolmen osa-alueen tulokset osoittavat johdonmukaisesti, että Elastic Net on paras malli PISA-oppimistulosten ennustamisessa taustamuuttujilla, XGBoost sijoittuu toiseksi ja Random Forest heikoimmaksi. Tämä järjestys toistui kaikilla kolmella osa-alueella ja kaikilla kolmella mittarilla (RMSE, MAE, R^2), mikä viittaa siihen että tulos on robusti eikä riipu valitusta osa-alueesta tai arviointimittarista. Mallien väliset erot olivat tilastollisesti merkitseviä kaikilla osa-alueilla ja kaikilla mittareilla, sillä mallien 95 %:n luottamusvälit eivät menneet päällekkäin (ks. taulukko 6). Elastic Netin ja XGBoostin välinen ero oli pieni (R^2:ssa korkeintaan 0,027 yksikköä), mutta sekin oli tilastollisesti merkitsevä luottamusvälien perusteella. Random Forestin ero parhaaseen malliin oli huomattavasti suurempi, noin 0,15–0,18 R^2-yksikköä osa-alueesta riippuen. Tulos viittaa siihen, että PISA-taustamuuttujien ja oppimistulosten välinen yhteys on pääosin lineaarinen, jolloin regularisoidusta lineaarisesta mallista saadaan paras ennustetarkkuus.

Kuvista 4–6 nähdään, että yksittäisten ajojen välinen vaihtelu on huomattavan pientä kaikilla malleilla, sillä laatikkokuvaajat ovat kapeita ja mediaanit vakaita. Tämä vahvistaa, että tulokset eivät riipu yksittäisestä taitoksesta, PV-arvosta tai satunnaissiemenestä. Random Forestin jakaumat ovat jonkin verran leveämpiä kuin Elastic Netin ja XGBoostin, mikä viittaa siihen että Random Forestin suorituskyky vaihtelee enemmän aineistojaosta riippuen. Elastic Netin ja XGBoostin laatikkokuvaajat ovat lähes yhtä kapeita ja sijoittuvat selvästi Random Forestin yläpuolelle kaikilla osa-alueilla. Elastic Netin ja XGBoostin jakaumat ovat osittain päällekkäisiä, mikä heijastaa näiden mallien lähellä toisiaan olevaa suorituskykyä. Ero on pieni mutta luottamusvälien perusteella tilastollisesti merkitsevä (ks. taulukko 6). Random Forestin jakaumat eivät sen sijaan mene päällekkäin kummankaan muun mallin kanssa millään osa-alueella, mikä vahvistaa että sen heikompi suorituskyky on tilastollisesti merkitsevä kaikilla osa-alueilla ja kaikilla mittareilla.

6.2 Valitut hyperparametrit

Sisemmässä ristiinvalidoinnissa valittujen hyperparametrien jakaumat olivat johdonmukaisia kaikilla osa-alueilla. Elastic Netissä α-parametri painottui voimakkaasti arvoihin 0 ja 0,1 kaikilla osa-alueilla, mikä tarkoittaa että lähes puhdas harjuregres-

sio tai hyvin lievästi lassoa kohti painottunut malli osoittautui parhaaksi regularisointimuodoksi. Tulos on yhdenmukainen Elastic Netin teoreettisen kehyksen kanssa, sillä kun selittäjät korreloivat voimakkaasti, lasso valitsee mielivaltaisesti yhden muuttujan korrelaatioryhmästä, kun taas harjuregressio jakaa painon tasaisemmin koko ryhmälle [10]. PISA-kyselymuuttujien voimakas keskinäinen korrelaatio tekee harju-tyyppisestä regularisoinnista odotuksenmukaisen valinnan tässä aineistossa. Parametri λ painottui suuriin arvoihin, mikä viittaa voimakkaaseen regularisointiin.

Random Forestissa suurin mtry-arvo hakutilasta voitti johdonmukaisesti lähes kaikissa ajoissa kaikilla osa-alueilla. Tulos kertoo mahdollisesti liian pienestä hakutilasta. Malli suosisi todennäköisesti vielä suurempaakin mtry-arvoa. Tämä on kuitenkin yhdenmukainen sen kanssa, että aineistossa on vain muutama vahva selittäjä, jotka malli haluaa nähdä mahdollisimman usein jakokohdissa [18]. Minimilehtikoko painottui pienimpään hakutilan arvoon selvästi useimmin kaikilla osa-alueilla, ja otantaosuus suurimpaan arvoon lähes aina, jolloin suurempi otantaosuus tuotti paremman ennustetarkkuuden tässä aineistossa.

XGBoostissa oppimisnopeus `eta` painottui mataliin arvoihin, mikä yhdistettynä suuriin iteraatiomääriin on tyypillinen optimaalinen yhdistelmä XGBoostissa [12]. Puiden maksimisyvyys `max_depth` painottui mataliin arvoihin eli yksinkertaiset puut yleistyvät parhaiten tässä aineistossa. Tämä on yhdenmukainen Elastic Netin harjuregressio tuloksen kanssa, molemmat viittaavat siihen että aineiston rakenne suosii yksinkertaisempia malleja monimutkaisempien vuorovaikutusten sijaan.

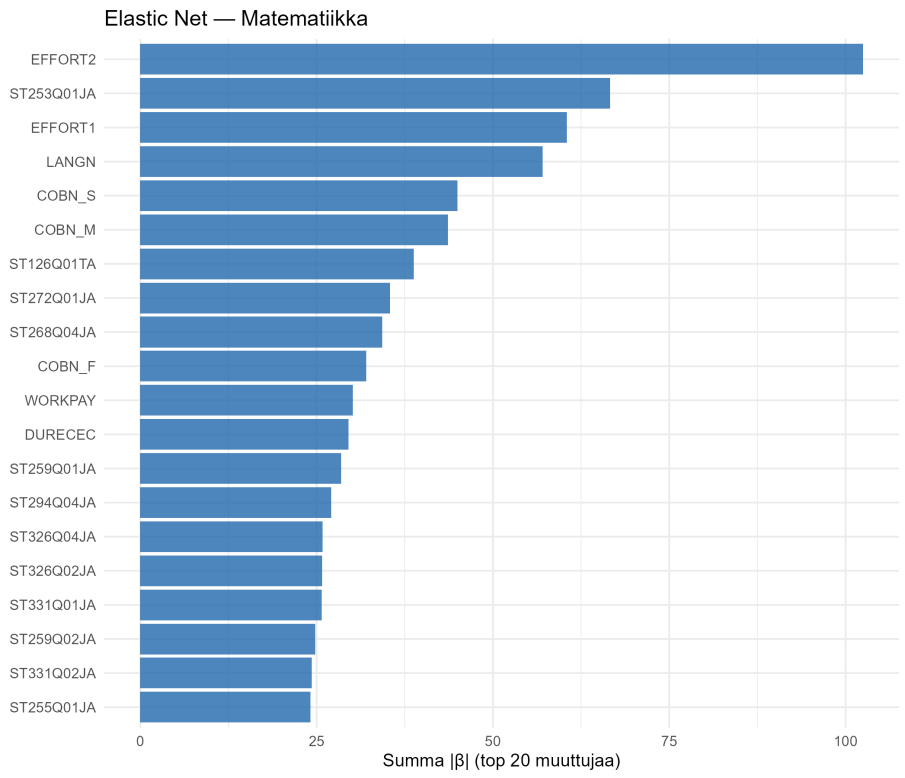
6.3 Tärkeimmät taustatekijät tulosten taustalla

6.3.1 Matematiikka

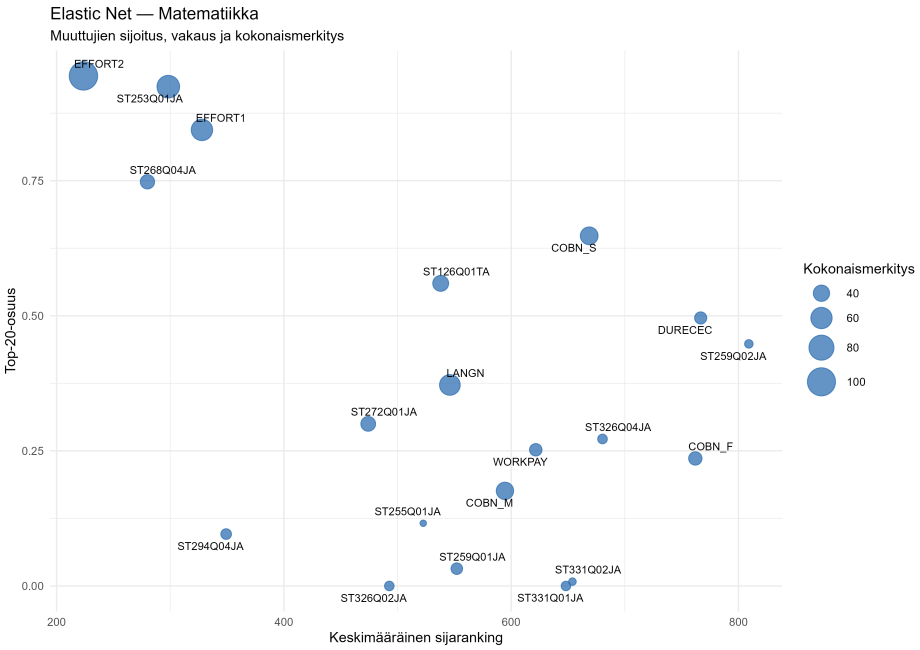
Tarkastellaan ensin muuttujia, jotka vaikuttavat eniten matematiikan tuloksiin. Kuvassa 7 esitetään Elastic Net -mallin 20 tärkeintä muuttujaa kokonaistärkeyden mukaan. Muuttuja `EFFORT2` on selvästi tärkein. Se kuvaa oppilaan itse arvioimaa ponnistusta PISA-testiä varten, ja sen kokonaistärkeys on lähes kaksinkertainen seuraavaksi tärkeimpään muuttujaan `ST253Q01JA` verrattuna. Kyseinen muuttuja kuvaa digitaalisten laitteiden määrää kotona ja toimii siten sosioekonomisen taustan välillisenä mittarina. Muuttuja `EFFORT1`, joka kuvaa oppilaan vaivannäköä koulussa ennen testiä, sijoittuu myös korkealle. Yhdessä `EFFORT2`:n kanssa tämä kertoo, että opiskelumotivaatio ja työskentelyasenne ovat keskeisesti yhteydessä matematiikan osaamisen kanssa. Lisäksi kotikieltä kuvaava `LANGN` sekä oppilaan (`COBN_S`), äidin (`COBN_M`) ja isän (`COBN_F`) syntymämaata kuvaavat muuttujat nousevat esiin. Nämä

muuttujat kuvaavat maahanmuuttajataustaa, jonka yhteys osaamiseen havaittiin jo kuvailevassa tarkastelussa (luku 3.3.3).

Kuvassa 8 tarkastellaan muuttujien tärkeyttä, vakautta ja kokonaismerkitystä yhtäaikaisesti. Vaaka-akseli kuvaa keskimääräistä sijoitusta (pienempi = tärkeämpi), pystyakseli top-20-esiintymisosuutta ja pisteiden koko kokonaistärkeyttä. Muuttuja `EFFORT2` sijoittuu selvästi vasempaan yläkulmaan, mikä tarkoittaa, että se on sekä tärkein että vakain muuttuja. Myös `ST253Q01JA` ja `EFFORT1` ovat vakaita ja merkittäviä. Kiinnostava havainto on `ST268Q04JA`:n korkea sijoittuminen esiintymistiheydessä top-20:ssa. Tämä muuttuja mittaa oppilaan omaa käsitystä matematiikan helppoudesta, ja sen korkea vakaus viittaa siihen, että minäpystyvyyskokemus näyttäytyy johdonmukaisesti tärkeänä tekijänä matematiikan osaamisen yhteydessä. Sen sijaan maahanmuuttajataustaa kuvaavat muuttujat (`LANGN`, `COBN_S`, `COBN_M`) ovat kokonaistärkeydeltään suuria mutta esiintyvät hieman harvemmin top-20-listalla, sillä niiden merkitys jakautuu usean keskenään korreloivan muuttujan kesken, jotka kuvaavat samaa ilmiötä eri näkökulmista.

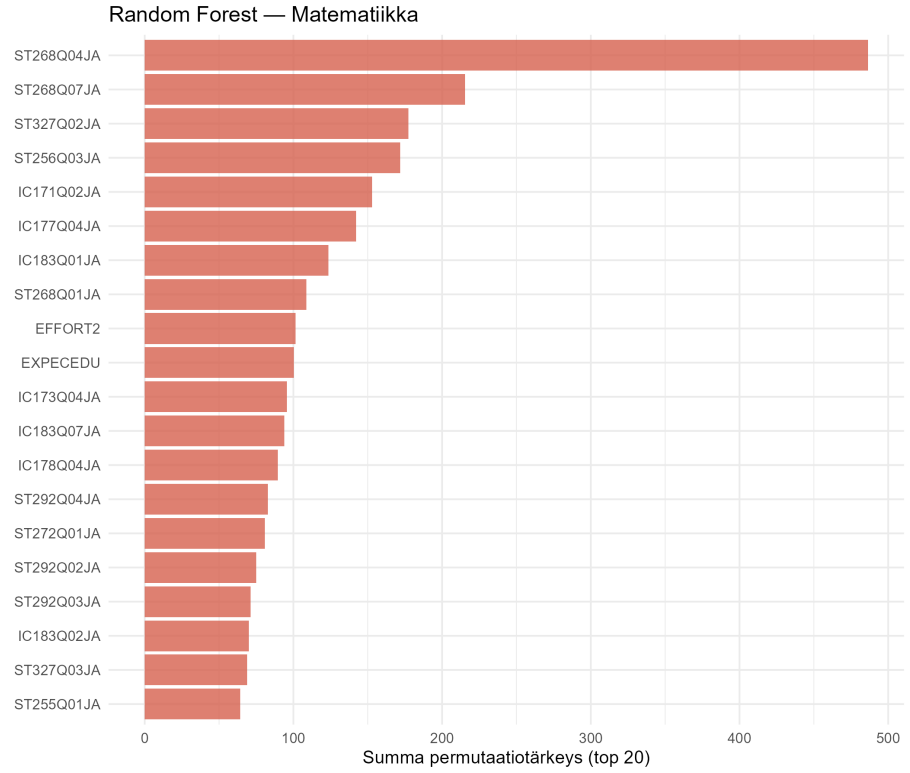


Kuva 7: Elastic Net -mallin tärkeimmät muuttujat matematiikassa.

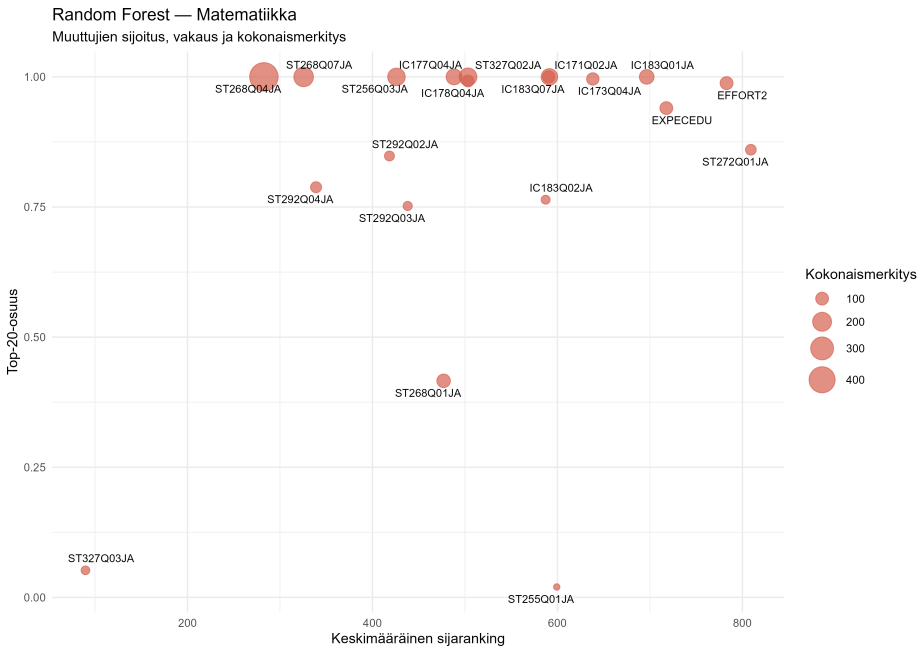


Kuva 8: Elastic Net -mallin muuttujien sijoitus, vakaus ja kokonaismerkitys matematiikassa.

Kuvassa 9 esitetään Random Forest -mallin 20 tärkeintä muuttujaa permutaatiotärkeyden perusteella. Muuttuja `ST268Q04JA` on selvästi tärkein, ja sen kokonaistärkeys on yli kaksinkertainen seuraavaan muuttujaan `ST268Q07JA` verrattuna. Molemmat kuuluvat samaan muuttujasarjaan ja kuvaavat oppilaan käsityksiä matematiikan helppoudesta ja halua menestyä matematiikan tunneilla eli puupohjaisessa mallissa matematiikka-asenteet ja minäpystyvyys nousevat tärkeiksi osaamiseen liittyviksi tekijöiksi. Lisäksi useat `IC`-alkuiset muuttujat, jotka kuvaavat tieto- ja viestintätekniikan käyttöä, nousevat esiin. Tieto- ja viestintätekniikan käytön yhteys matematiikan osaamiseen voi heijastaa sekä sosioekonomista taustaa että oppimiseen liittyviä toimintatapoja. Muuttuja `EFFORT2` esiintyy myös Random Forestissa tärkeimpien muuttujien joukossa, mutta selvästi alemmalla sijalla kuin Elastic Netissä. Sen sijaan `EXPECEDU`, joka kuvaa koulutusodotuksia, nousee tärkeämmäksi, mikä viittaa siihen, että oppilaan tulevaisuuden odotukset voivat liittyä osaamiseen epälineaarisella tavalla, jota lineaarinen malli ei tavoita yhtä hyvin. Kuvan 10 mukaan useat keskeiset muuttujat esiintyvät top-20-listalla lähes kaikissa ajoissa, mikä kertoo tärkeysrakenteen vakaudesta. Toisaalta kokonaismerkitys jakautuu tasaisemmin muuttujien kesken kuin Elastic Netissä, jossa yksittäinen muuttuja dominoi selvästi. Osa muuttujista esiintyy vain satunnaisesti, mikä viittaa epävakaampaan rooliin mallissa.

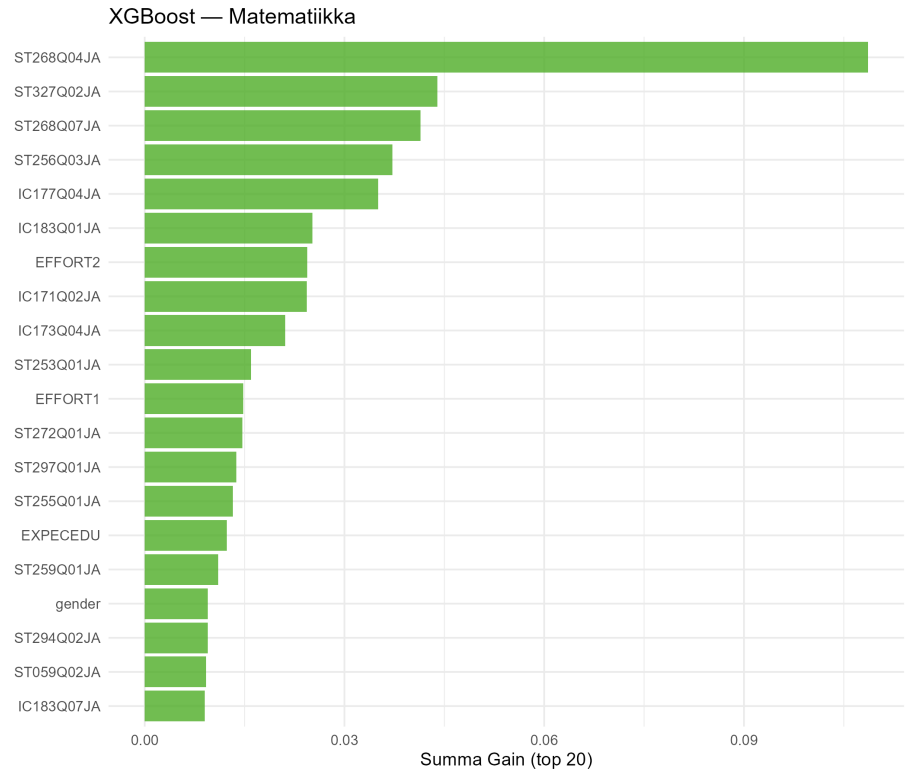


Kuva 9: Random Forest -mallin tärkeimmät muuttujat matematiikassa.

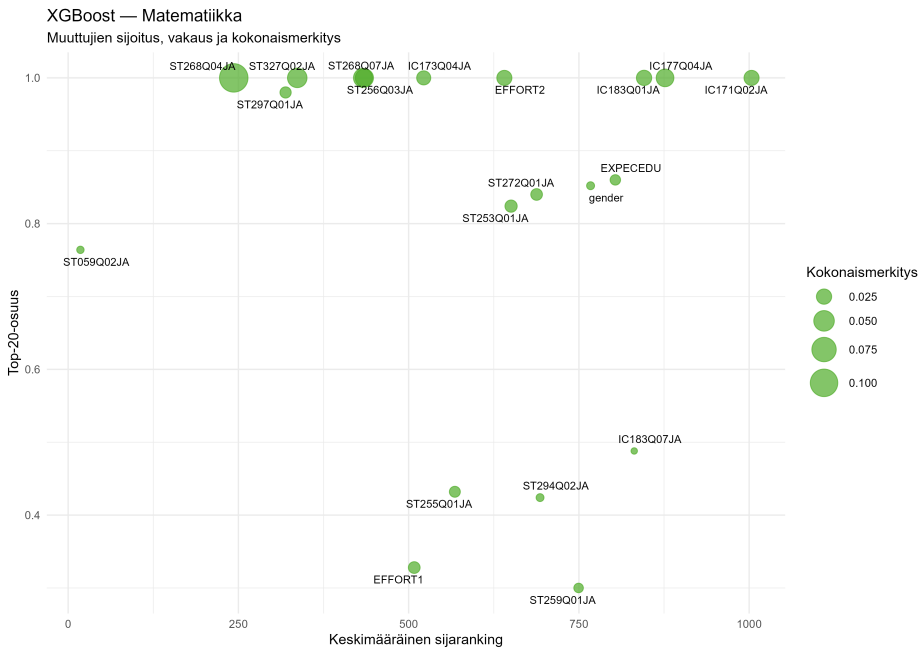


Kuva 10: Random Forest -mallin muuttujien sijoitus, vakaus ja kokonaismerkitys matematiikassa.

Kuvassa 11 esitetään XGBoost -mallin tärkeimmät muuttujat, jotka ovat hyvin samansuuntaisia kuin Random Forest -mallissa. Myös tässä mallissa `ST268Q04JA` on selvästi tärkein, mikä vahvistaa havaintoa minäpystyvyyden keskeisestä roolista puupohjaisissa malleissa. Seuraaviksi tärkeimmät muuttujat liittyvät niin ikään asenteisiin, motivaatioon sekä tieto- ja viestintätekniikan käyttöön. Motivaatiota kuvaavat muuttujat `EFFORT2` ja `EFFORT1` ovat mukana top-20:ssa, mutta niiden suhteellinen merkitys on pienempi kuin Elastic Netissä. Lisäksi muuttuja `gender` nousee esiin, mikä viittaa mahdolliseen epälineaariseen yhteyteen sukupuolen ja matematiikan osaamisen välillä, jota muut mallit eivät löytäneet. Kuvassa 12 useat muuttujat esiintyvät hyvin vakaasti tärkeimpien joukossa erityisesti muuttuja `ST268Q04JA`, vaikka yksittäisten muuttujien merkitys vaihtelee jonkin verran ajojen välillä.



Kuva 11: XGBoost -mallin tärkeimmät muuttujat matematiikassa.



Kuva 12: XGBoost -mallin muuttujien sijoitus, vakaus ja kokonaismerkitys matematiikassa.

Yhteenvetona matematiikan tuloksista keskeisimmiksi osaamiseen liittyviksi tekijöiksi nousevat oppilaan motivaatio ja asenne, erityisesti ponnistelua ja minäpystyvyyttä kuvaavat muuttujat. Lisäksi sosioekonomista taustaa kuvaavat tekijät ovat johdonmukaisesti yhteydessä matematiikan osaamiseen. Nämä tulokset viittaavat siihen, että sekä yksilön oma suhtautuminen oppimiseen että erilaiset taustatekijät muodostavat keskeisen selityspohjan matematiikan tuloksille.

Eri mallien tulokset ovat pääosin yhteneviä keskeisten muuttujien osalta, vaikka niiden keskinäinen järjestys ja suhteellinen painotus vaihtelevat. Tulokset tukevat havaintoa, että havaittu signaali on aineistossa itsessään eikä yksittäisen mallin tuottama ilmiö. Samat muuttujaryhmät erityisesti motivaatio, minäpystyvyys ja sosioekonomiset tekijät nousevat toistuvasti esiin kaikissa malleissa, mikä vahvistaa tulosten tulkinnan luotettavuutta.

Sen sijaan osa muuttujista, kuten kieli- ja syntymämaata kuvaavat tekijät, nousevat esiin selkeämmin vain tietyissä malleissa. Niiden yhteys matematiikan osaamiseen voi olla heikompi, epäsuorempi tai mallin rakenteesta riippuva, eikä yhtä vahvasti aineistossa toistuva kuin keskeisimmillä muuttujilla. Lisäksi näitä samaa ilmiötä kuvaavia muuttujia on useita, mikä voi hajauttaa niiden merkitystä erityisesti puupohjaisissa malleissa, joissa keskenään korreloivat muuttujat jakavat selitysvoimaa.

6.3.2 Lukutaito

Tarkastellaan seuraavaksi lukutaidon tuloksiin liittyviä keskeisiä muuttujia. Elastic Net -mallissa (kuvat 21 ja 22, liite C) tärkeimmäksi muuttujaksi nousee selvästi `EFFORT2`, joka kuvaa oppilaan itse arvioimaa ponnistelua PISA-testiä varten. Myös `EFFORT1` sijoittuu korkealle, mikä korostaa motivaation ja työskentelyasenteen merkitystä lukutaidossa. Lisäksi maahanmuuttotaustaa kuvaavat muuttujat, kuten `COBN_S` ja `LANGN`, ovat keskeisessä asemassa. Tulokset tukevat havaintoa, että kieliympäristö ja taustatekijät ovat yhteydessä lukutaitoon. Kodin resursseja kuvaavat `ST253Q01JA`, `ST259Q01JA` ja `ST258Q01JA` sekä työntekoon ja varhaiskasvatukseen liittyvät muuttujat (`WORKPAY` ja `DURECEC`) esiintyvät myös tärkeimpien muuttujien joukossa. Kuplakuvaajien perusteella erityisesti `EFFORT2` ja `EFFORT1` ovat sekä merkittäviä että vakaita muuttujia, kun taas maahanmuuttotaustaa kuvaavien muuttujien merkitys jakautuu usean keskenään korreloivan muuttujan kesken samalla tavalla kuin matematiikassa.

Puupohjaisissa malleissa painotus muuttuu osittain. Random Forest- ja XGBoost-malleissa (kuvat 23–26, liite C) keskeiseksi muuttujaksi nousee `ST256Q03JA`, joka kuvaa nykykirjallisuuden teosten määrää kotona. Lisäksi useat `IC`-alkuiset muuttujat korostuvat molemmissa malleissa, mikä viittaa tieto- ja viestintätekniikan käyttöön liittyvien toimintatapojen yhteyteen lukutaitoon. Toisaalta nämä molemmat muuttujaryhmät voidaan mieltää myös sosioekonomisen taustan mittareina. Myös koulutusodotuksia kuvaavat muuttujat, kuten `EXPECEDU` ja `ST327Q02JA`, sijoittuvat korkealle erityisesti XGBoost-mallissa. Vaikka motivaatiota kuvaavat `EFFORT1`- ja `EFFORT2`-muuttujat kuuluvat edelleen merkittävimpien muuttujien joukkoon, niiden suhteellinen merkitys jää pienemmäksi kuin Elastic Net -mallissa. Kuplakuvaajista nähdään lisäksi, että puupohjaiset mallit tunnistavat varsin johdonmukaisesti samat keskeiset muuttujat eri ajoissa. Tärkeys kuitenkin jakautuu useamman muuttujan kesken kuin Elastic Net -mallissa, eikä yksittäinen muuttuja dominoi yhtä selvästi. Erityisesti tieto- ja viestintätekniikkaan sekä koulutusodotuksiin liittyvät muuttujat muodostavat vakaasti toistuvan ryhmän.

Yhteenvetona lukutaidon tuloksista keskeisimmiksi osaamiseen liittyviksi tekijöiksi nousevat motivaatio, kodin resurssit sekä oppimisympäristöön ja arjen toimintaan liittyvät tekijät. Elastic Net korostaa erityisesti ponnistelua ja työskentelyasennetta, kun taas puupohjaiset mallit painottavat enemmän kodin kulttuurisia resursseja, koulutusodotuksia ja tieto- ja viestintätekniikan käyttöä. Mallien välillä on eroja painotuksissa, mutta keskeiset muuttujaryhmät toistuvat varsin johdonmu-

kaisesti, mikä viittaa siihen, että havaittu signaali on pääosin aineistossa itsessään eikä yksittäisen mallin tuottama ilmiö.

6.3.3 Luonnontieteet

Luonnontieteiden tuloksissa eri mallien painotukset ovat hyvin samankaltaisia kuin lukutaidon tuloksissa. Elastic Net -mallissa (kuvat 27 ja 28, liite C) keskeisimmäksi muuttujaksi nousee `EFFORT2`, joka kuvaa oppilaan itse arvioimaa ponnistelua PISA-testiä varten. Myös `EFFORT1` kuuluu tärkeimpien muuttujien joukkoon, mikä korostaa motivaation merkitystä luonnontieteiden osaamisessa. Lisäksi kieliympäristöön ja maahanmuuttotaustaan liittyvät muuttujat, kuten `LANGN` sekä `COBN`-muuttujat, sijoittuvat korkealle. Tämä kertoo siitä, että taustatekijöillä on selvä yhteys myös luonnontieteiden tuloksiin. Mukana keskeisten muuttujien joukossa ovat myös kodin resursseja ja oppilaan ajankäyttöä kuvaavat muuttujat, kuten `ST253Q01JA` ja `WORKPAY`. Kuplakuvaajien perusteella erityisesti `EFFORT1`, `EFFORT2` ja `LANGN` nousevat esiin varsin vakaasti tärkeimpien muuttujien joukossa.

Puupohjaisissa malleissa keskeiset muuttujat painottuvat osittain eri tavalla kuin Elastic Net -mallissa. Random Forest- ja XGBoost-malleissa (kuvat 29–32, liite C) tärkeimmäksi muuttujaksi nousee `ST268Q04JA`, joka kuvaa oppilaan käsitystä matematiikan helppoudesta. Lisäksi useat tieto- ja viestintätekniikan käyttöön liittyvät `IC`-muuttujat sijoittuvat korkealle molemmissa malleissa. Kodin kulttuurisia resursseja kuvaava `ST256Q03JA` sekä koulutustavoitteisiin liittyvät muuttujat, kuten `ST327Q02JA` ja `EXPECEDU`, ovat myös keskeisessä asemassa erityisesti XGBoost-mallissa. Motivaatiota kuvaavat `EFFORT1`- ja `EFFORT2`-muuttujat esiintyvät edelleen tärkeimpien joukossa, mutta niiden suhteellinen merkitys jää pienemmäksi kuin Elastic Net -mallissa. Kuplakuvaajien perusteella puupohjaiset mallit tunnistavat varsin johdonmukaisesti samat keskeiset muuttujat eri ajoissa. Selitysvoima jakautuu kuitenkin useamman muuttujan kesken, eikä yksittäinen muuttuja dominoi yhtä voimakkaasti kuin Elastic Net -mallissa. Erityisesti tieto- ja viestintätekniikan käyttöön sekä koulutustavoitteisiin liittyvät muuttujat muodostavat vakaasti toistuvan ryhmän.

Yhteenvetona luonnontieteiden tuloksista voidaan todeta, että motivaatio, oppimisympäristö sekä kodin resurssit ovat keskeisiä osaamiseen liittyviä tekijöitä kuten muissakin aihealueissa. Elastic Net nostaa tärkeäksi erityisesti vaivannäön ja motivaation, kun taas puupohjaiset mallit painottavat enemmän tieto- ja viestintätekniikan käyttöä, oppilaan kokemuksia sekä koulutustavoitteita. Tulokset viittaavat

siihen, että luonnontieteiden osaaminen rakentuu useiden toisiinsa liittyvien tekijöiden yhteisvaikutuksesta, jossa sekä yksilölliset asenteet että ympäristötekijät ovat merkittävässä roolissa.

6.4 Kansainvälinen vertailu: Ruotsi ja Yhdysvallat

Seuraavaksi tarkastellaan lyhyesti matematiikan tuloksiin liittyviä keskeisiä tekijöitä kansainvälisessä vertailussa. Vertailuun Suomen kanssa valitaan Ruotsi ja Yhdysvallat, joista ensimmäinen edustaa koulutusjärjestelmältään ja yhteiskunnalliselta rakenteeltaan Suomea muistuttavaa kontekstia, kun taas jälkimmäinen tarjoaa vertailukohdan selvästi erilaisesta ympäristöstä.

Vertailu rajataan Elastic Net- ja XGBoost -malleihin, jotka osoittautuivat aiemmassa tarkastelussa parhaiten suoriutuviksi. Näiden mallien avulla pyritään tarkastelemaan, missä määrin matematiikan osaamiseen liittyvät keskeiset muuttujat ovat samankaltaisia eri maiden välillä ja missä määrin niiden merkitys vaihtelee kontekstin mukaan.

Tavoitteena ei ole tehdä kattavaa maiden välistä analyysia, vaan tuoda esiin yleisellä tasolla, kuinka keskeiset selittävät tekijät sijoittuvat erilaisissa koulutus- ja yhteiskuntaympäristöissä.

Ruotsin tulokset osoittavat, että Elastic Net- ja XGBoost -mallit tuottavat osittain samankaltaisen, mutta myös selvästi erilaisia painotuksia sisältävän kuvan matematiikan osaamiseen liittyvistä tekijöistä. Kuvista 13 ja 15 nähdään, että keskeiset muuttujaryhmät ovat osin yhteisiä, mutta yksittäisten muuttujien tärkeys vaihtelee mallien välillä.

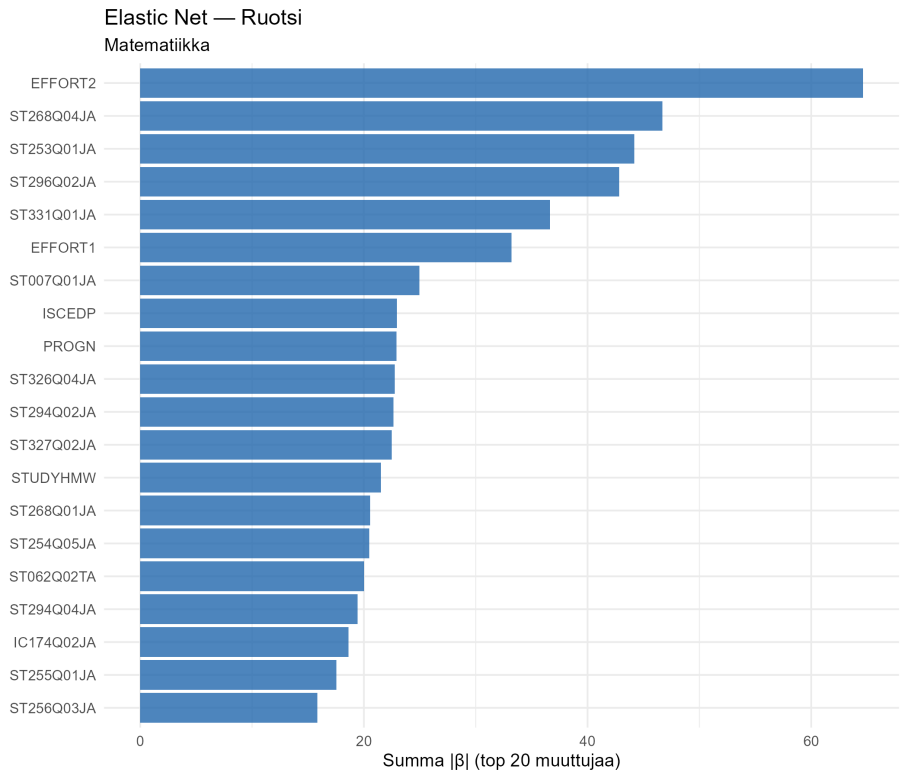
Elastic Net -mallissa selvästi merkittävimmäksi nousee `EFFORT2`, mikä korostaa oppilaan vaivannäön ja työskentelyasenteen keskeistä roolia. Lisäksi useat muuttujat, kuten `ST268Q04JA` (matematiikan koettu helppous), `ST253Q01JA` (digitaalisten laitteiden määrä kotona) ja `ST296Q02JA` (kotitehtäviin käytetty aika), sijoittuvat korkealle. Myös `EFFORT1` kuuluu keskeisten muuttujien joukkoon. Tämä rakenne muistuttaa Suomen tuloksia, joissa motivaatiota ja ponnistelua kuvaavat muuttujat olivat keskeisessä asemassa lineaarisessa mallissa. Sen sijaan maahanmuuttotaustaa kuvaavien muuttujien rooli näyttäytyy Ruotsissa heikompana kuin Suomessa.

XGBoost -mallissa painotus on selvästi erilainen. Tärkeimmäksi muuttujaksi nousee `ST268Q04JA`, joka kuvaa oppilaan käsitystä matematiikan helppoudesta. Lisäksi useat samaan muuttujasarjaan kuuluvat asenne- ja minäpystyvyysmuuttujat sekä `IC`-alkuiset tieto- ja viestintätekniikan käyttöä kuvaavat muuttujat painottu-

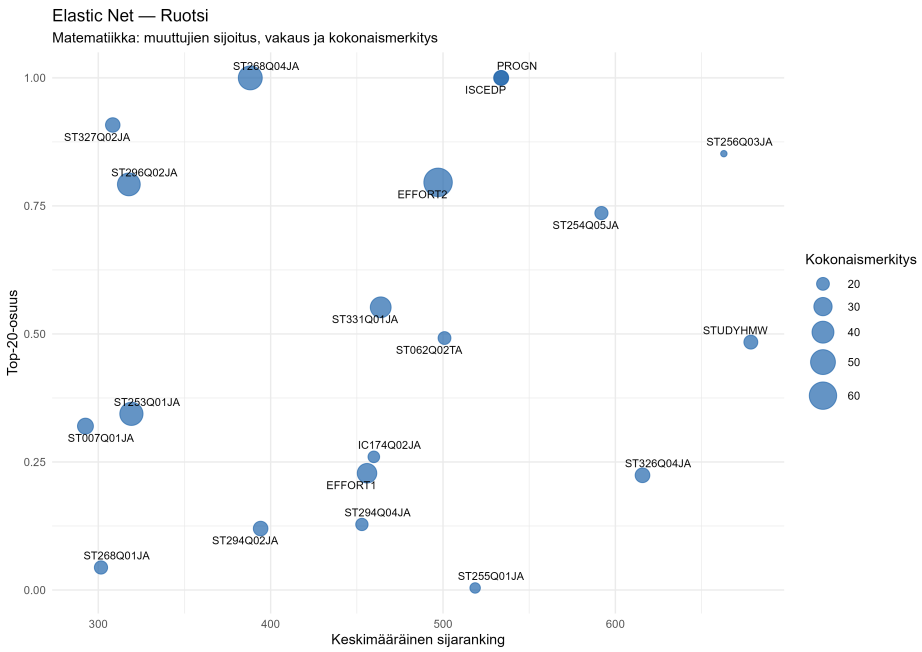
vat selvästi enemmän kuin Elastic Net -mallissa. Epälineaarinen malli tunnistaa erityisesti oppilaan kokemuksiin ja toimintaympäristöön liittyviä tekijöitä, jotka eivät nouse yhtä vahvasti esiin lineaarisessa mallissa. Tältä osin tulokset ovat hyvin samansuuntaisia Suomen tulosten kanssa.

Kuplakuvaajista (kuvat 14 ja 16) nähdään lisäksi, että vaikka osa muuttujista on yhteisiä, niiden vakaus ja sijoittuminen vaihtelevat. Elastic Net -mallissa `EFFORT2` on merkittävä ja melko vakaa muuttuja, mutta sen vakaus on hieman heikompi kuin Suomessa. XGBoost -mallissa puolestaan useat minäpystyvyyteen liittyvät muuttujat sijoittuvat erittäin vakaasti tärkeimpien muuttujien joukkoon. Tämä tuo esiin mallien välistä eroa siinä, millaisia rakenteita ne aineistosta tunnistavat.

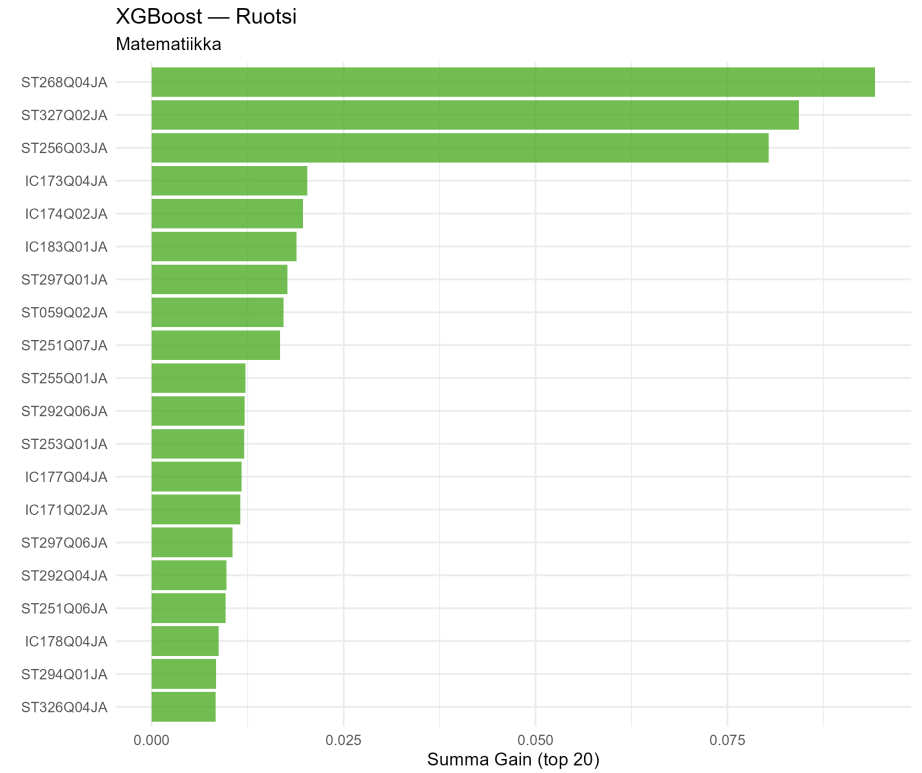
Suomeen verrattuna keskeiset muuttujaryhmät ovat edelleen samankaltaisia, mutta niiden suhteellinen merkitys vaihtelee. Erityisesti maahanmuuttotaustan vaikutus näyttäytyy Ruotsissa pienempänä, ja myös ponnistelua kuvaavat muuttujat ovat hieman vähemmän vakaita Elastic Net -mallissa. Sen sijaan XGBoost -mallin tulokset ovat hyvin lähellä Suomen vastaavia tuloksia. Kokonaisuutena tulokset viittaavat siihen, että matematiikan osaamisen taustalla vaikuttavat keskeiset ilmiöt ovat samansuuntaisia molemmissa maissa, mutta niiden painotus ja mallinnuksessa korostuvat rakenteet vaihtelevat jonkin verran kontekstin mukaan.



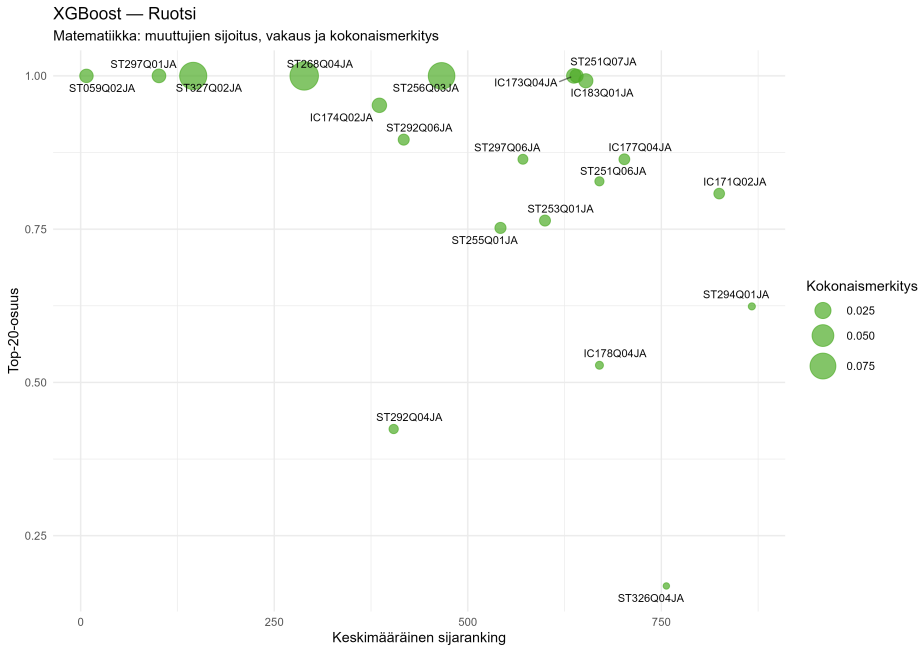
Kuva 13: Elastic Net -mallin tärkeimmät muuttujat matematiikassa Ruotsissa.



Kuva 14: Elastic Net -mallin muuttujien sijoitus, vakaus ja kokonaismerkitys matematiikassa Ruotsissa.



Kuva 15: XGBoost -mallin tärkeimmät muuttujat matematiikassa Ruotsissa.



Kuva 16: XGBoost -mallin muuttujien sijoitus, vakaus ja kokonaismerkitys matematiikassa Ruotsissa.

Yhdysvaltojen tulokset osoittavat, että Elastic Net- ja XGBoost -mallit tuottavat osittain samankaltaisen, mutta myös selvästi erilaisia painotuksia sisältävän kuvan matematiikan osaamiseen liittyvistä tekijöistä. Kuvista 17 ja 19 nähdään, että keskeiset muuttujaryhmät ovat osin yhteisiä, mutta niiden keskinäinen merkitys vaihtelee mallien välillä.

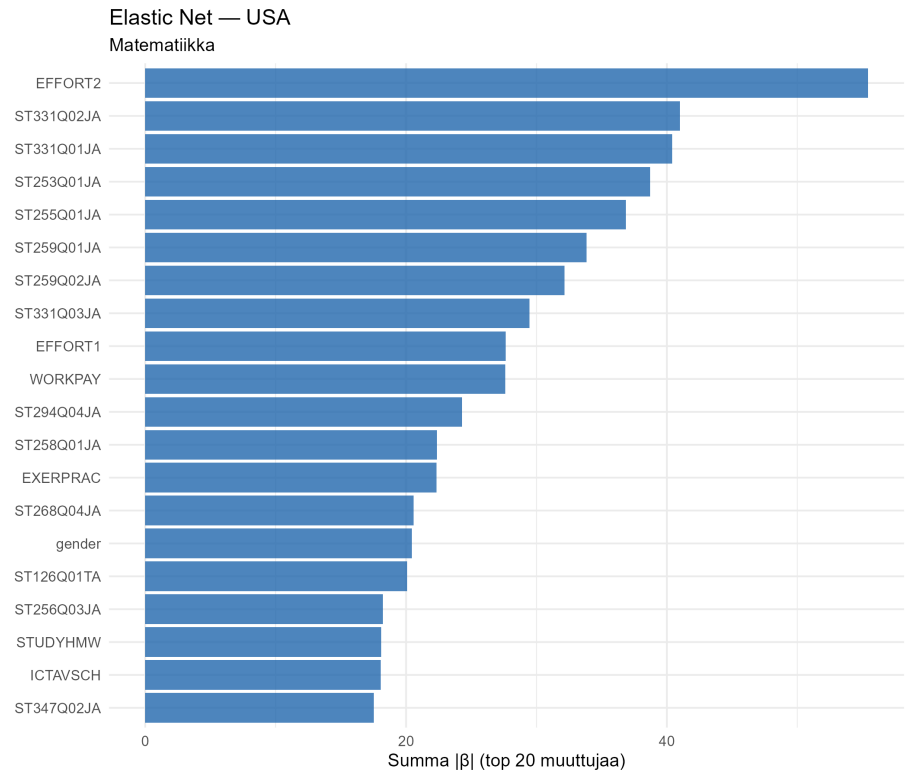
Elastic Net -mallissa tärkeimmäksi muuttujaksi nousee `EFFORT2`, mikä korostaa oppilaan ponnistelun ja työskentelyasenteen keskeistä roolia myös Yhdysvalloissa. Lisäksi muut vaivannäköä kuvaavat muuttujat, kuten sarjan `ST331` muuttujat, jotka kuvaavat ponnistelua antaa täsmälliset vastaukset ja `ST253Q01JA` (digitaalisten laitteiden määrä kotona), sijoittuvat korkealle, mikä viittaa asenteiden, oppimiseen liittyvien kokemusten sekä kodin resurssien merkitykseen. Myös `EFFORT1` kuuluu keskeisten muuttujien joukkoon, vaikka sen vakaus on kuplakuvaajan perusteella selvästi heikompi kuin `EFFORT2`:n. Kokonaisuutena Elastic Net -mallin rakenne muistuttaa Suomen tuloksia, joissa motivaatio ja ponnistelu olivat keskeisessä asemassa, mutta Yhdysvalloissa niiden rinnalle nousee enemmän muita tekijöitä kuten sukupuoli ja muuttujat `ST294Q04JA` ja `WORKPAY`, jotka kuvaavat työntekoa koulun ohella.

XGBoost -mallissa painotus on kuitenkin selvästi erilainen. Merkittävin muuttuja on `ST059Q02JA` (oppituntien kokonaismäärä), jonka merkitys on huomattavasti suurempi kuin muiden muuttujien. Lisäksi `ST268Q04JA` (matematiikan koettu help-

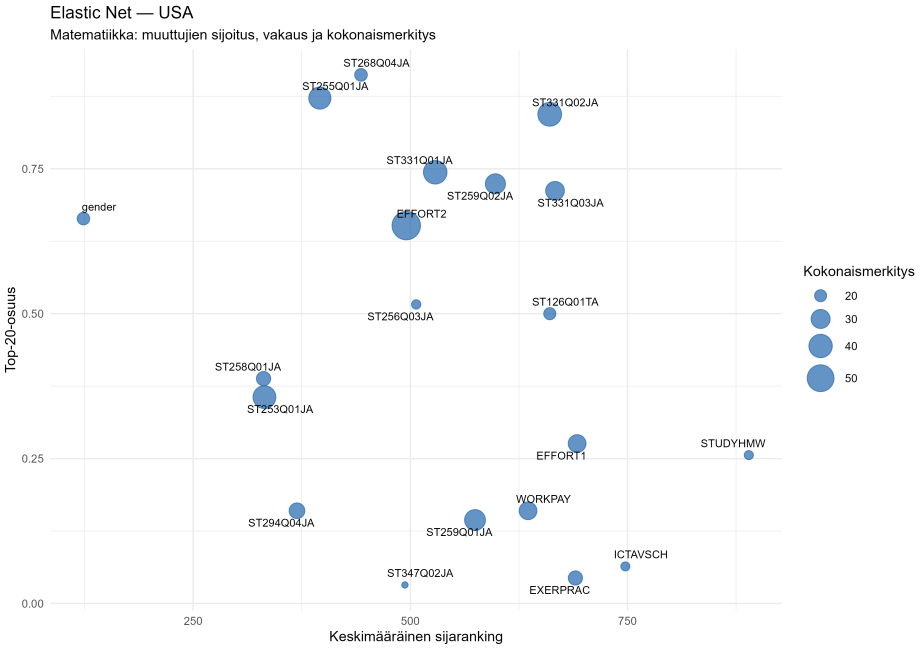
pous) sijoittuu korkealle. Huomionarvoista on, että useat FL-alkuiset muuttujat, jotka kuvaavat oppilaan talousosaamista ja taloudellisia kokemuksia, kuten yrittäjyyteen liittyvää tietoa, rahankäyttöä (esim. lahjarahat) sekä talousaiheiden käsittelyä koulussa, ovat keskeisessä asemassa. Lisäksi useat IC-alkuiset muuttujat ovat tärkeitä, mikä viittaa tieto- ja viestintätekniikan käyttöön ja oppimisympäristöön liittyvien tekijöiden merkitykseen. Kokonaisuutena tulokset viittaavat siihen, että matematiikan osaaminen kytkeytyy Yhdysvalloissa vahvasti oppilaan koulunkäyntiin liittyviin rakenteellisiin tekijöihin, kuten opetuksen määrään, sekä arjen toimintaan, kuten taloudellisiin kokemuksiin ja digitaalisiin käytäntöihin.

Kuplakuvaajista (kuvat 18 ja 20) nähdään lisäksi, että muuttujien vakaus vaihtelee mallien välillä. Elastic Net -mallissa EFFORT2 on keskeinen ja melko vakaa muuttuja, mutta monet muut muuttujat esiintyvät selvästi epävakaammin, mikä viittaa siihen, että selitysvoima jakautuu useamman tekijän kesken. XGBoost -mallissa puolestaan keskeiset muuttujat, kuten ST059Q02JA, esiintyvät erittäin vakaasti tärkeimpien joukossa, ja lisäksi lukemiseen ja tieto- ja viestintätekniikan käyttöön liittyvät muuttujat muodostavat johdonmukaisen ryhmän.

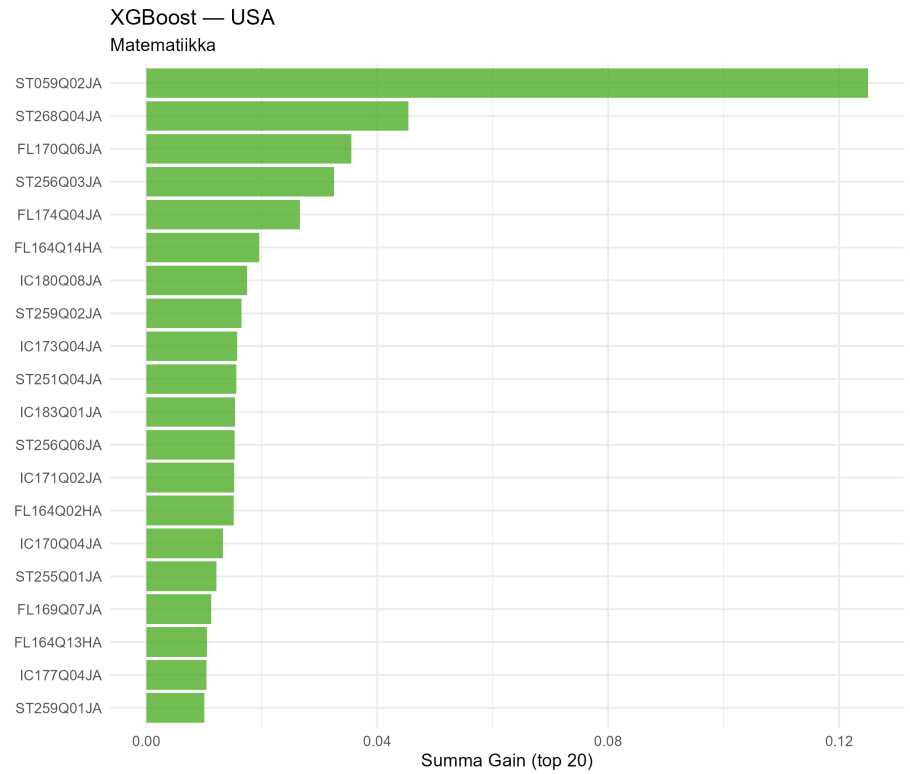
Suomeen verrattuna keskeinen ero on se, että vaikka Elastic Net -mallissa motivaatio painottuu myös Yhdysvalloissa, XGBoost -mallissa painopiste siirtyy selvästi koulunkäyntiin ja arjen toimintaan liittyviin tekijöihin. Erityisesti oppituntien määrää kuvaava muuttuja nousee Yhdysvalloissa keskeiseen rooliin, kun taas Suomessa vastaavat rakenteelliset tekijät eivät nouse esiin yhtä selvästi. Tämä voi viitata siihen, että opetuksen määrä vaihtelee Yhdysvalloissa enemmän oppilaiden välillä, kun taas Suomessa koulutuksen rakenne on yhtenäisempi. Suomen tuloksissa motivaatiomuuttujat olivat keskeisiä molemmissa mallityypeissä, kun taas Yhdysvalloissa mallien välinen ero on selvästi suurempi. Tämä vahvistaa havaintoa, että matematiikan osaamisen taustalla olevat ilmiöt ovat osittain samoja, mutta niiden keskinäiset suhteet ja mallinnuksessa korostuvat rakenteet vaihtelevat kontekstin mukaan.



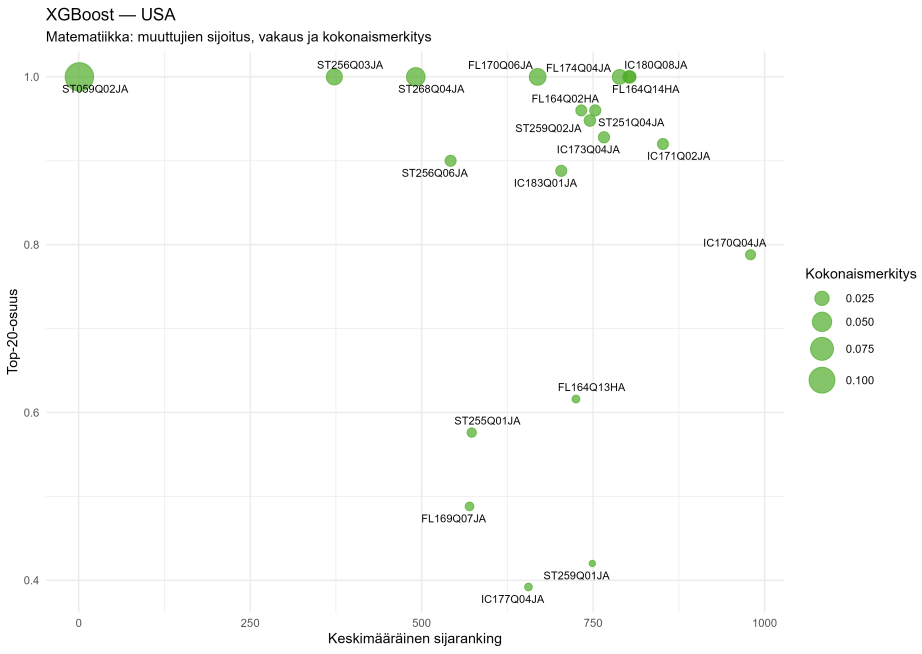
Kuva 17: Elastic Net -mallin tärkeimmät muuttujat matematiikassa Yhdysvalloissa.



Kuva 18: Elastic Net -mallin muuttujien sijoitus, vakaus ja kokonaismerkitys matematiikassa Yhdysvalloissa.



Kuva 19: XGBoost -mallin tärkeimmät muuttujat matematiikassa Yhdysvalloissa.



Kuva 20: XGBoost -mallin muuttujien sijoitus, vakaus ja kokonaismerkitys matematiikassa Yhdysvalloissa.

7 Päätelmät ja pohdintaa

Tässä työssä tarkasteltiin PISA 2022 -aineiston avulla, mitkä tekijät ovat yhteydessä oppilaiden matematiikan, lukutaidon ja luonnontieteiden osaamiseen, sekä sitä, miten eri mallinnusmenetelmät kuvaavat näitä yhteyksiä. Analyysissä verrattiin lineaarista Elastic Net -mallia sekä kahta puupohjaista menetelmää, Random Forestia ja XGBoostia, ja lisäksi tarkasteltiin tulosten eroja Suomen, Ruotsin ja Yhdysvaltojen välillä.

Tulosten perusteella eri mallien välillä havaittiin selkeitä eroja sekä ennustetarkkuudessa että siinä, millaisia muuttujia ne korostavat. Testiaineiston perusteella Elastic Net suoriutui kokonaisuutena parhaiten, ja myös XGBoost tuotti varsin kilpailukykyisiä tuloksia. Random Forest jäi sen sijaan selvästi heikoimmaksi.

Kun tuloksia verrataan opetusaineiston suorituskykyyn, havaitaan kuitenkin merkittäviä eroja mallien yleistettävyyskyvyssä. Erityisesti Random Forest ja XGBoost saavuttivat erittäin korkean selitysasteen opetusaineistossa, mutta niiden suorituskyky heikkeni selvästi testiaineistossa. Tämä viittaa ylisovittamiseen, joka oli voimakkainta Random Forest -mallissa. Elastic Net -mallissa ero opetus- ja testiaineiston välillä oli selvästi pienempi, mikä viittaa parempaan yleistettävyyteen. Puupohjaisten mallien ylisovittaminen liittyy niiden suureen joustavuuteen. Tässä työssä käytetyissä malleissa erityisesti Random Forest kärsi tästä. Minimisolmukoon hakutila oli suhteellisen pieni (`min_node_size` = 5–40), ja erityisesti pienin arvo mahdollistaa hyvin pienten havaintoryhmien muodostumisen. Tällöin puut voivat kasvaa syviksi ja sovittaa yksittäisiä havaintoja tarkasti, mikä lisää riskiä oppia aineiston kohinaa todellisen signaalin sijaan. XGBoostissa ylisovittaminen oli vähäisempää, mutta sitä esiintyi silti. Mallissa ei esimerkiksi rajoitettu puiden syvyyttä kovin tiukasti, eikä oppimisnopeutta tai regularisaatiota kasvatettu maksimaalisesti. Tämä mahdollistaa monimutkaisten rakenteiden oppimisen, mutta samalla lisää riskiä, että malli sovittaa myös satunnaisvaihtelua. Vaikka hyperparametrit valittiin sisäkkäisellä ristiinvalidoinnilla, valintakriteerinä ollut ennustetarkkuus voi suosia hieman monimutkaisempia malleja. Tämän vuoksi lopulliset mallit eivät välttämättä ole yleistettävyyden kannalta optimaalisimpia. Ylisovittamista olisi voitu vähentää esimerkiksi kasvattamalla minimisolmukokoa, rajoittamalla puiden syvyyttä tai lisäämällä regularisaatiota, mikä olisi pakottanut mallit yksinkertaisempaan rakenteeseen.

Muuttujien näkökulmasta tulokset osoittivat, että keskeiset oppimistuloksiin liittyvät tekijät ovat pitkälti samankaltaisia eri malleissa, mutta niiden painotus vaih-

telee. Motivaatio ja ponnistelu (EFFORT-muuttujat) nousivat erityisesti Elastic Net -mallissa keskeisiksi kaikilla osa-alueilla. Lisäksi maahanmuuttotaustaa kuvaavat muuttujat (kuten LANGN sekä COBN-muuttujat) tulivat esiin erityisesti Elastic Net -mallissa. Tämä liittyy osittain siihen, että samaa ilmiötä kuvaa useampi keskenään korreloiva muuttuja. Lineaarinen malli voi antaa näille muuttujille samanaikaisesti painoa, jolloin ilmiö näkyy selkeämmin kokonaisuutena. Puupohjaisissa malleissa tilanne on erilainen, sillä yksittäinen jako perustuu kerrallaan vain yhteen muuttujaan, jolloin keskenään korreloivista muuttujista valitaan tyypillisesti vain yksi. Tämän seurauksena maahanmuuttotaustan vaikutus jakautuu usean muuttujan kesken eikä yksittäinen muuttuja nouse yhtä selvästi esiin. Puupohjaiset mallit painottivat sen sijaan enemmän oppilaan kokemuksiin ja oppimisympäristöön liittyviä tekijöitä, kuten käsitystä matematiikan helppoudesta, tieto- ja viestintätekniikan käyttöä ja koulutusodotuksia. Tämä viittaa siihen, että ei-lineaariset menetelmät tunnistavat laajempia ja mahdollisesti monimutkaisempia yhteyksiä, joita lineaarinen malli ei kuvaa yhtä vahvasti.

Kansainvälisessä vertailussa havaittiin, että Suomi ja Ruotsi tuottavat hyvin samankaltaisen kuvan oppimistuloksiin liittyvistä tekijöistä. Molemmissa maissa motivaatio ja ponnistelu olivat keskeisiä muuttujia, ja mallien väliset erot olivat suhteellisen pieniä. Yhdysvalloissa tilanne oli erilainen, vaikka Elastic Net -mallissa motivaatio säilyi keskeisenä tekijänä. XGBoost -mallissa korostuivat selvästi oppituntien määrä, talousosaamiseen liittyvät muuttujat sekä arjen toimintaan ja oppimisympäristöön liittyvät tekijät. Tämä kertoo siitä, että Yhdysvalloissa oppimistulokset kytkeytyvät vahvemmin koulunkäynnin rakenteisiin ja oppilaan arjen kokemuksiin, kun taas Suomessa ja Ruotsissa ilmiö näyttäytyy tasaisempana ja yhtenäisempänä.

Työn keskeinen havainto on, että eri mallinnusmenetelmät eivät tuota täysin samaa kuvaa ilmiöstä. Elastic Net tarjoaa selkeämmän ja vakaamman tulkinnan keskeisistä tekijöistä, kun taas puupohjaiset mallit tuovat esiin laajempia ja monimutkaisempia yhteyksiä, mutta samalla niiden tulokset ovat alttiimpia ylisovittamiselle. Tämä tarkoittaa, että tulosten tulkinnassa on tärkeää huomioida käytetty menetelmä eikä tarkastella yksittäisen mallin tuloksia irrallaan.

Työhön liittyy myös useita rajoitteita. Ensinnäkin analyysi perustuu poikkileikkausaineistoon, joten tulokset kuvaavat yhteyksiä eivätkä syy-seuraussuhteita. Lisäksi aineisto ei sisällä ajallista ulottuvuutta, joten oppimistulosten kehitystä ajan suhteen ei voida tarkastella. Toiseksi monet keskeiset muuttujat, kuten motivaatio ja ponnistelu, perustuvat itsearviointiin, mikä voi sisältää systemaattista harhaa. Lisäksi muuttujien suuri määrä ja niiden välinen korrelaatio vaikeuttavat mallinnus-

ta ja lisäävät ylisovittamisen riskiä. Myös aineiston esikäsittely, kuten puuttuvien arvojen perusteella tehty karsinta, voi vaikuttaa tuloksiin. Lisäksi kansainvälinen vertailu rajattiin vain kahteen maahan Suomen lisäksi, joten tuloksia ei voida yleistää laajemmin kaikkiin koulutusjärjestelmiin. Eri maiden välillä voi olla merkittäviä rakenteellisia eroja, joita tässä tarkastelussa ei täysin pystytä huomioimaan.

Kokonaisuutena tulokset viittaavat siihen, että oppimistulosten taustalla olevat keskeiset ilmiöt ovat pitkälti samankaltaisia eri konteksteissa, mutta niiden keskinäiset suhteet ja merkitys vaihtelevat. Samalla työ korostaa, että mallinnusmenetelmän valinnalla on merkittävä vaikutus siihen, millainen kuva aineistosta muodostuu. Erityisesti puupohjaisten menetelmien kohdalla on tärkeää kiinnittää huomiota ylisovittamisen hallintaan, jotta tulokset säilyvät luotettavina ja yleistettävinä.

Viitteet

1. OECD. *PISA 2022 Results (Volume I): The State of Learning and Equity in Education* Luettu 5.3.2025. https://doi.org/10.1787/53f23881-en (OECD Publishing, Paris, 2023).
2. OECD. *PISA 2022 Technical Report* Luettu 11.2.2025. https://doi.org/10.1787/01820d6d-en (OECD Publishing, Paris, 2024).
3. OECD. *PISA 2022 Results (Volume I and II) — Country Notes: Finland* Luettu 4.2.2026 (Organisation for Economic Co-operation ja Development, 2023). https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/11/pisa-2022-results-volume-i-and-ii-country-notes_2fca04b9/finland_ac04e1eb/6991e849-en.pdf.
4. OECD. *PISA 2022 Dataset* Luettu 10.2.2026 (Organisation for Economic Co-operation ja Development, 2023). https://webfs.oecd.org/pisa2022/index.html.
5. OECD. *PISA 2022 Codebook* Luettu 11.2.2026 (Organisation for Economic Co-operation ja Development, 2023). https://www.oecd.org/en/data/datasets/pisa-2022-database.html.
6. OECD. *PISA Data Analysis Manual: SPSS, Second Edition* Luettu 5.3.2026. https://doi.org/10.1787/9789264056275-en (OECD Publishing, Paris, 2009).
7. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning: with Applications in R* 2. painos. Luettu 4.3.2026. https://www.statlearning.com/ (Springer, New York, NY, 2021).
8. Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12.** Luettu 26.1.2026, 55–67. https://doi.org/10.1080/00401706.1970.10488634 (1970).
9. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58.** Luettu 26.1.2026, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x (1996).
10. Zou, H. & Hastie, T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67.** Luettu 26.1.2026, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x (2005).

11. Breiman, L. Random Forests. *Machine Learning* **45.** Luettu 29.10.2025, 5–32. https://doi.org/10.1023/A:1010933404324 (2001).

12. Chen, T. & Guestrin, C. *XGBoost: A Scalable Tree Boosting System* teoksessa *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* Luettu 23.10.2025 (Association for Computing Machinery, New York, NY, 2016), 785–794. https://doi.org/10.1145/2939672.2939785.

13. XGBoost Developers. *XGBoost Parameters* Luettu 3.3.2026. https://xgboost.readthedocs.io/en/stable/parameter.html.

14. Hyndman, R. J. & Koehler, A. B. Another look at measures of forecast accuracy. *International Journal of Forecasting* **22.** Luettu 04.03.2026, 679–688. https://robjhyndman.com/publications/another-look-at-measures-of-forecast-accuracy/ (2006).

15. Cawley, G. C. & Talbot, N. L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research* **11.** Luettu 5.3.2026, 2079–2107. https://jmlr.org/papers/v11/cawley10a.html (2010).

16. Wainer, J. & Cawley, G. C. Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications* **182.** Luettu 3.3.2026, 115222. https://doi.org/10.1016/j.eswa.2021.115222 (2021).

17. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33.** Luettu 3.3.2026, 1–22. https://doi.org/10.18637/jss.v033.i01 (2010).

18. Probst, P., Wright, M. N. & Boulesteix, A.-L. Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery* **9.** Luettu 3.3.2026, e1301. https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1301 (2019).

19. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **77.** Luettu 5.3.2026, 1–17. https://doi.org/10.18637/jss.v077.i01 (2017).

20. Bergstra, J. & Bengio, Y. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* **13.** Luettu 3.3.2026, 281–305. https://jmlr.org/papers/v13/bergstra12a.html (2012).

21. Robitzsch, A. & Lüdtke, O. On the Treatment of Missing Item Responses in Educational Large-Scale Assessment Data: An Illustrative Simulation Study and a Case Study Using PISA 2018 Mathematics Data. *European Journal of Investigations in Health, Psychology and Education* **11.** Luettu 5.3.2026, 1653–1687. https://doi.org/10.3390/ejihpe11040117 (2021).

Liitteet

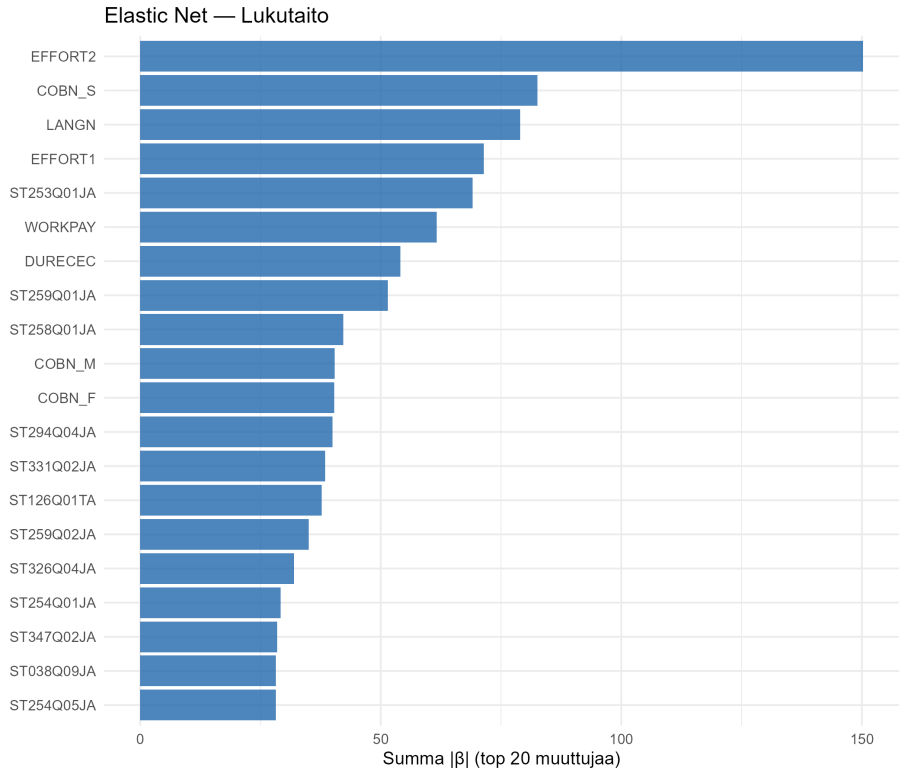
A Muuttujien selitykset (Excel-tiedosto)

Avaa Excel-liite

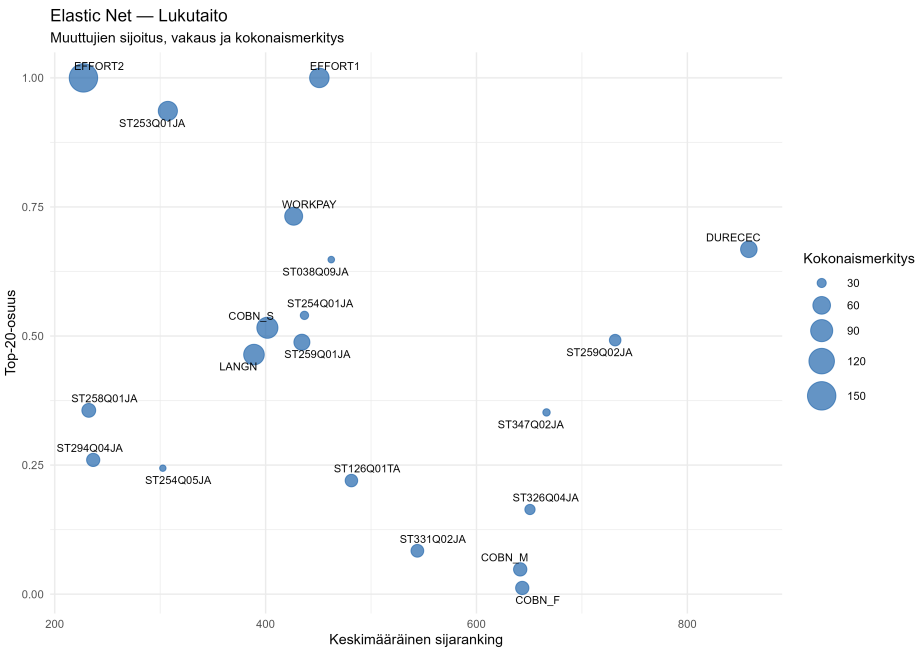
B Koodit (Github)

Avaa Github-sivu

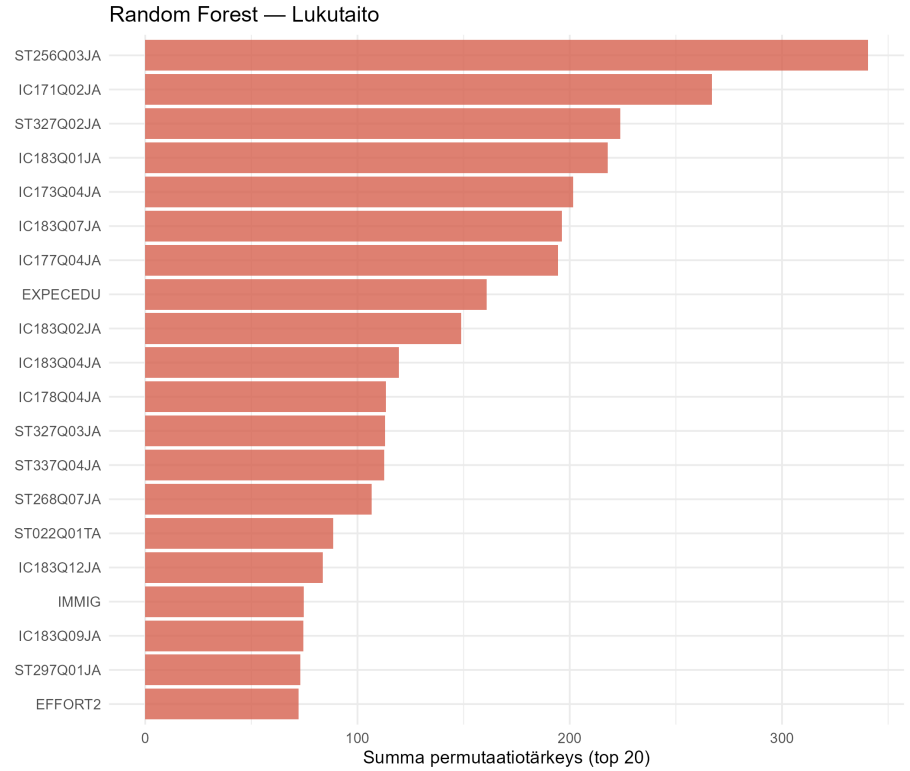
C Kuvat



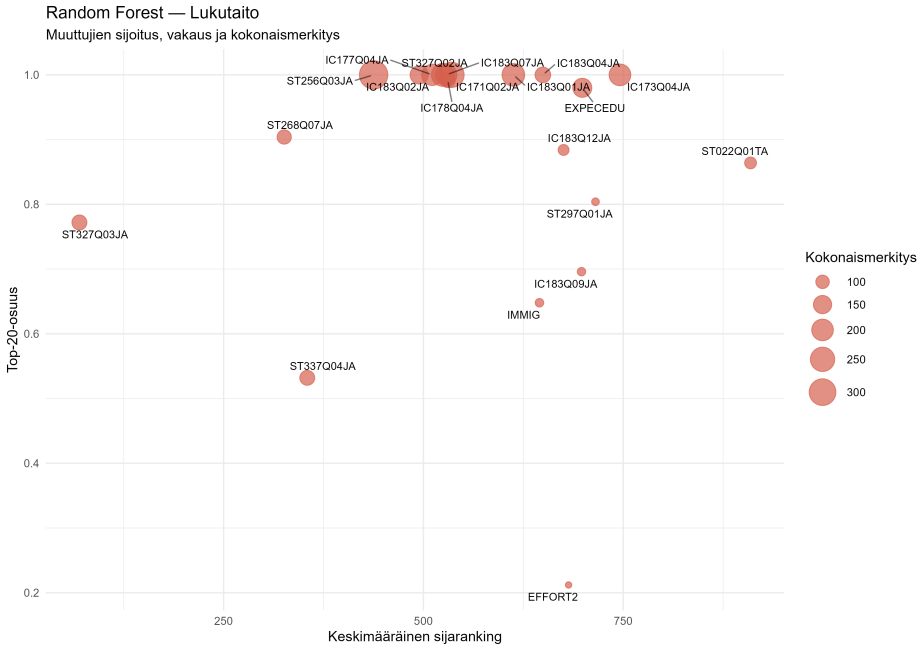
Kuva 21: Elastic Net -mallin tärkeimmät muuttujat lukutaidossa.



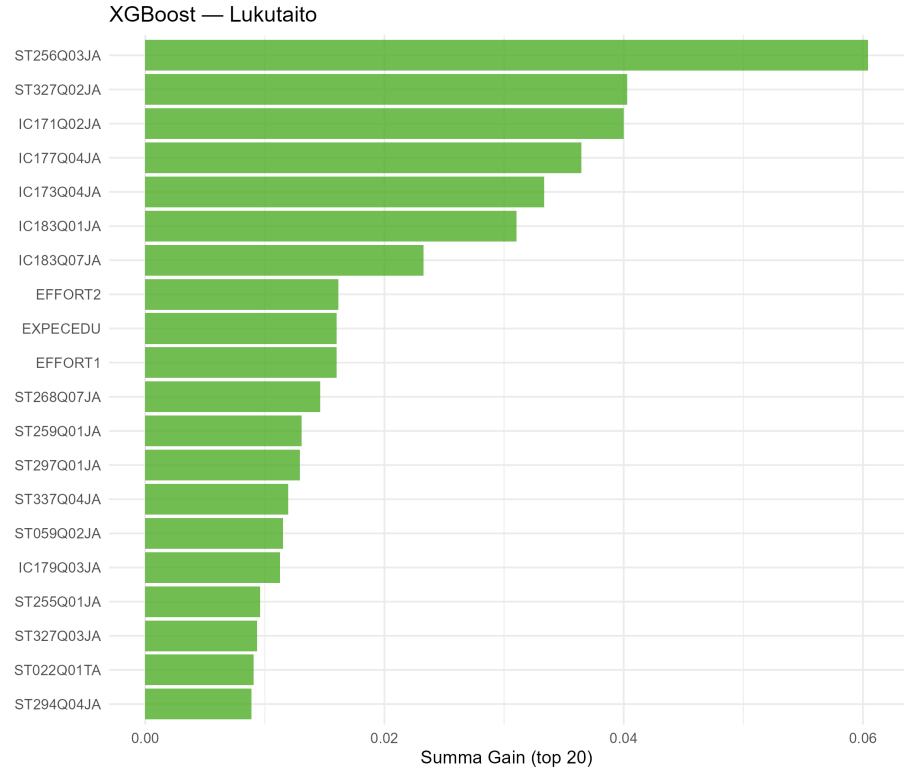
Kuva 22: Elastic Net -mallin muuttujien sijoitus, vakaus ja kokonaismerkitys lukutaidossa.



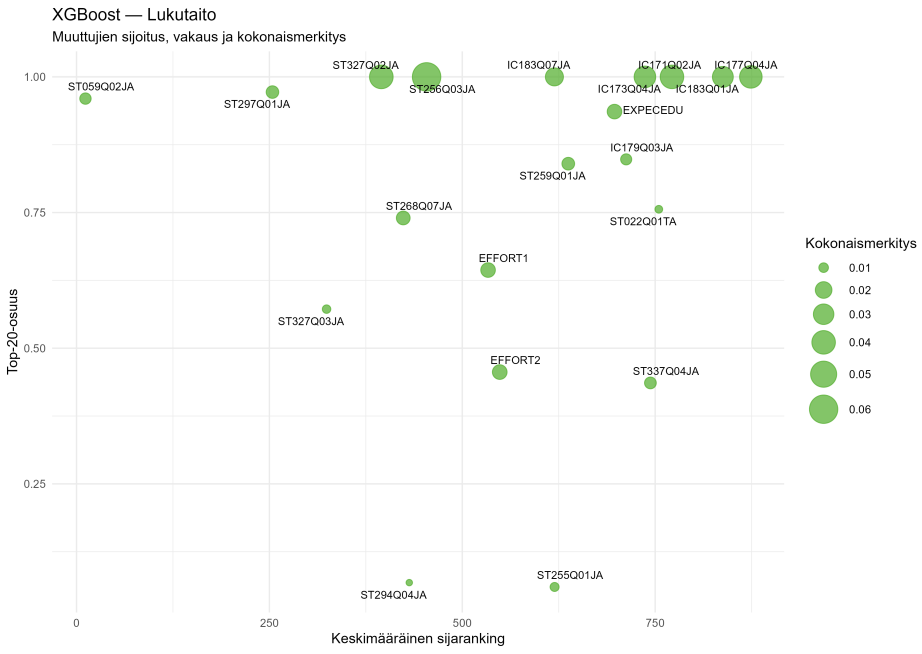
Kuva 23: Random Forest -mallin tärkeimmät muuttujat lukutaidossa.



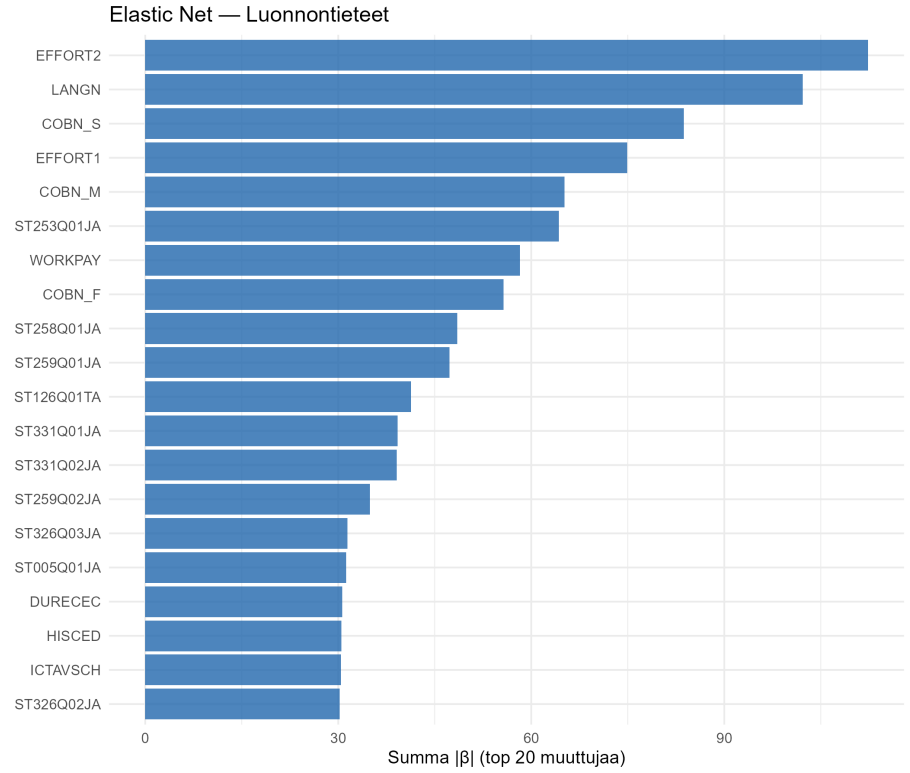
Kuva 24: Random Forest -mallin muuttujien sijoitus, vakaus ja kokonaismerkitys lukutaidossa.



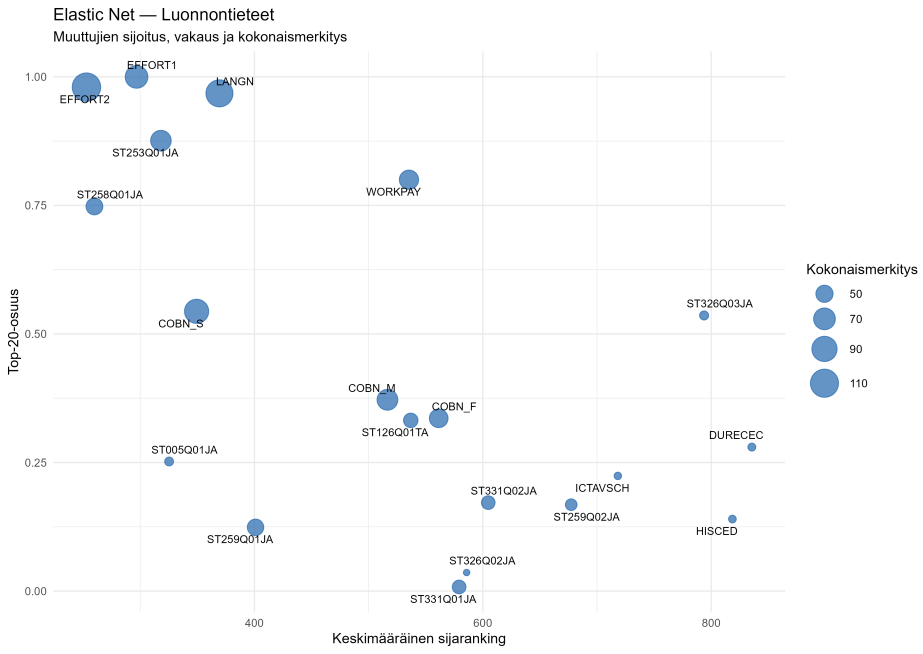
Kuva 25: XGBoost -mallin tärkeimmät muuttujat lukutaidossa.



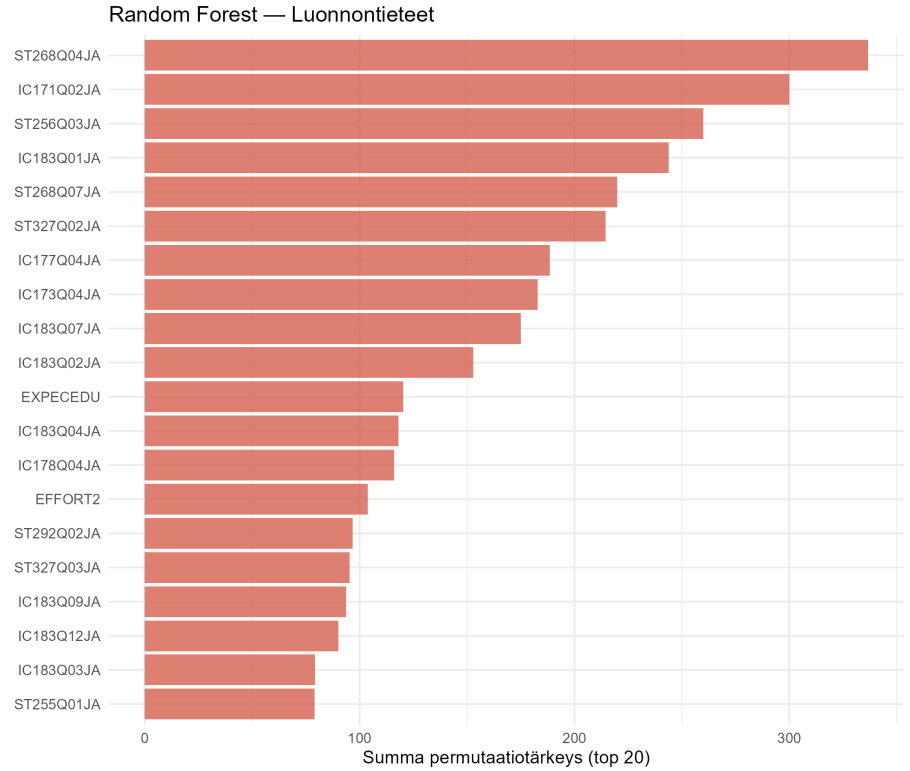
Kuva 26: XGBoost -mallin muuttujien sijoitus, vakaus ja kokonaismerkitys lukutaidossa.



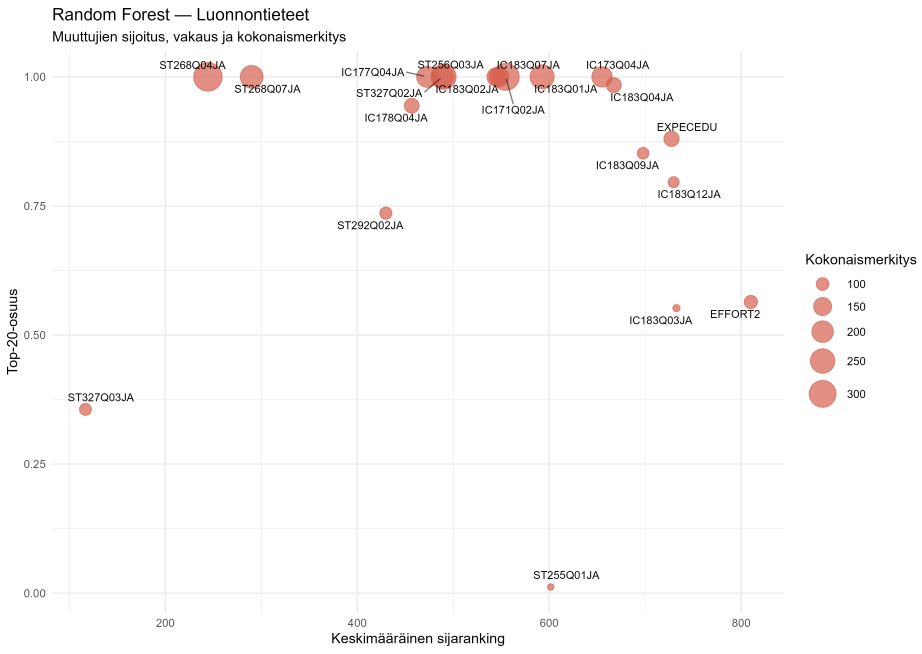
Kuva 27: Elastic Net -mallin tärkeimmät muuttujat luonnontieteissä.



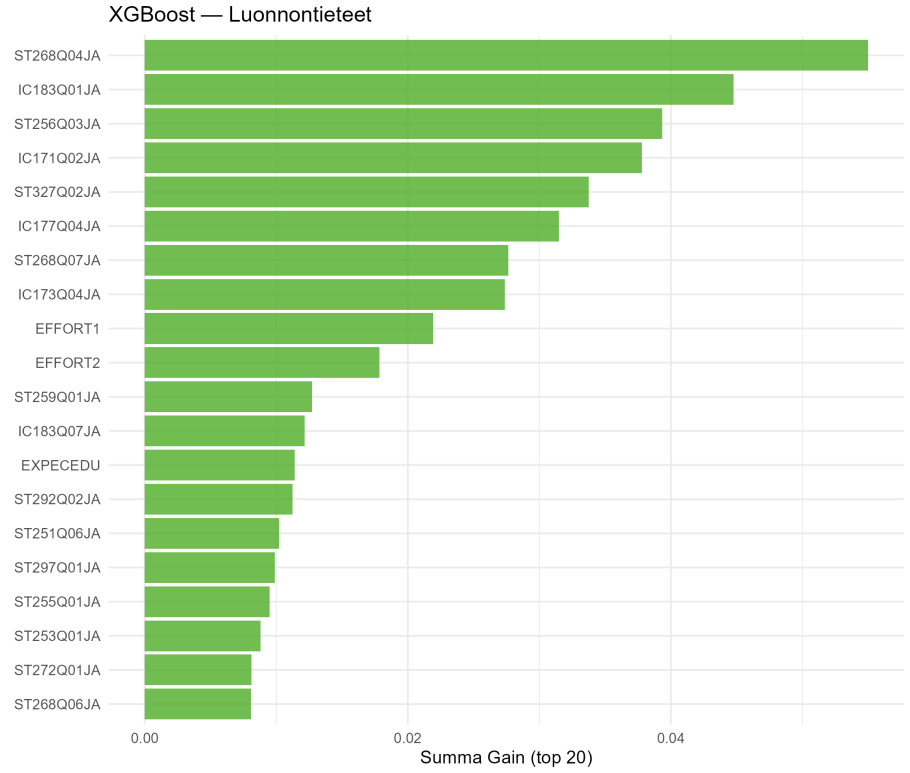
Kuva 28: Elastic Net -mallin muuttujien sijoitus, vakaus ja kokonaismerkitys luonnontieteissä.



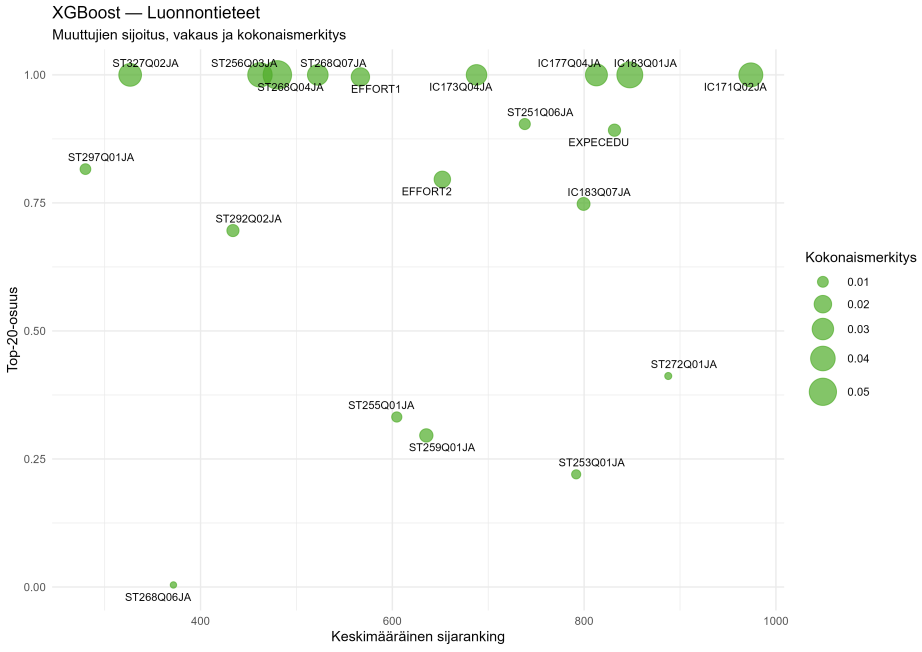
Kuva 29: Random Forest -mallin tärkeimmät muuttujat luonnontieteissä.



Kuva 30: Random Forest -mallin muuttujien sijoitus, vakaus ja kokonaismerkitys luonnontieteissä.



Kuva 31: XGBoost -mallin tärkeimmät muuttujat luonnontieteissä.



Kuva 32: XGBoost -mallin muuttujien sijoitus, vakaus ja kokonaismerkitys luonnontieteissä.