Maisterintutkielma

Tietojenkäsittelytieteen maisteriohjelma

Pudokkaiden automatisoitu tunnistaminen

Sanna Korpi

8.3.2025

Matemaattis-luonnontieteellinen tiedekunta

Helsingin yliopisto

Yhteystiedot

PL 68 (Pietari Kalmin katu 5)
00014 Helsingin yliopisto

Sähkopostiosoite: info@cs.helsinki.fi
URL: http://www.cs.helsinki.fi/

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Study programme
Matemaattis-luonnontieteellinen tiedekunta	Tietojenkäsittelytieteen maisteriohjelma

Tekijä — Författare — Author

Sanna Korpi

Työn nimi — Arbetets titel — Title

Pudokkaiden automatisoitu tunnistaminen

Ohjaajat — Handledare — Supervisors

FT Kjell Lemström, FT Matti Luukkainen

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Maisterintutkielma	8.3.2025	58 + 35 sivua

Tiivistelmä — Referat — Abstract

Tässä työssä tutkitaan koneoppimismenetelmin, miten opintojen keskeytyminen, eli ”pudokkuus”, näkyy Helsingin yliopiston kandiohjelman opiskelijoiden kurssisuoritus- ja ilmoittautumisaineistossa. Lisäksi koneoppimismalleilla pyritään ennustamaan, ketkä opiskelijat ovat vaarassa tulla pudokkaiksi. Aineistona tässä työssä on vuosina 2017-2023 opinto-oikeuden saaneiden opiskelijoiden em. aineistot sekä hieman taustatietoa opiskelijasta. Ennustamista varten aineistosta kootaan kolme eri versiota. Lisäksi jokainen aineisto katkaistaan ajallisesti useammasta eri kohdasta, jotta saadaan selville, missä kohtaa opiskeluita ennustaminen kannattaisi suorittaa. Jokaisella aineistolla koulutetaan kaksi koneoppimismallia: XGBoost-malli ja yksittäisen päätöspuu. Parhaiten pärjäävän mallin ympärille rakennetaan myös ohjelmisto, jolla ennuste on helppo suorittaa uudelle aineistolle. Mallin voi myös kouluttaa uudelleen valitsemillaan parametreillä.

XGBoost-mallit koulutettuna suurimman aineiston aliaineistoilla, eli 2017-2022 opinto-oikeuden saaneilla, suoriutuivat parhaiten. Pudokkuutta ennustavia tekijöitä olivat parhaiden mallien näkökulmasta mm. tarkasteluajankohdan loppupuolella saatujen opintopisteiden määrä, ja noin puoli vuotta ennen loppua saadun ajanjakson keski-arvo. Miessukupuoli ja korkeampi ikä olivat myös jossain määrin vaikuttavia tekijöitä todennäköisempään pudokkuuteen. Mallien luotettavuus oli hyvin matala, kun opintoja oli takana vasta yhden vuoden verran. Siitä eteenpäin tulokset kohenivat tasaisesti. Kun opintoja oli takana 1,5 tai 2,5 vuotta, malli pystyi saamaan kiinni 90% pudokkaista, joskin ”väärien hälytysten” osuus pudokkaiksi leimatuista oli 1,5 vuoden aineistolla jopa 41%. 2,5 vuoden aineistolla osuus putosi 34%:iin. Mallin kokonaistarkkuudeksi jäi 1,5 vuoden tapauksessa 0,68 ja 2,5 aineistolla 0,75. Epätarkoista tuloksista voidaan päätellä, ettei pudokkuus välttämättä näy opintomenestyksessä kaikissa tapauksissa.

ACM Computing Classification System (CCS)
Computing methodologies → Machine learning → Machine learning approaches → Classification and regression trees
General and reference → Cross-computing tools and techniques → Empirical studies

Avainsanat — Nyckelord — Keywords

Koneoppiminen, päätöspuut

Säilytyspaikka — Förvaringsställe — Where deposited

Helsingin yliopiston kirjasto

Muita tietoja — övriga uppgifter — Additional information

Algoritmien opintosuunta

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Study programme
Faculty of Science	Master's Programme in Computer Science

Tekijä — Författare — Author

Sanna Korpi

Työn nimi — Arbetets titel — Title

Automatic detection of dropouts

Ohjaajat — Handledare — Supervisors

Dr. Kjell Lemström, Dr. Matti Luukkainen

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's thesis	March 8, 2025	58 + 35 pages

Tiivistelmä — Referat — Abstract

In this study, machine learning methods are used to investigate how student dropouts are reflected in the course performance and registration data of students in the undergraduate programme at the University of Helsinki. Additionally, the aim is to predict which students are at risk of becoming dropouts. The data used in this study includes records of students who were granted study rights between 2017 and 2023, as well as some background information about the students. Three different versions of the data are compiled for prediction. Furthermore, each dataset is temporally truncated at several points to determine when it would be most effective to make predictions during a student's studies. Two machine learning models are trained on each dataset: an XGBoost model and a single decision tree.

The XGBoost models performed the best, when trained on the subdatasets of the largest dataset, i.e. students who were granted study rights between 2017 and 2022. Factors that predicts dropout, according to the best models, included the number of credits accumulated towards the end of the observation period, and the average number of credits earned around six months before the end of the observation period. Male gender and higher age were also somewhat influential factors for a higher likelihood of dropping out. The reliability of the models was very low when the students had studied only one year. After that, the results steadily improved. Once the students had completed 1.5 or 2.5 years of study, the model was able to identify 90% of the dropouts, although the proportion of false positives among those identified as potential dropouts was as high as 41% with the 1.5-year dataset. This proportion decreased to 34% with the 2.5-year dataset. The overall accuracy of the model was 0.68 for the 1.5-year dataset and 0.75 for the 2.5-year dataset. The inaccuracies suggest that dropping out may not always be reflected in academic performance in every case.

ACM Computing Classification System (CCS)
Computing methodologies → Machine learning → Machine learning approaches → Classification and regression trees
General and reference → Cross-computing tools and techniques → Empirical studies

Avainsanat — Nyckelord — Keywords

Machine learning, Decision trees

Säilytyspaikka — Förvaringsställe — Where deposited

Helsinki University Library

Muita tietoja — övriga uppgifter — Additional information

Algorithms study track

Sisällys

Tekoälyä on käytetty apuna tämän työn tekemisessä.

Tekoälyä on käytetty synonyymien etsimiseen ja ohjelmoinnissa ohjelmointikirjastojen syntaksin opetteluun hakukonehakujen sijasta.

1 Johdanto

Helsingin tietojenkäsittelytieteen kandiohjelmassa moni jättää opintonsa kesken. Opetushallinnon tilastopalvelu Vipusesta selviää, että tietojenkäsittelytieteen opiskelijoista keskeyttää useampi kuin muilta luonnontieteen aloilta. Syitä on lukemattomia erilaisia, mikä tekee varhaisen puuttumisen haastavaksi. Voi kuitenkin olla mahdollista, että ”pudokkuus” näkyy jollain tavalla opiskelijoiden toiminnassa syistä riippumatta jo hieman ennen opintojen keskeytymistä. Jos opintojen keskeytyminen näkyisi esimerkiksi kurssisuoritus- ja ilmoittautumisaineistossa riittävän aikaisin, jopa automatisoitu puuttuminen voisi olla mahdollista.

Tämän maisterintutkielman tavoitteena on tutkia, millä tapaa opintojen keskeytyminen näkyy kurssisuoritus- ja ilmoittautumisaineistossa ennen opintojen keskeytymistä, ja onko keskeytymistä mahdollista ennustaa koneoppimismenetelmin. Tavoitteena on myös luoda järjestelmä, joka ennustaa, onko opiskelija potentiaalinen opintojen keskeyttäjä tähän astisten kurssisuoristustensa ja ilmoittautumistensa perusteella.

Empiiriset tutkimukset antavat viitteitä siitä, että opiskelijan kokema integroituminen yliopistoon ja sen sosiaalisiin yhteisöihin on yksi tärkeimmistä tekijöistä keskeyttämisen kannalta [28]. Opintomenestyksen puolestaan katsotaan korreloivan jonkin verran integroitumisen kanssa [28]. Toisissa empiiririssä tutkimuksissa onkin havaittu, että opintomenestys on yksi suurimmista tekijöistä opintojen keskeytymisessä. Vaikka opintoemenestys ei suoraan olisikaan niin sanottu vaikuttava tekijä keskeyttämiseen, ainakin keskeyttämisvaara näkyy opintomenestyksessä useassa tutkimuksessa [3, 4, 11, 14, 21]. Tämän takia voidaan olettaa, että kurssisuoritus- ja ilmoittautumisaineisto on sopivaa materiaalia ennusteen tekemiseen.

Aihetta sivuavia aikaisempia tutkimuksia on olemassa. Koneoppimismenetelmin on yritetty ennustaa opiskelijoiden opintojen keskeyttämistä useissa eri tutkimuksissa, mutta suuressa osassa aineisto sisälsi myös merkittävän määrän taustatietoja opiskelijoista. Helsingin yliopiston tapauksessa käytettävissä on vain anonymisoitua kurssisuoritus- ja ilmoittautumisaineistoa sekä minimaaliset taustatiedot opiskelijasta.

Nykyisistä sovellutuksista ei myöskään ole kovin käytännönläheisiä ratkaisuja olemassa. Chengin et al. mukaan tähän astiset mallit ovat ongelmallisia, koska ne vaativat niin paljon prosessointitehoa ja -aikaa, sekä ovat vaikeita ottaa käyttöön käytännössä, kun dataa

pitää manuaalisesti käsitellä paljon ennen kuin ennusteen voi tehdä [3]. Oman ongelmansa tuo myös datan kerääminen. Käytännössä kurssidata lienee ainoa tietolähde, joka on yliopiston vapaasti käytettävissä tällaiseen tarkoitukseen, vaikka joissain tutkimuksissa olikin käytetty jopa klikkausvirtadataa, kuten Nagrecha et al. tekivät [19].

Aihetta on siis käsitelty tutkimuksissa, mutta ei täysin samanlaisella aineistolla kuin Helsingin yliopistolla on saatavilla.

Tässä työssä tutkitaan ensin, minkälainen on sopiva pudokkaan määritelmä käytössä olevalle aineistolle. Kun pudokas on määritelty, tarkastellaan yleisesti minkälaisia muuttujia aineisto tarjoaa, ja mitkä mahdollisesti voisivat selittää pudokkuutta. Tämän jälkeen valitaan aineistoon sopivat koneoppimismallit ja suoritetaan malleilla useampi koe, jossa aineisto on käsitelty hieman eri tavoin. Malleilla pyritään ennustamaan, onko opiskelija tuleva pudokas vai ei. Mallien avulla suoritetaan myös tiedonlouhintaa, eli selvitetään, mitkä tekijät mallien mukaan selittävät pudokkuutta.

2 Tausta

Tässä osiossa selvitetään, miten opintojen keskeytymistä on tutkittu koneoppimisen näkökulmasta muissa tutkimuksissa. Tehdään pieni katsaus siihen, mitä koneoppiminen on, miten pudokkuus on tutkimuksissa määritelty, ja minkälaisia pudokkuuteen vaikuttavia tekijöitä tutkimuksissa on löytynyt. Lisäksi tarkastellaan lähemmin, mitä koneoppimismalleja on käytetty ja miten aineistoja on käsitelty ennen malleille syöttämistä.

Opintojen keskeytymistä on tutkittu koneoppimismenetelmin pääosin kahdella tavalla: joko on yritetty vain löytää tekijöitä, jotka ennustavat opintojen keskeytymistä (tiedonlouhinta) tai sitten on pyritty rakentamaan malli, jolla voi ennustaa onko tietty opiskelija keskeyttämisvaarassa. Jotkut tutkimukset ovat pyrkineet molempiin tavoitteisiin. Viitteet on eritelty taulukossa 2.1.

Vaikuttavia tekijöitä on yleisimmin etsitty Bayes-verkoilla. Ennustemalleihin puolestaan on sovellettu enimmäksen päätöspuita, gradienttitehostettuja päätöspuita, logistista regressiota ja Bayes-verkkoja. Myös neuroverkkoja on käytetty. Eri tutkimuksissa käytettyjä malleja on listattu taulukkoon 2.1. Otoskoot ovat vaihdelleet eri tutkimuksissa merkittävästi, ja se on vaikuttanut myös mallin valintaan. Esimerkiksi neuroverkkoja on käytetty vain, kun otoskoko on ollut suuri, kun taas päätöspuita ja logistista regressiota on käytetty myös pienempiin otoksiin. Viitteet on eritelty taulukossa 2.1.

2.1 Koneoppiminen lyhyesti

Koneoppimisella tarkoitetaan menetelmiä, joissa opetusaineiston perustella määritellään ohjelmallisesti sääntöjä, joita pystytään myöhemmin soveltamaan uuteen vastaavaan aineistoon [6]. Sääntöjen avulla aineisto voidaan jakaa eri luokkiin, tai määrittää arvo muttujalle, jonka arvoa ei tiedetä. Menetelmät pyritään kehittämään niin, että niitä on helppo soveltaa yleisesti erilaisiin aineistoihin. Oppiminen on siis ohjelmallinen tapa löytää automaattisesti rakenteita ja toistuvuutta aineistoista. Opitut säännöt voivat olla erittäin monimutkaisia algoritmejä, tai jotain hyvin yksinkertaista, kuten ensimmäisen asteen yhtälö, johon aineiston havainnot sovitetaan [8].

Koneoppimista käytetään hyväksi useimmiten silloin, kun aineistosta ei ole suoraan löy-

Malli	Aineiston koko	Tiedonlouhinta	Ennustaminen	Molemmat
Päätöspuu	802 - 62 375	[17]	[27]	[4, 1, 21, 11, 28]
Neuroverkot	9 195 - 248 000		[27, 3, 26, 17, 28]	
Logistinen regressio	802 - 24 770		[26]	[4, 21, 11, 9]
XGBoost	331 - 24 770	[3]		[18, 13, 9]
Bayes-verkot	383 - 62 375	[14, 17]	[27]	
Satunnaismetsä	331 - 16 807		[26]	[18, 4]
Naiivi Bayes	802-6800			[21, 28]
Tukivektorikone	16 807		[26]	
RIPPER-algoritmi	6800			[28]
CatBoost	331			[18]

Taulukko 2.1: Tutkimuksissa käytetyt koneoppimismallit ja aineistojen koot (opiskelijaa aineistossa) viitteineen. Taulukossa mainittua mallia on käytetty tutkimuksessa joko tiedonlouhintaan, pudokkuuden ennustamiseen tai molempiin. Taulukko on järjestetty siten, että ylimpänä on tutkimuksissa eniten käytetyt mallit.

dettävissä tekijöitä, joiden perusteella se voitaisiin jakaa eri luokkiin, tai määrittää jollekin muuttujalle arvoa [6]. Opittujen sääntöjen soveltamista uuteen aineistoon kutsutaan yleensä ennustamiseksi. Algoritmia, jolla sääntöjä sovelletaan aineistoon, kutsutaan puolestaan malliksi.

Koneoppimismenetelmät voidaan jakaa karkeasti ohjattuihin ja ohjaamattomiin menetelmiin [6]. Ohjatuissa menetelmissä on ennalta tiedossa esimerkkejä mahdollisista luokista tai muuttujan arvoista, joita havainnoilla voi olla ja joita mallilla halutaan myöhemmin ennustaa. Malli opetetaan näiden esimerkkien avulla. Tiedossa voi esimerkiksi olla, mikä pilvisten päivien määrä on ollut missäkin kuussa. Tällöin aineiston jokainen rivi vastaa jotakin kuukautta ja pilvisten päivien määrää tuossa kuussa. Malli opetetaan näillä tiedoilla. Opettamisen jälkeen opetetulle mallille annetaan vain pilvisten päivien määrä, ja malli käyttää opittuja sääntöjä sen ennustamiseen, mikä kuukausi mahdollisesti on kyseessä. Ohjaamattomassa oppimisessa taas ei etukäteen kerrota mallille, mitä sen pitäisi aineistosta löytää. Tällainen malli voi olla esimerkiksi luokittelija, joka ryhmittelee aineistoa eri luokkiin löytämiensä rakenteiden mukaisesti. Lopputulos voi tällöin olla esimerkiksi havaintojen ryhmittely sellaisten ominaisuusyhdistelmien perusteella, jotka yllättävät tutkijan, ja tuovat esiin uutta tietoa aineistosta.

Käytännössä koneoppimismalleja käytetään yleensä valmiiksi ohjelmoitujen kirjastojen välityksellä. Hieman mallista riippuen, kirjastot sallivat useiden eri hyperparametrien säätä-

misen. Mallin suoritukseen vaikuttavia käyttäjän asettamia parametrejä kutsutaan hyperparametreiksi ja niiden säätämisellä voi olla erittäin suuri vaikutus mallin suorituskykyyn [22]. Kirjastoihin on säädetty valmiiksi yleisimpiin tapauksiin sopivat oletushyperparametrit, mutta ne eivät useimmiten tarjoa parasta suorituskykyä, koska jokainen aineisto on yksilöllinen.

2.2 Mikä on pudokas?

Pudokas on yleisen käsityksen mukaan opiskelija, joka keskeyttää opintonsa eikä valmistu siitä koulutusohjelmasta, jossa opiskeli. Tässä työssä pudokas täytyy määritellä hieman eri tavalla, koska ei ole saatavilla riittävää määrää aineistoa, joka käsittäisi opiskelijan suoritukset koko ajalta, jolloin opinto-oikeus on voimassa. Suomen yliopistoissa opinto-oikeus on oletuksena 7 vuotta kandidaatin ja maisterin tutkintoon yhteensä, mutta tähän voi hakea lisäaikaa. Tämän vuoksi ei voida myöskään olla koskaan täysin varmoja, jääkö opiskelijan opinnot varmasti kesken, vaikka opinto-oikeus olisikin kulunut loppuun.

Eri tutkimuksissa pudokkaan määrittely vaihtelee suuresti. Tan ja Shao määrittelivät tilastoanalyysin perusteella pudokkaaksi opiskelijan, joka ei osallistunut kahteen lukukauteen lukukauden päättökokeisiin, koska sillä oli vahva korrelaatio valmistumattomuuden kanssa [27]. Toisaalta taas Solis et al. käyttivät kahta eri määritelmää; ensimmäisenä ”opiskelijat, jotka eivät ole kahteen lukukauteen ilmoittautuneet kursseille”, ja toisena ”kaikki ei-valmistuneet” [26]. Jälkimmäinen määritelmä sisälsi siis jopa tällä hetkellä aktiiviset opiskelijat, jotka eivät ole valmistuneita. Kemper et al. puolestaan määrittelivät pudokkaan sellaiseksi, joka ei ollut valmistunut opintoaikanaan [11]. Heillä oli käytössään laajat aineistot koko opiskeluajalta. Miranda et al. taas määrittelivät pudokkaaksi opiskelijan, joka oli epäaktiivisena kolme vuotta [17]. Perez et al. tutkimuksessa pudokas oli opiskelija, joka ei valmistunut kuudessa vuodessa edes kandidaatiksi [21]. Yhteistä kaikille määritelmille oli kuitenkin se, että ne liittyivät siihen, onko opiskelija tehnyt suorituksia tai ilmoittautumisia – aikaväli vain vaihteli. Tässä työssä sopivaa aikaväliä etsitään analysoimalla käytössä olevaa aineistoa.

2.3 Pudokkuutta ennustavat tekijät

Osassa tutkimuksia käytettiin suoraan koneoppimismenetelmiä pudokkuuden ennustamiseen, eikä raportoitu tarkemmin, mitkä tekijät lopulta osoittautuivat tärkeiksi ennusteen

muodostamisessa [27, 26]. Toiset tutkimukset taas keskittyivät enemmän nimenomaan etsimään pudokkuutta ennustavia tekijöitä sen sijaan, että olisivat yrittäneet saada aikaan mahdollisimman tarkan ennustuksen keskeytysvaarasta [17, 14]. Tässä aliluvussa kerrotaan tarkemmin, miten pudokkuutta ennustavia tekijöitä on selvitetty ja minkälaisia tuloksia tutkimuksissa on saatu.

Miranda ja Guzmán selvittivät tiedonlouhinnalla, mitkä tekijät ennustivat opintojen keskeytymistä heidän aineistossaan [17]. Aineisto koostui chileläisen yliopiston opiskelijoiden kurssisuoritustiedoista, ja heidän toisen asteen koulutuksen (lukiota vastaava koulutustaso) arvosanoistaan. Miranda ja Guzmán käyttivät kolmea eri luokittelijaa: Bayes-verkkoja, päätöspuita ja monikerroksista perseptroniverkkoa (neuroverkkoa). Toisen asteen keskiarvo osoittautui yhdeksi tärkeimmistä ennustavista tekijöistä opintojen keskeytymisen kannalta. Perez et al. puolestaan ennustivat pudokkuutta yksinkertaisilla malleilla: päätöspuulla, logistisella regressiolla ja Naivi-Bayes-mallilla [21]. Heidän aineistossaan tärkeimmiksi tekijöiksi nousivat opiskeltujen lukukausien määrä, arvosanakeskiarvo, uudelleenilmoittautumisten määrä ja kumulatiivinen arvosanakeskiarvo. Hieman yllättävästi kurssien uudelleenottamisen vähäinen määrä ennusti pudokkuutta, eikä suuri määrä, kuten voisi odottaa.

Costa et al. ennustivat opintojen keskeytymistä kolmen ensimmäisen lukukauden opintomenestyksen perusteella kolmella eri mallilla; päätöspuulla, logistisella regressiolla ja satunnaismetsällä [4]. Heidän aineistostaan nousi tärkeimmäiksi akateemiseksi tekijäksi kolmannen lukukauden keskiarvo, ja sosiaalisista tekijöistä opiskelijan saamat hyödyt, kuten stipendit. Aineisto sisälsi stipenditietojen ja lukukausien keskiarvojen lisäksi myös opintojen aloitusajankohdan, iän aloittaessa, opintomäärät lukukausittain, kokonaiskeskiarvon ja kokonaisopintomäärän. Myös Huo et al. aineistosta tärkeimmiksi muuttujiksi XGBoost-mallilla nousivat taloudelliset tekijät, mutta myös lukukausi-ilmoittautumisen tyyppi: oliko opiskelija osa-aikainen ja ilmoittautunut vuoden molemmille lukukausille vai ei [9].

Lacave et al. [14] tutkivat opintojen keskeytymisen syitä Bayes-verkoilla. Heidän aineistonsa Castilla-La Manchan yliopistosta koostui melko laajoista opiskelijoiden taustatiedoista sekä opintojen suoritustiedoista. Lopulta tärkeimmiksi tekijöiksi osoittautuivat suoritettujen opintojen määrä ja se, miten pitkälle opinnoissa oli edetty. Jos näitä tietoja ei ollut saatavilla, keskeyttämistä ennustivat parhaiten pääsykokeen arvosana sekä uusittujen tenttien määrä. Tyypilliseksi pudokkaaksi osoittautui miespuolinen 18-vuotias opiskelija, joka on saanut stipendin ja valinnut koulutusohjelman ykkösvaihtoehdokseen, mutta on

pärjännyt pääsykokeessa huonosti, ja ei ole suorittanut yhtään kurssia, vaikka on ilmoittautunut 6-10 kurssille. Lacave et al. arvelevat kyseen olevan siitä, että suuri osa pudokkaista keskeyttää opintonsa ennen ensimmäisiä lopputenttejä ja tutkimuksessa se näkyy näin [14].

Abu-Oda et al. [1] lähestyivät ongelmaa eri tavalla kuin muut. He käsittelivät aineistonsa siten, että jokaisesta opiskelijasta oli yksi tietue, joka sisälsi opintomenestyksen kursseittain. Kursseiksi oli valittu keskeiset 7 tietojenkäsittelytieteen kurssia. Lisäksi aineistossa oli tietoa opiskelijan sukupuolesta, lukioaikaisesta keskiarvosta, asuinalueesta (etelä/pohjoinen) ja opintojen keskiarvosta. He käyttivät koneoppimismenetelmänään päätöspuuta ja saivat tulokseksi, että kaksi kurssia nousivat muita tärkeämmäksi keskeyttämisen ennustamisessa. Kursseilla ”digital design” ja ”algorithm analysis” hyvin pärjänneet keskeyttivät epätodennäköisemmin kuin muut. Myös sukupuolella oli väliä siinä, johtiko huono arvosana kurssista lopulta keskeyttämiseen. Pojat keskeyttivät todennäköisemmin saatuaan huonon arvosanan jommasta kummasta kurssista. Myös Moreira da Silva et al. lähestyivät pudokkuutta kurssikohtaisen aineiston kautta, ja saivat yhtä lailla tulokseksi, että muutamalla kurssilla oli selkeästi enemmän tekemistä pudokkuuden kanssa kuin toisilla [18]. Toisin kuin missään muussa tutkimuksessa, heidän kokeissaan ikä oli kuitenkin yksi tärkeimmistä tekijöistä pudokkuuden ennustamisessa. He käyttivät gradienttitehostettuja menetelmiä sekä satunnaismetsää.

Kemper et al. [11] ennustivat saksalaisen yliopiston tenttiaineistolla opintojen keskeytymistä. Aineistossa oli useita tietueita per opiskelija. Yksi tietue esitti yhden tentin tuloksia. Tietue sisälsi kuitenkin myös samalla koostettua opintomenestysdataa, kuten siihen astisen keskiarvon, hyväksyttyjen ja hylättyjen tenttien lukumäärät, tehtyjen tenttien keskiarvot, ja lisäksi hieman taustatietoa, kuten sukupuoli, kansalaisuus, ilmoittautumisaika ja ikä. He käyttivät logistista regressiota ja päätöspuuta sen vuoksi, että halusivat saada samalla selville tekijät, jotka ennustavat opintojen keskeytymistä. He määrittelivät pudokkaan ei-valmistuneeksi, ja jättivät aineistosta pois kaikki ne, jotka olivat edelleen aktiivisia opinnoissaan. Lopulta molemmat mallit antoivat saman suuntaisia tuloksia; tärkeimmäksi ennustavaksi tekijäksi muodostui tenttien keskiarvosta ja hyväksyttyjen tenttien määrästä muodostettu yhteismuuttuja. Koordinaatistoon piirrettynä muuttuja antaakin jo heti ymmärtää, että se erottelee keskeyttävät opiskelijat hyvin heidän aineistossaan.

Kaiken kaikkiaan yleisimmät tekijät, joilla oli suurin vaikutus opintojen keskeytymisen ennustamiseen, liittyivät opintojen keskiarvoihin tai suoritusten lukumääriin. Eri tutkimuksissa ilmenneitä vaikuttavia tekijöitä on eritelty tarkemmin taulukossa 2.2.

Vaikuttava tekijä	Mallit	Viitteet
Opintojen keskiarvo	Päätösp., log. regr., naiivi-Bayes, s.metsä, RIPPER, XGBoost	[4, 21, 11] [28, 9]
Hyväksytyt kurssit lkm	Log. regr., päätösp., Bayes-ver., XGBoost	[11, 14, 3]
Lukukausien määrä	Päätösp., log. regr., naiivi-Bayes	[21, 14]
Aikaisempi opintomenestys	Päätösp., Bayes-ver.	[17, 14]
Uudelleenilm. lkm	Päätösp., log. regr., Naivi-Bayes, Bayes-ver.	[21, 14]
Tietyn kurssin arvosana	Päätösp., XGBoost, CatBoost	[1, 18]
Kurssin ajoitus	XGBoost	[3]
Hylätyt kurssit lkm	Log. regr., päätöspuu	[11]
Lukukausi ilm. tyyppi	XGBoost	[9]
Opiskelijan ikä	XGBoost, CatBoost, s.metsä	[18]

Taulukko 2.2: Keskeyttämistä ennustavat tekijät eri tutkimuksissa. Taulukko on järjestetty siten, että ylimpänä on yleisimmät tutkimuksissa löytyneet vaikuttavat tekijät.

2.4 Aineiston käsittely

Tutkimuksissa on ollut hyvin erilaisia ja eri kokoisia aineistoja käytössä. Yleisin tapa on kuitenkin ollut, että yhtä opiskelijaa kohden on yksi tietue, johon on kerätty koostettuja tietoja hänen opintosuorituksistaan, mahdollisten henkilötietojen lisäksi. Joissain tutkimuksissa on kuitenkin käytetty myös suoraan kurssiaineistovirtaa, jossa yksi tietue on vastannut yhtä kurssisuoritusta, eli yhtä opiskelijaa kohden on ollut useampi tietue [1, 3, 18].

Tan ja Shao ennustivat opintojen keskeytymistä laajalla aineistolla ja kolmella eri koneoppimismallilla [27]. Aineisto sisälsi paljon tietoa opiskelijoiden taustoista, jopa poliittisesta suuntautumisesta ja opintomenestyksestä sekä nykyisissä että aikaisemmissa opinnoissa. Jokaista opiskelijaa kohden oli vain yksi tietue, eli ei aikasarjaa. Malleista käytössä oli päätöspuut, neuroverkot ja Bayes-verkot. Cheng et al. puolestaan käyttivät myös aikasarjatyyppistä aineistoa, joka ei sisältänyt opiskelijasta taustatietoja [3]. He pyrkivät ennustamaan takaisinkytkeytyvällä neuroverkolla (engl. recurrent neural network), keskeyttääkö opiskelija opintonsa seuraavan lukukauden aikana. Heidän aineistonsa yksi tietue sisälsi yhden opiskelijan kurssi-ilmoittautumisen tiedot, sekä paljon tietoa kurssista itsestään, esimerkiksi kuka oli luennoitsija, ja montako kertaa opiskelija oli jo yrittänyt kurssia aiemmin. Eli yhtä opiskelijaa kohden aineistossa oli useita tietueita, jotka muodostivat opis-

kelijan kurssi-ilmoittautumisista aikasarjan. He saivat kuitenkin lopulta parempia tuloksia ilman aikasarjaa, käyttämällä koostettua opintomenestysdataa, kuten pääosassa aihealueen tutkimuksissa on tehty. Tässä aineistossa yksi tietue sisälsi yhden opiskelijan siihen astisen opintomenestyksen tietoja, kuten keskiarvon, suoritetun kurssimäärän, viimeikaisen suoritusmäärän ja viimeaikaisen keskiarvon.

Kemper et al. aineistosta selvisi, että saksalaisen yliopiston Karlsruhe Institute of Technology:n opiskelijoilla keskeytysuhka on suurin ensimmäisten kolmen lukukauden aikana [11]. Tämän vuoksi tarkasteluajanjakso ei voinut olla kovin pitkä, koska nämä keskeytysuhan alla olevat opiskelijat tulisi tunnistaa jo hyvin vähällä aineistolla. Toisaalta, keskeyttämistä tapahtui myös tasaiseen tahtiin alkupiikin jälkeen. Ennustusten tarkkuus nousi, mitä pidemmältä ajalta aineistoa on käytössä. Tästä huolimatta he rajoittivat tenttiaineistoa enimmillään kolmeen ensimmäiseen lukukauteen sen vuoksi, että heidän aineistossaan suurin keskeytymisuhka oli juuri ensimmäisten kolmen kuukauden aikana ja näin olleen ennusteen tarkkuutta tärkeämpi peruste. He koostivat kuusi eri aineistoa lähtöaineistosta. Aineistot käsittivät lukukaudet 1, 1-2, 1-3 ja kukin aineisto oli vielä kertaalleen joko tasapainotettu tai alkuperäinen. Tasapainotetulla tarkoitetataan sitä, että tällainen binäärinen aineisto pyritään saamaan sellaiseksi, että kumpiakin luokkia on yhtä paljon. Tässä tapauksessa siis pudokkaiden määrää tuli joko lisätä tai ei-pudokkaiden vähentää. He päätyivät koneellisesti lisäämään pudokkaita. Tämä tapahtui siten, että aineistosta eroteltiin tietueet, jotka koskivat pudokkaita ja koneellisesti arvottiin datapisteiden väliin uusia datapisteitä, eli täydennettiin aineistoa sellaisilla tietueilla, joiden esiintyminen aineistossa voisi olla todennäköistä. Tulosten perusteella tasapainottaminen kannatti. Tasapainotetuilla aineistoilla koulutetut malllit ennustivat pudokkaan tarkemmin. Toisaalta väärien positiviivisten ennustusten määrä nousi huomattavasti, eli malli herkemmin ennusti ei-pudokkaan pudokkaaksi kuin ei-tasapainotulla aineistolla koulutetut mallit. Tämä on silti todennäköisesti parempi lopputulos mallin käyttökohteen kannalta, koska ei-tasapainotetulla aineistolla koulutetut mallit puolestaan ennustivat virheellisesti monen pudokkaan ei-pudokkaaksi [11]. Muissa tutkimuksissa aineistoja ei juurikaan tasapainotettu. Toisaalta joissain aineistoissa pudokkaita olikin melkein yhtä paljon kuin ei-pudokkaita, joten se ei mahdollisesti olisi ollut tarpeenkaan.

Solis et al. [26] tutkivat eri lähestymistapoja keskeyttämisen ennustamiseen. He testasivat neljää eri koneoppimismallia: satunnaismetsää, neuroverkkoja, tukivektorikonetta ja logistista regressiota. Lisäksi he määrittelivät neljä eri lähestymistapaa, joissa vaihteli pudokkaan määritelmä, sekä tarkasteltava opintojen ajanjakso. Pudokas oli joko kaikki ”ei-

valmistuneet” tai ”opiskelijat, jotka eivät ole kahteen lukukauteen ilmoittautuneet kursseille”. Opintosuoritusaineiston osalta tarkasteltava ajanjakso oli joko koko opiskeluaika tai sitten viimeinen lukukausi juuri ennen opintojen keskeytymistä. Tässä tapauksessa ei-pudokkaiden osalta käytettävä lukukausi oli arvottu. Parhaiten pärjäsivät mallit, jotka ottivat huomioon opiskelijan koko opiskeluajan, ja joissa pudokas oli määritelty ”ei-valmistunut”. Koneoppimismalleista satunnaismetsä antoi tarkimman ennusteen.

Monissa tutkimuksissa parhaat tulokset oli saatu aineistolla, jossa yhtä opiskelijaa kohden oli vain yksi tietue, johon oli koottu opintomenestystä ja taustatietoja. Tietueessa saattoi olla esimerkiksi eri kentissä keskiarvot tai suoritettujen kurssien määrä eri lukukausilta.

Aineiston käsittely on myös tässä tutkielmassa erittäin tärkeässä roolissa. Tässä työssä tarkoitus olisi ennustaa mahdollisimmaan ajankohtaisesti, onko opiskelija keskeyttämisvaarassa juuri nyt. Tämän takia on mahdotonta saada opiskelijan koko opintohistoriaa tietueeseen, koska sitä ei välttämättä ole vielä paljoa. Tavoitteena olisi pystyä ennustamaan opiskelijan ensimmäisen – tai viimeistään toisen – vuoden aikana, onko opiskelija potentiaalinen tuleva pudokas. Toisin sanoen voi olla turha tieto, että keskeyttämisvaarassa on opiskelija, joka on saanut huonoja arvosanoja opintojen ensimmäiset 3 vuotta ja ollut sen jälkeen ilmoittautumatta yhdellekään kurssille kokonaiseen vuoteen. Olisi paljon tarkoitukenmukaisempaa päästä kiinni siihen, mitä tapahtuu mikrotasolla, eli näkyisikö keskeyttämisvaara jo aivan ensimmäisten kurssien tai ilmoittautumisten datassa. Koska suurimmassa osassa tutkimuksista on käytetty koostettua opintomenestysdataa, kannattaa menetelmää kokeilla myös tässä työssä. Suoritukset on todennäköisesti hyvä jakaa pienempiin yksiköihin kuin vuosittaisiin tai lukukausittaisiin opintopistemääriin ja keskiarvoihin, jotta opintojen alkuvaiheen suoritusten vaihtelu tulisi esiin.

3 Aineisto

Tässä osiossa tarkastellaan aineistoa, joka tätä tutkielmaa varten on koottu. Sen jälkeen määritellään pudokas analyysin perusteella, ja tehdäään yleiskatsaus siihen, minkä muotoista aineisto on ja mitä aineisto pitää sisällään.

Tätä tutkielmaa varten koottu aineisto sisältää opiskelija- ja kurssitietoja Helsingin yliopiston tietojenkäsittelytieteen kandiohjelman opiskelijoista, jotka ovat saaneet opinto-oikeuden lukuvuosina 2017-2023. Lisäksi vuodesta 2020 alkaen on saatavilla myös ilmoittautumistietoja. Aineisto on koottu syyskuussa 2024. Mitään kursseihin liittyviä tietoja ei ole saatavilla tämän jälkeen, mutta lukukausi-ilmoittautumiset on tiedossa vuoden 2025 kevääseen saakka.

Opiskelijatiedot sisältävät opinto-oikeuden alkamis- ja päättymisajat, opintojen aloitusajankohdan, lukukausien läsnäolotiedot ja tiedon siitä, onko opiskelija valmistunut kandidaatiksi. Opiskelijoista on tiedossa myös ikä ja sukupuoli. Kurssitiedoissa näkyy suoritettujen kurssien kurssikoodit, suorituspäivät, opintopistemäärät ja kurssien arvosanat. Ilmoittautumistiedoissa puolestaan ilmoittautumispäivämäärä, lukukausi ja kurssikoodit. Kurssi- ja ilmoittautumistiedot käsittävät kaikki kyseisten opiskelijoiden kurssit Helsingin yliopiston opintotietojärjestelmässä, eli niitä ei ole karsittu sen mukaan, minkä koulutusohjelman kursseja ne ovat.

Opiskelijatietojen formaatti:

```
opisknro;opinto-oik_alku;opinto-oik_loppu;aloituspvm;valmistunut;lukukausi-ilmot;
syntynyt;sukupuoli
a123;2018-08-01;2024-06-19;2018-08-01;true;137:1,138:1,139:3 ... ,1998,1
b124;2017-08-01;2024-06-19;2017-08-01;false;137:1,138:1,139:1 ... ,1980,2
```

Kurssitietojen formaatti:

```
opisknro;arvosana;opintopisteet;suoritus_pvm;kurssi
a123;4;10;2019-05-08;TKT20001
a123;5;5;2018-12-21;TKT10001
b124;4;5;2018-12-21;TKT10001
```

Ilmoittautumistietojen formaatti:

```
opisknro;lukukausi;ilm-pvm;kurssi
a123;138;2018-11-01;TKT10001
a123;143;2019-03-01;TKT20001
b124;138;2018-11-01;TKT10001
```

3.1 Aineiston esikarsinta ja ajanjaksojen määrittely

Aineisto vaatii jonkin verran käsittelyä ennen kuin se on käyttökelpoista data-analyysiä ja koneoppimista varten. Kurssi- ja ilmoittautumisaineistot sisältävät tietoja myös jo valmistuneilta, ja koulutusohjelmaa vaihtaneilta opiskelijoilta. Näitä tietoja ei haluta mukaan varsinaiseen aineistoon. Opiskelija-aineistossa on saatavilla ”opinto-oikeuden päättymispäivämäärä”. Kyseisen päivämäärän merkitys vaihtelee riippuen opiskelijan tilanteesta. Jos opiskelija on jo ehtinyt valmistua kandidaatiksi, päivämäärä on valmistumisen päivämäärä. Muilla päivämäärä on opinto-oikeuden päättymispäivämäärä tietojenkäsittelytieteen koulutusohjelmassa. Jos siis opiskelija on vaihtanut esimerkiksi farmasian koulutusohjelmaan, päivämäärä on vaihdon päivämäärä. Kyseisen päivämäärän avulla kurssi- ja ilmoittautumisaineistosta saadaan poistettua kaikki rivit, joiden ilmoittautumis- tai suorituspäivämäärä on opinto-oikeuden, valmistumisen tai koulutusohjelman vaihtamisen jälkeen.

Aineisto tarjoaa suorituspäivämäärien lisäksi tietoa lukukaudesta, jonka aikana suoritus on tehty. Kuitenkin data-analyysiä ja koneoppimista varten opintopistemäärien tai keskiarvon laskeminen pelkästään lukukausittain antaisi melko karkean arvion opintojen etenemisestä. Tässä työssä on toiveena päästä käsiksi mahdollisimman ajankohtaisesti opintojen ongelmakohtiin, joten hienojakoisemmalle jaottelulle voi olla tarvetta. Kuukausittain jaottelu puolestaan olisi epäkäytännöllinen, koska yksi kurssi kestää yleensä kahdesta neljään kuukautta, joten nolla-arvojen määrä kasvaisi suureksi. Helsingin yliopisto noudattaa sykliä, jossa yhdessä vuodessa on neljä periodia ja lisäksi kesäperiodi, jolloin suurin osa opiskelijoista viettää kesälomaa. Opintosuoritusten jaottelu periodin pituisiin jaksoihin tuntuu järkevimmältä vaihtoehdolta. Periodien tarkkoja alku- ja loppupäivämääriä on kuitenkin hankala yhdistää tähän aineistoon. Tämän työn pohjalta rakennettavaa järjestelmää on tarkoitus käyttää myös tulevaisuudessa, eikä tulevaisuuden periodien alkamis- ja loppumisaikoja voida tietää etukäteen ja päivämäärät pitäisi arvata. Tämän vuoksi vuoden jakamista periodeihin päivämäärätarkasti ei nähdä tarpeellisena. Edellä mainittujen

syiden takia tässä työssä vuosi jaetaan viiteen aikajaksoon taulukon 3.2 mukaisesti. Jako takaa kurssisuoritusaineiston jakautumisen suhteellisen tasaisesti ympäri vuoden. Tasainen jakutuminen mahdollistaa myös koneoppimismallin käyttämisen edelleen, vaikka sisäänottoa tehtäisiin joskus myös alkuvuodesta. Jokaiselle ajanjaksolle lasketaan erikseen opintopisteiden määrä, ajanjakson keskiarvo, hylättyjen kurssien määrä kullakin ajanjaksolla sekä ilmoittautumisaineiston ollessa mukana, myös keskeytettyjen kurssien määrä per ajanjakso. Suurimmillaan aineistossa voi siis esiintyä yhteensä 38 ajanjaksoa (2017 aloittaneista voi olla aineistoa 7,5 vuoden ajalta). Ennen opintojen aloituspäivää tehdyt suoritukset kerätään 0-jakso -kenttään. Käytännössä tämä kenttä kertoo opiskelijan opintomenestyksestä edeltävissä opinnoissaan.

arvosana	numero
HT	5
L	5
ECLA	5
KH	4
MCLA	4
NSLA	3
TT	3
Hyv.	3
CL	3
LA	2
HYL	0
Hyl.	0
EISA,	-1/0 *
LUOP,	-1/0 *

Taulukko 3.1: Kirjallisten arvosanojen muuntaminen välille 0-5. * Arvosanat ”EISA” ja ”LUOP” käsitellään keskeytettynä kurssina (-1), mikäli ilmoittautumisaineisto on mukana aineistoissa, muutoin 0.

Kurssiaineiston arvosana-kentässä on hyvin monen tyyppisiä arvosanoja, koska osa suorituksia on vanhoja hyväksilukuja, ja osasta kursseista saa hyväksytty/hylätty -tyyppisiä merkintöjä. Arvosanat muutetaan välille $0-5$, jossa nolla vastaa hylättyä kurssia. Kirjallisia arvosanoja ei haluttu jättää aineistosta pois – kuten opintotietojärjestelmässä tehdään esim. keskiarvoa laskiessa – koska ainesto jaotellaan lyhyisiin ajanjaksoihin, ja opiskelijan edistymisestä kertoo hyvinkin paljon se, että on saatu esimerkiksi kahdesta 3 opintopisteen kurssista ”hyväksytty” sen sijaan, että ajanjakson keskiarvo olisi 1.0 jonkin kolman-

Kuukaudet
Tammi-, helmikuu
Maalis-, huhtikuu
Touko-, kesä-, heinäkuu
Elo-, syys-, lokakuu
Marras-, joulukuu

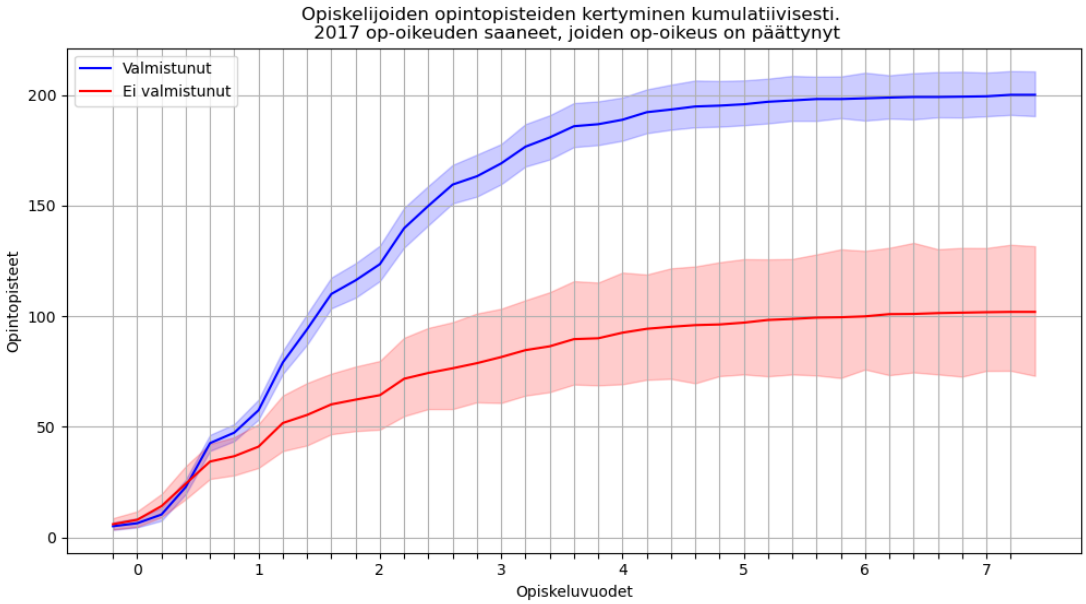
Taulukko 3.2: Vuoden jako viiteen ajanjaksoon. Yksi ajanjakso käsittää 2-3 kuukautta.

nen huonosti menneen yhden opintopisteen kurssin vuoksi. Tämän vuoksi kirjalliset arvosanat on muutettu numeerisiksi, vaikkei keskiarvon tarkkuus ole tällöin absoluuttinen. Arvosanojen muunnokset on nähtävissä taulukossa 3.1.

3.2 Yleiskatsaus

Aineisto sisältää 143 sellaista opiskelijaa, jotka saivat opinto-oikeuden 2017 ja joiden opinto-oikeus on päättynyt. Näistä opiskelijoista valmistui 86 henkilöä, eli 60 prosenttia. Valmistuminen/ei-valmistuminen tiedetään näiden opiskelijoiden kohdalla melko varmaksi, koska opinto-oikeus on päättynyt. On kuitenkin mahdollista, että opiskelija vielä anoo oikeuden takaisin ja valmistuu joskus. Tämän määrittelyn suhteen joudutaan siis sietämään pientä epävarmuutta. Nimitetään ei-valmistuneita kuitenkin ”ei-valmistuneiksi”, vaikkei 100% varmuutta voida taata. Analysoidaan hieman näiden opiskelijoiden kurssiaineistoa.

Opintopisteiden kertyminen alkaa hidastua ei-valmistuneiden osalta selvästi ensimmäisen opiskeluvuoden jälkeen. Viivakaaviossa 3.1 nähdään että ei-valmistuneet ja valmistuneet alkavat erottua toisistaan selkeästi noin 6.-7. ajanjakson kohdalla (syksyllä aloittaneilla tämä vastaa toisen vuoden kevättä). Viivakaaviossa näkyy haaleammalla värillä otoksen keskiarvon 95% luottamusvälin raja-arvot ja vahvalla värillä otoksen keskiarvo. Kurssien yhteiskeskiarvo puolestaan on yllättävän sama molemmilla joukoilla. Kuitenkin suurin piirtein 7. ajanjakson kohdalla yhteiskeskiarvokin erottelee opiskelijat ei-valmistuneisiin ja valmistuneisiin hienoisella erolla. Yhteiskeskiarvo on kuvattu kaaviossa 3.2. Ajanjaksokohtainen keskiarvo puolestaan ei erottele kovinkaan selkeästi ei-valmistuneita valmistuneista. Ainoastaan valmistuneilla vaihteluväli on hieman maltillisempi, mikä todennäköisesti kertoo suuremmasta määrästä suoritettuja kursseja, jolloin keskiarvokin pysyy tasaisempana. Lisäksi tulee huomioida, että ei-valmistuneet ovat aineistossa tehneet hyvin vähän



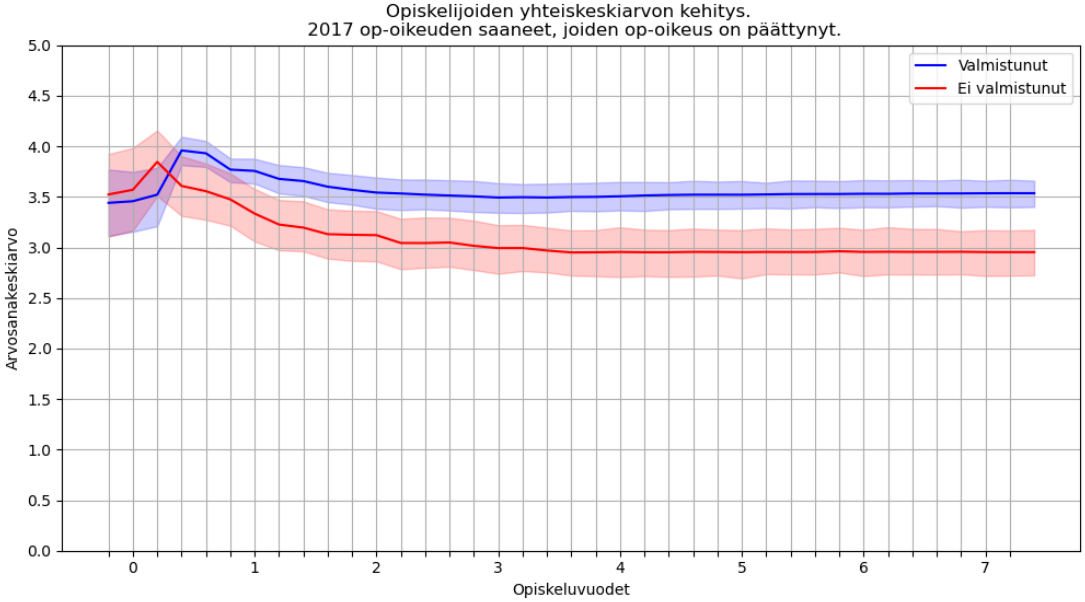
Kuva 3.1: Opintopisteiden kertyminen

suorituksia viimeisimmillä ajanjaksoilla, joten yksilöiden väliset erot näkyvät selkeämmin. Ajanjaksojen keskiarvot ovat nähtävissä kaaviossa 3.3.

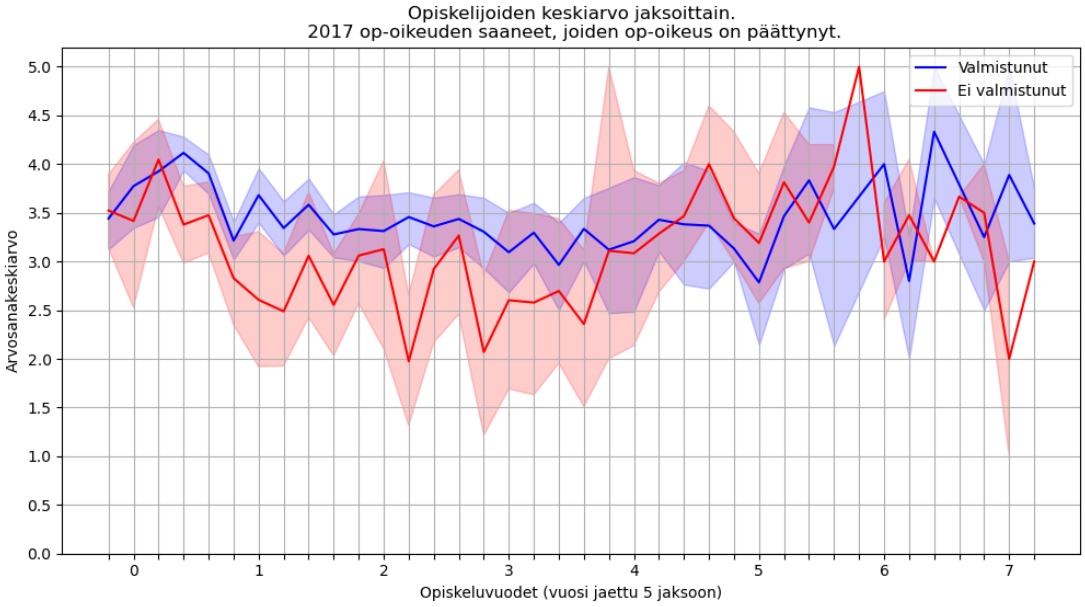
Jos pudokkuuden määritelmä muissa tutkimuksissa ei ole ollut ”ei-valmistunut”, se on määritelty usein aikavälinä, jolloin opiskelijalta ei ole tullut kurssisuorituksia. Tämän vuoksi tarkastellaan vielä ajanjaksoittain opiskelijoiden määrää, joilla on nolla suoritusta kyseiseltä jaksolta. Valmistuneilla on selvästi vähemmän nollasuoritusjaksoja opintojen alussa. Viivakaaviosta 3.4 nähdään, että esimerkiksi ensimmäsen vuoden opintojen jälkeen valmistuvista opiskelijoista vain noin viidellä prosentilla on nolla suoritusta, kun taas ei-valmistuvista tässä vaihessa opintoja nollasuorituksia on noin 50 prosentilla. Eli myös nollasuoritusjaksojen määrä erottelee joukot toisistaan. Tätä kaaviota tarkastellessa tulee huomioida, että opinto-oikeudellisten opiskelijoiden määrä vähenee ajan kasvaessa, kun osa valmistuu ja osa lopettaa opiskelun kokonaan.

3.3 Pudokkaan määrittely

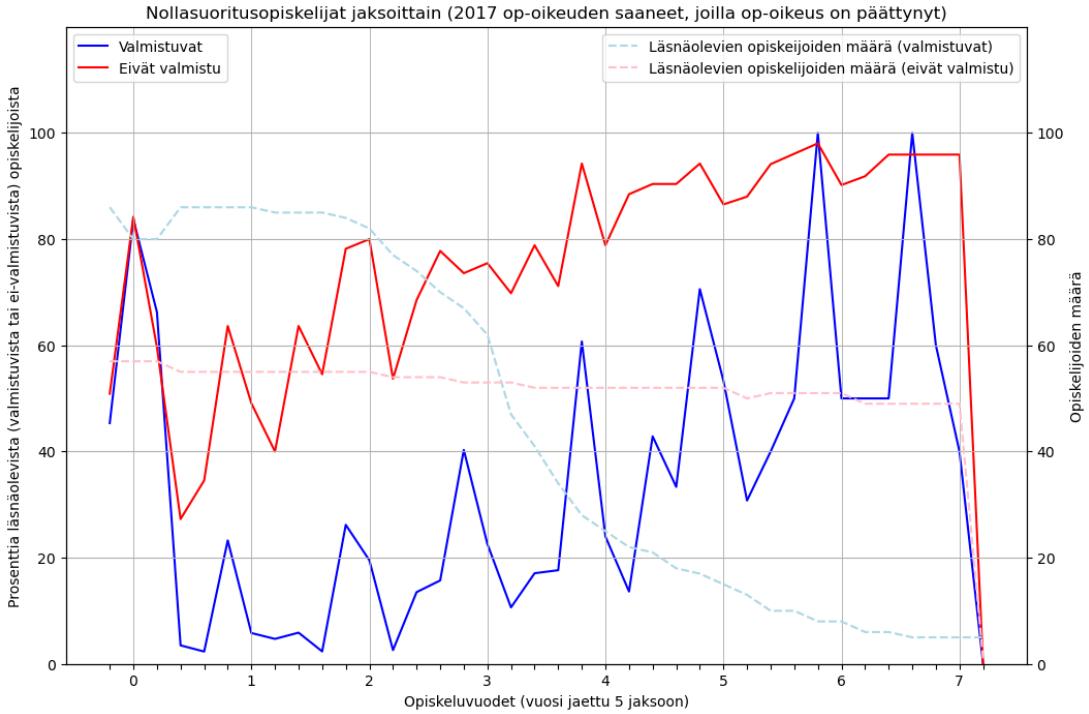
Aluksi aineistoa tulee analysoida, jotta saadaan selville, sopiiko aineistoon sama pudokkaan määritelmä, jota esimerkiksi tutkimuksessa [26, 27] käytettiin (ei suorituksia kahteen lukukauteen, eli vuoteen). Jos mahdollista, aikaväli saisi olla mahdollisimman lyhyt, jotta aineistoon saataisiin mukaan mahdollisimman monta opiskelijaa, koska aineisto si-



Kuva 3.2: Yhteiskeskiarvon kehittyminen



Kuva 3.3: Keskiarvot ajanjaksoittain



Kuva 3.4: Opiskelijat, joilla on nolla suoritusta ajanjaksolla. Prosenttiosuus on nollasuoritusopiskelijoiden osuus kyseisen ajanjakson aikana läsnäolevien opiskelijoiden määrästä per osajoukko (valmistuneet/ei valmistuneet).

sältää myös vain vähän aikaa opiskelleita opiskelijoita. Analysointia varten kurssitiedoista ja opiskelijatiedoista johdetaan aineisto, jossa yhtä opiskelijaa kohden on vain yksi tietue. Sarakkeissa on tiedot opinto-oikeudesta ja kronologisessa järjestyksessä opiskeluajan ajanjaksoista ja niiden suorituksista. Aineiston muoto on esillä taulukossa 3.3.

Johdetun aineiston avulla tutkitaan, onko opiskelijoilla, jotka eivät ole valmistuneet koko opinto-oikeutensa aikana ja eri pituisilla nollasuoritusjaksoilla vahva riippuvuus. Näin pyritään löytämään sopiva rajaus pudokkaan määritelmälle. ”Ei ole valmistunut 7 vuoden aikana” tietysti takaisi hyvin suurella todennäköisyydellä varmuuden siitä, että kyseessä on pudokas, mutta tämä määritelmä ei palvele käyttötarkoitusta tässä tapauksessa, koska kurssidataa on saatavilla vain rajatulta ajalta, ja vain hyvin pieni osa aineistosta kattaa opiskelijan koko opinto-oikeusajan. Lisäksi opiskelija voi hakea opinnoilleen lisäaikaa vaikka opinto-oikeus olisi jo päättynyt. Todennäköisesti kuitenkin pienelläkin otoksella hyvin varmoja pudokkaita (v. 2017 aloittaneet, ei valmistuneet 7 vuodessa), saadaan johdettua parempi muuttuja – joka esiintyy aineistossa laajemmin – vaikkei tarjoakaan yhtä hyvää varmuutta siitä, että kyseessä on pudokas. Tan et al. määrittelivät pudokkaan vastaavalla tavalla analyysin tuloksena [27]. He päätyivät lopulta käyttämään määritelmää ”ei osallis-

Sarakkeen nimi	Numeroskaala	Selite
opisknro	0 - max(int32)	Opiskelijan yksilöivä tunniste
opinto-oik_alku	0 - max(int32)	Vuosi, jolloin opinnot on aloitettu
opinto-oik_loppu	0 - max(int32)	Vuosi, jolloin opinto-oikeus loppuu
valmistunut	0-1	Onko opiskelija valmistunut kandidaatiksi (0=ei, 1=kyllä)
lk_ilmo_0	-1-3	1. lukukauden ilmoittautumistieto
lk_ilmo_1	-1-3	2. lukukauden ilmoittautumistieto
lk_ilmo_x	-1-3	seuraavien lukukausien ilmoittautumistiedot
		(-1=NaN, 0=luvalla p., 1=läsnä, 2=poissa, 3=laiminlyöty)
...		yht. 16 lk_ilmo -saraketta (8 lukuvuotta)
op_0	0 - max(float32)	aikaisempien opintojen opintopistemäärä tai NaN
op_1	0 - max(float32)	1. ajanjakson opintopistemäärä tai NaN
op_x	0 - max(float32)	seuraavien ajanjaksojen opintopistemäärät tai NaN
...		yht. 39 op_ -saraketta (7,5 lukuvuotta)

Taulukko 3.3: Aineisto pudokkaan määritelmää varten. Selitteet: NaN tarkoittaa puuttuvaa tietoa kentässä. Lukukausi-ilmoittautumisten osalta se tarkoittaa sitä, että ei ole opinto-oikeutta. ”Luvalla p.” on poissaolo joka ei kuluta opinto-oikeutta. op_x kentissä puuttuva arvo tarkoittaa sitä, että poissaolevaksi ilmoittautuneella opiskelijalla ei ole suorituksia.

tunut kahteen lukukauteen päättökokeisiin", jolla oli vahvin korrelaatio "ei valmistunut" -muuttujan kanssa.

Johdetusta aineistosta eristetään osajoukko, joka sisältää vain ne opiskelijat, joiden opintooikeus on päättynyt. Tämän osajoukon koko on 201 opiskelijaa. Tästä aineistosta eristetään useampi osajoukko, jossa opiskelijat eivät ole tehneet yhtään suoritusta tietyllä aikavälillä, vaikka he eivät ole ilmoittaneet poissaolostaan. Tällaisia joukkoja ovat esimerkiksi "kaksi peräkkäistä ajanjaksoa ilman suorituksia", "kolme peräkkäistä ajanjaksoa ilman suorituksia" jne. Näiden joukkojen ja ei-valmistuneiden väliltä etsitään merkitseviä riippuvuuksia. Tavoitteena on löytää sellainen puuttuvien suoritusten aikaväli, joka parhaiten selittää valmistumista, ja jonka riippuvuus valmistumiseen on merkitsevä, mutta puuttuvien suoritusten aikaväli mahdollisimman lyhyt. Puuttuvien suoristusten aikavälin ollessa mahdollisimman lyhyt, saadaan varsinaisesta aineistosta suurempi osa määriteltyä pudokkaaksi, koska aineistossa on suuri osa vastikään aloittaneita opiskelijoita.

3.3.1 Khiin neliö -testi

Nollasuoritusjaksojen ja valmistumattomuuden välisiä riippuvuuksia testataan aluksi Khiin neliöllä, joka sopii kategoristen muuttujien välisten lineaaristen riippuvuuksien testaamiseen [23].

Khiin neliö -testi perustuu odotetun ja havaitun frekvenssijakauman vertailuun [23]. Tarkastellaan siis käytännössä sitä, onko kahden joukon muodostamat frekvenssijakaumat samanmuotoiset. Testissä nollahypoteesi on H_0: Kaksi muuttujaa ovat riippumattomia toisistaan. Vaihtoehtoinen hypoteesi on H_1: Kaksi muuttujaa eivät ole riippumattomia toisistaan, eli niiden välillä on havaittavissa riippuvuus.

Khiin neliön testisuure TS saadaan seuravaalla kaavalla:

$$TS = \sum_{i=1}^{k} E_{ij} = \sum_{i=1}^{k} \frac{(N_i - e_i)^2}{e_i},$$

jossa E_{ij} on odotettu frekvenssi, N_i on ensimmäisen muuttujan frekvenssi ja e_i toisen muuttujan frekvenssi.

Khiin neliö -testisuure TS kuvaa eroja havaittujen ja odotettujen frekvenssien välillä. Testisuuretta verrataan teoreettiseen Khiin neliö -jakauman arvoon halutulla luottamusvälillä. Jos TS ylittää kriittisen arvon, nollahypoteesi hylätään ja päätellään, että muuttujien väillä on tilastollisesti merkitsevä riippuvuus valitulla luottamusvälillä [23].

Testi toimii kun otoskoko on ”riittävän suuri” [23]. Tässä tapauksessa 201 opiskelijan otoskokoa voidaan pitää käyttötarkoitukseen nähden riittävän suurena. Lasketaan ensin frekvenssit, eli kuinka moni koko otoksen opiskelijoista valmistuu ja kuinka moni ei valmistu. Tämän lisäksi lasketaan kuinka monella on ”x jaksoa ilman suorituksia” ja kuinka monella ei. Odotettu frekvenssi E_{ij} lasketaan siis siten, että N_i on ”ei suorituksia aikavälillä x”-muuttujan frekvenssi ja e_i on ”ei valmistunut”-muuttujan frekvenssi.

Laskennan nopeuttamiseksi jokaisesta osajoukosta (”x jaksoa ilman suorituksia”, jossa $x = \{1 - 13\}$) muodostetaan ristiintaulukko, ja ajetaan erikseen `scipy.stats`-kirjaston `chi2_contingency`-funktiolla Khiin neliö -testi. Testeistä selviää, että itse asiassa kaikilla valituilla nollasuoritusaikaväleillä, jossa peräkkäisiä jaksoja on enemmän kuin yksi, on 99% luottamusvälillä merkittävä riippuvuus valmistumattomuuden kanssa. Testitulokset on esitetty taulukossa 3.4. Tämä vahvistaa oletuksen siitä, että nollasuoritusjaksoja on hyvä käyttää pudokkuuden määritelmänä. Käyttötarkoituksen vuoksi pyritään löytämään lyhyin nollasuoritusten aikaväli, jotta aineistossa näkyisi pudokkaita mahdollisimman paljon. Testeistä selviää, että suurin riippuvuus on tärkeysjärjestyksessä muuttujilla 6,7,8 ja 5 ”peräkkäistä jaksoa ilman suorituksia”. Muuttujalla ”6 peräkkäistä jaksoa ilman suorituksia” ja ”7 peräkkäistä jaksoa ilman suorituksia” on suurin Khiin neliön arvo ja pienin p-arvo. 6 tai 5 ”peräkkäistä jaksoa ilman suorituksia” sopisi pudokkaan määritelmäksi parhaiten, koska jaksojen määrä on pieni mutta merkitsevyys suuri.

3.3.2 Yhteisinformaatio

Tarkastellaan vielä muuttujien yhteisinformaatiota varmistuksen vuoksi. Yhteisinformaatio paljastaa muuttujat, jotka tarjoavat eniten informaatiota kohteeksi valitusta muuttujasta (tässä: valmistumattomuus). Yhteisinformaatiota käytetään usein muuttujien valitsemiseen koneoppimisessa [20]. Yhteisinformaatio kertoo, kuinka samanlaiset muuttujien yhteistodennäköisyysjakauma $P(x, y)$ ja tulojakauma $P(x)P(y)$ ovat [12]. Yhteisinformaatio $I(X;Y) \geq 0$. $I(X;Y) = 0$ jos ja vain jos $P(x)P(y) = P(x, y)$. Eli yhteisinformaatio on nolla, jos muuttujat ovat toisistaan riippumattomat. Siten yhteisinformaatio $I(X;Y)$ voidaan määritellä ehdollisen todennäköisyyden avulla seuraavasti:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X),$$

jossa $H(Y|X)$ on ehdollinen entropia $\sum_x p(x)H(Y|X = x)$ [12].

Yhteisinformaation laskemista varten yhdistetään samaan taulukkoon tieto valmistumisesta (0 tai 1) ja seuraaviin sarakkeisiin tieto ”x peräkkäistä jaksoa ilman suorituksia” (0 tai

Aika ilman suorituksia	P-arvo	Vapausasteet	Khiin neliö	Merkittävä 99%
1 jakso	0.125	1	2.350	Ei
2 peräk. jaksoa	$8.600 \cdot 10^{7}$	1	24.219	Kyllä
3 peräk. jaksoa	$1.082 \cdot 10^{20}$	1	87.006	Kyllä
4 peräk. jaksoa	$1.228 \cdot 10^{21}$	1	91.310	Kyllä
5 peräk. jaksoa	$3.534 \cdot 10^{22}$	1	93.775	Kyllä
6 peräk. jaksoa	$1.078 \cdot 10^{23}$	1	**100.684**	Kyllä
7 peräk. jaksoa	$1.078 \cdot 10^{23}$	1	**100.684**	Kyllä
8 peräk. jaksoa	$4.127 \cdot 10^{23}$	1	98.022	Kyllä
9 peräk. jaksoa	$3.773 \cdot 10^{21}$	1	89.900	Kyllä
10 peräk. jaksoa	$1.600 \cdot 10^{20}$	1	86.232	Kyllä
11 peräk. jaksoa	$1.032 \cdot 10^{18}$	1	77.996	Kyllä
12 peräk. jaksoa	$3.927 \cdot 10^{18}$	1	75.357	Kyllä
13 peräk. jaksoa	$1.457 \cdot 10^{17}$	1	72.770	Kyllä

Taulukko 3.4: Khiin neliö -testit nollasuoritusjaksojen ja valmistumattomuuden välillä. Sarake ”merkittävä 99%” kertoo, onko tulos merkittävä 99% luottamusvälillä.

1), jossa $x = \{1 - 13\}$. Laskemiseen käytetään `sklearn`-kirjaston `mutual_info_classif` -funktiota, jota kirjasto tarjoaa muuttujien valitsemiseen. Funktio ajetaan edellä mainitulle taulukolle. Tulokset ovat samanlaiset kuin Khiin neliön testeillä, eli eniten informaatiota valmistumattomuudesta tarjoaa tieto siitä, onko opiskelijalla 6 tai 7 peräkkäistä jaksoa ilman suorituksia (eli hieman yli vuosi tai 1,5 vuotta ilman suorituksia). Tämä ei ole yllättävää, koska otoskoko on sama jokaisessa Khiin neliön testissä, joten Khiin neliö -arvot ovat jo itsessään vertailukelpoisia keskenään, ja niiden tulisi antaa samat tulokset muuttujien tärkeydestä kuin yhteisinformaatio. Yhteisinformaation laskemisella saatiin tarkistettua, ettei Khiin neliö -testeissä ole tapahtunut laskuvirheitä. Tulokset on esitelty taulukossa 3.5.

3.3.3 Pudokkaan määritelmä

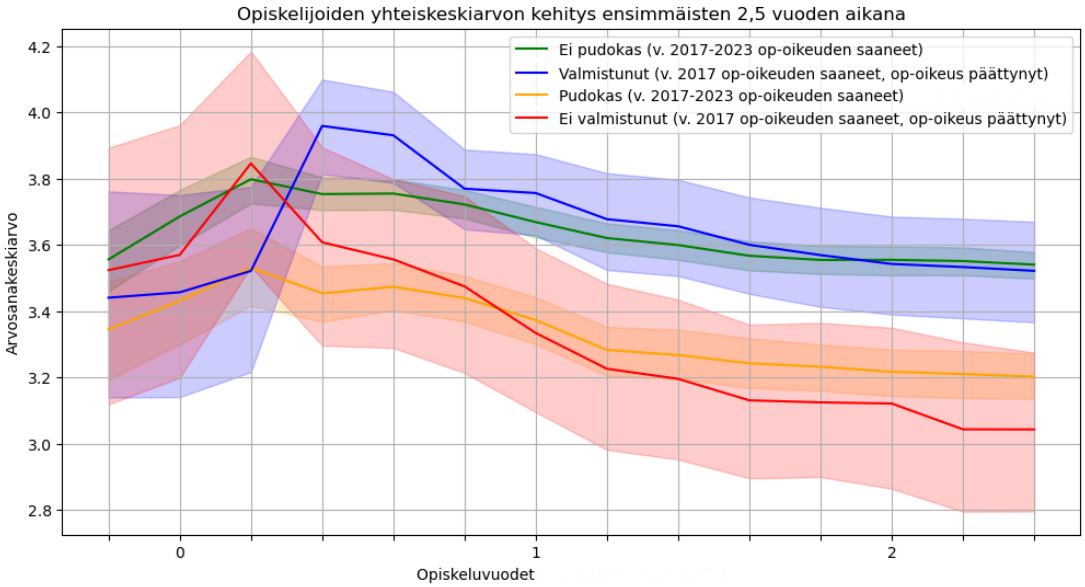
Khiin neliö -testien ja yhteisinfomaation tulosten valossa valitaan tähän työhön pudokkuuden määritelmäksi ”6 peräkkäistä jaksoa ilman suorituksia”. Vuosi on tässä työssä jaettu viiteen ajanjaksoon (kts. 3.1), joten 6 jaksoa tarkoittaa siis yhtä vuotta ja 2-3 kuukautta, eli karkeasti arvioituna 14-15 kuukautta ilman suorituksia ja poissaoloilmoitusta.

Aika ilman suorituksia	Yhteisinfomaatio
6 peräk. jaksoa	**0.397615**
7 peräk. jaksoa	**0.397615**
8 peräk. jaksoa	0.389741
5 peräk. jaksoa	0.362622
4 peräk. jaksoa	0.354381
9 peräk. jaksoa	0.351416
3 peräk. jaksoa	0.342510
10 peräk. jaksoa	0.339438
11 peräk. jaksoa	0.305518
12 peräk. jaksoa	0.294813
13 peräk. jaksoa	0.284379
2 peräk. jaksoa	0.120702
1 jakso	0.019163

Taulukko 3.5: Yhteisinformaatio nollasuoritusjaksojen ja valmistumattomuuden välillä. Järjestetty merkittävimmästä muuttujasta vähiten merkittävään.

3.4 Katsaus aineistoon pudokkaan näkökulmasta

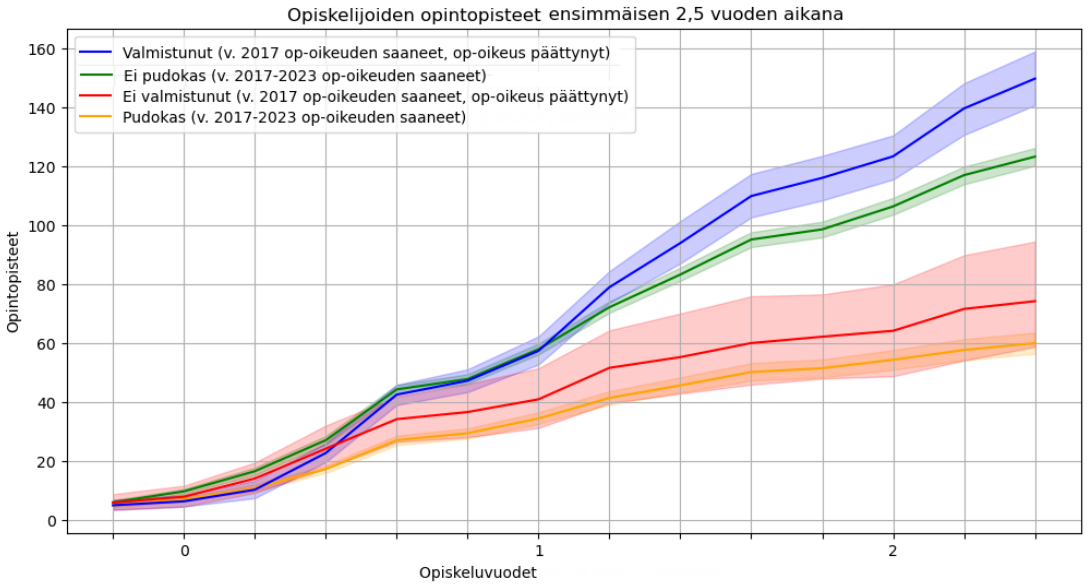
Nyt kun pudokkaaksi on määritelty ”opiskelijat, joilla on 6 peräkkäistä jaksoa ilman suorituksia ja ei ole ilmoittanut poissaolostaan”, voidaan tarkastella pudokkuutta koko aineistossa (2017-2023 opintonsa aloittaneet). Koko aineisto käsittää 1787 opiskelijaa. Aineistosta pudotetaan kuitenkin pois ne, jotka eivät ole vielä opiskelleet 14 kuukautta, koska he eivät voi olla pudokkaita. Käytännössä 2023 aloittaneet jäävät pois. Jäljelle jää 1549 opiskelijaa, joista määritelmän mukaisia pudokkaita on 695, eli noin 45 prosenttia. Ei-valmistuneita vuonna 2017 aloittaneiden aineistossa oli noin 40 prosenttia. Luvut ovat samaa kokoluokkaa, joten tämä vahvistaa ajatusta siitä, että valittu pudokkaan määritelmä sopii aineistoon. Viivakaavioista 3.6 ja 3.5 nähdään, että myös opintopisteiden kertyminen ja arvosanakehitys noudattelevat samaa linjaa kun verrataan ei-valmistuneita ja valittua pudokkaan määritelmää. Kaavioita tarkastellessa täytyy kuitenkin huomioida, että koko aineistossa (2017-2022 opinto-oikeuden saaneet) on mukana myös opiskelijoita, joilla on opinnot vielä huomattavasti kesken. Tämän takia tarkasteltavaksi ajaksi on valittu 2,5 vuotta. Lisäksi käyriä vertaillessa tulee huomioida, että otoskoot ovat huomattavan eri kokoiset. Esimerkiksi kaavioissa 3.5 ja 3.6 punainen viiva (”ei valmistunut”) on 57 opiskelijan



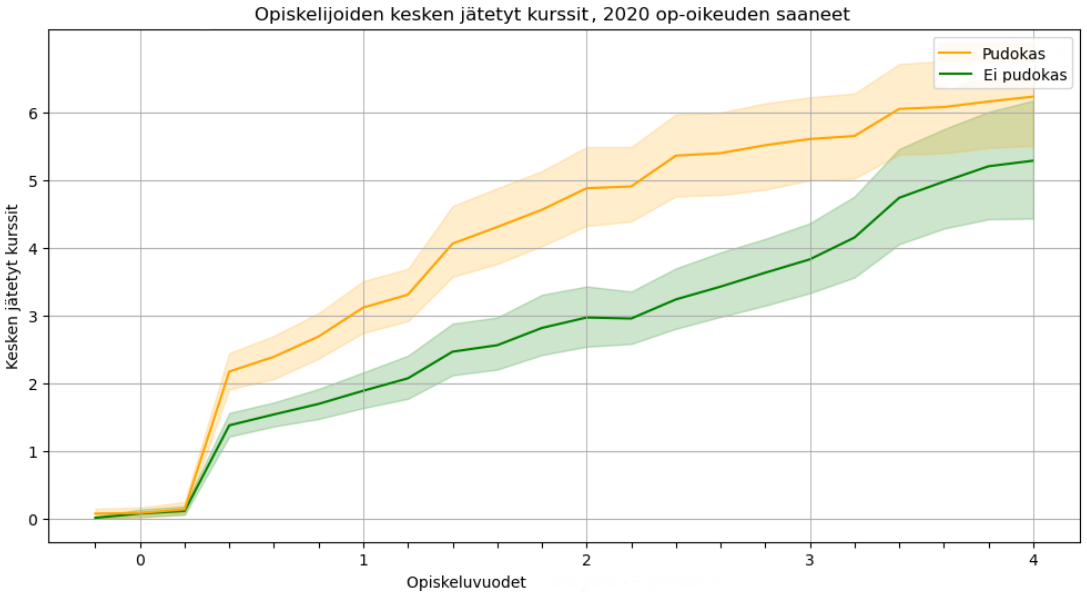
Kuva 3.5: Yhteiskeskiarvon kehittyminen ensimmäisten 2,5 vuoden aikana

keskiarvo, kun taas oranssi viiva (”pudokas”) on 695 opiskelijan keskiarvo. Punaisen ja sinisen käyrän (”valmistunut”/”ei-valmistunut”) paikallisille arvovaihteluille ei siis kannata antaa kovin suurta painoarvoa, vaan tarkastella kokonaiskuvaa.

Nyt päästään tarkastelemaan myös ilmoittautumisaineiston tuomaa lisäarvoa. Koska ilmoittautumisaineistoa on saatavissa vasta vuodesta 2020 lähtien, sitä ei voitu tarkastella lainkaan valmistumisen näkökulmasta, mutta pudokkuuden näkökulmasta voidaan. Vuonna 2020 aloittaneista on tiedossa opiskelijoiden ilmoittautumiset neljän vuoden ajalta. Näitä opiskelijoita on 290 henkilöä. Kun ilmoittautumisia verrataan suoritettuihin kursseihin, nähdään, että kesken jätettyjen kurssien keskimääräinen määrä on aavistuken isompi pudokkailla (kts. viivakaavio 3.7). Ero on kuitenkin melko pieni, joten tähän kaavioon kannattaa suhtautua pienellä varauksella, kuinka paljon informaatiota se tarjoaa todellisuudessa. Kurssi- ja ilmoittautumisaineistosta selviää myös, että tietyillä kursseilla pudokkaat suoriutuvat selvästi huonommin. Esimerkiksi pudokkaiden mediaanisuoritus kursseista TKT20005 (Laskennan mallit), TKT20001 (Tietorakenteet ja algoritmit) ja MAT11003 (Raja-arvot) onkin arvosanan sijaan kurssin keskeyttäminen, kun taas ei-pudokkailla arvosanamediaani kahdella ensimmäisellä kurssilla on 3 ja viimeisellä kurssilla 2. Useimmilla kursseilla suorituksissa ei näy näin suurta eroa – tai ei eroa silmämääräisesti lainkaan.



Kuva 3.6: Opintopisteiden kumulatiivinen kertyminen ensimmäisten 2,5 vuoden aikana



Kuva 3.7: Kesken jääneiden kurssien määrän kumulatiivinen kertyminen (ilmoittautuminen tehty, mutta ei ole kurssisuoritusta).

4 Ohjelmisto

Tämän työn toisena tavoitteena on luoda järjestelmä, jolla voi ennustaa opiskelijoiden mahdollisuutta päätyä pudokkaaksi. Ohjelmistolle syötetään kurssiaineisto sekä opiskelija-aineisto, ja ohjelmisto kertoo, ketkä opiskelijat ovat potentiaalisia tulevia pudokkaita. Ohjelmistolla voi myös kouluttaa mallin uudelleen uudella aineistolla. Ohjelmistolle koulutetaan kuitenkin yksi oletusmalli, jotta ohjelmistoa voi käyttää heti ennustamiseen, ilman hitaahkoa koulutusprosessia. Oletusmalli ja sen parametrit valitaan tämän työn kokeiden tulosten perusteella.

Valmiiksi koulutettu oletusmalli annetaan ainoastaan Helsingin yliopiston henkilökunnan edustajan käyttöön, koska se voi paljastaa tietoja koulutusaineistosta.

Ohjelmisto asetetaan saataville GitHub-palveluun. Repositoriosta löytyy ohjelmiston tarkemmat käyttöohjeet. Repositorion osoite on: https://github.com/Skorp7/Dropout-predictor

4.1 Vaatimukset ja käyttötarkoitus

Ohjelmiston päätavoite on tarjota helppokäyttöinen tapa suorittaa tässä työssä tehtyjä ennustuksia ilman koneoppimisohjelmointiosaamista tai erityisempiä aineistonkäsittelytaitoja. Riittää, että saa opiskelijoiden kurssitiedot ja taustatiedot siirrettyä .csv-tiedostona ohjelmistossa oikeaan kansioon. Aineiston käsittely, kouluttaminen ja ennustaminen ohjelmoidaan tässä työssä skriptityylisesti. Tämä ohjelmisto on siis koneoppimisskriptien ympärille rakennettava sovellus.

Toinen tärkeä lähtökohta ohjelmiston vaatimusmäärittelyssä on se, että sen tulee toimia yhtä hyvin tulevaisuuden opintoaineistolla, kuin nykyisellä tutkimusaineistolla. Tämän vuoksi aineisto on käsitelty niin, että esimerkiksi lukukaudet ja ajanjaksot saavat järjestysnumerot sen sijaan, että käytettäisiin esimerkiksi lukukausien numeroita. Aineiston käsittelyssä on otettu huomioon myös se, että sisäänottoaika saattaa olla syksyn lisäksi tai sijasta kevätlukukausi. Joustavan aineiston käsittelyn vuoksi ohjelmistoa on teoriassa mahdollista käyttää myös muissa koulutusohjelmissa. Tällöin tulee kuitenkin huomioida, että pudokkuus voi näkyä eri koulutusohjelmissa eri tavalla, eikä tähän työhön valittu pudokkuuden määritelmä välttämättä sovi yhtä hyvin kaikkiin koulutusohjelmiin.

Ohjelmiston tulee myös toimia helposti saatavilla olevilla ohjelmointikielillä ja -kirjastoilla. Sillä olisi hyvä pystyä myös jollain tapaa kokeilemaan, onko ohjelmistoympäristö kunnossa. Tällöin vian ilmetessä vikaa ymmärtää etsiä aineistosta, mikäli ympäristö on testien mukaan kunnossa.

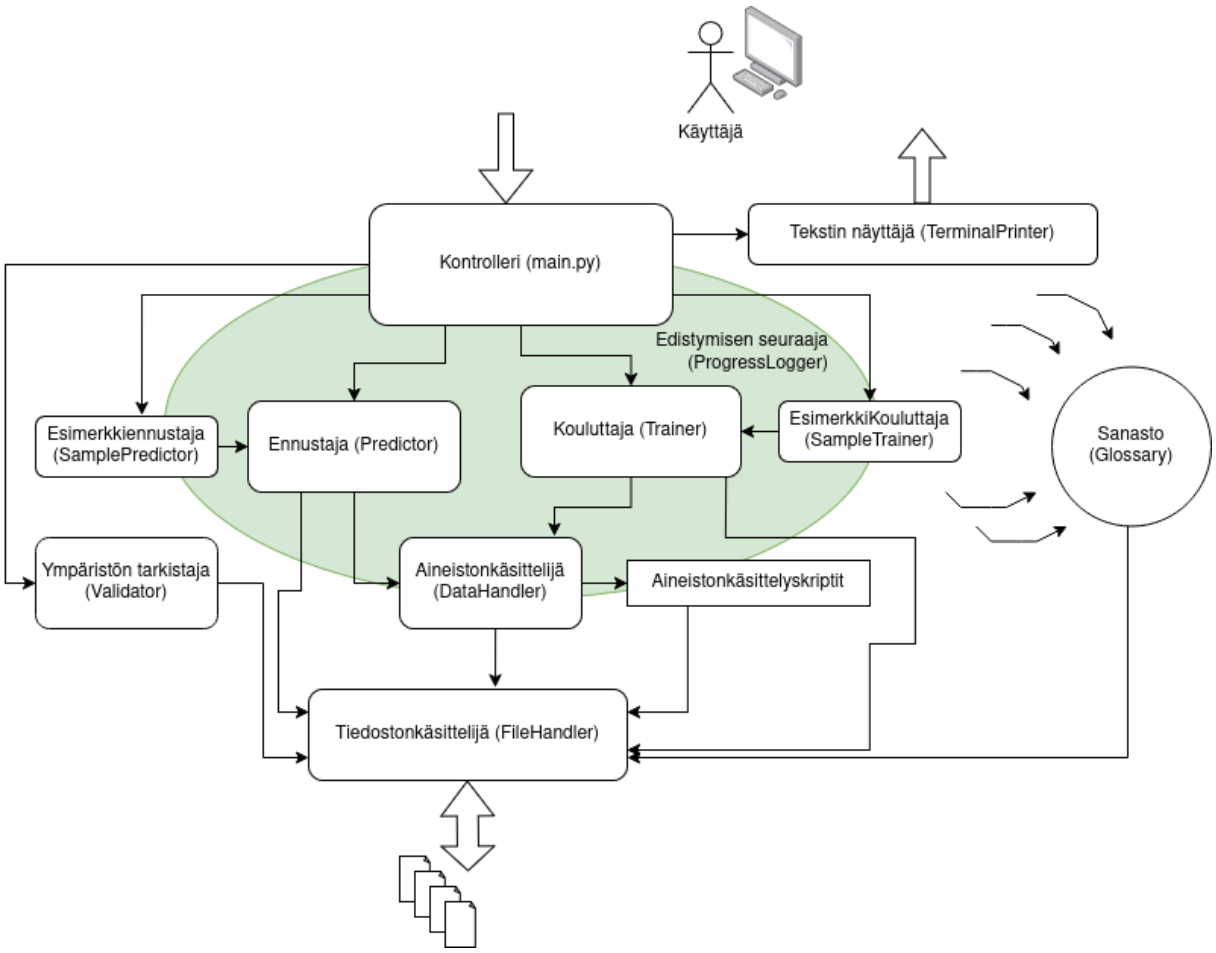
Ohjelmistolla ennustetaan pudokkuutta jossakin vaiheessa opintoja. Tarkkaa ajankohtaa ei ole tiedossa, milloin ennustus halutaan tehdä, joten ohjelmistoon luodaan mahdollisuus päättää ennustuksen ajankohta jokaiselle opetetulle mallille erikseen. Ohjelmistossa voi myös valita kielen suomen ja englannin väliltä, koska loppukäyttäjä ei välttämättä aina ole suomenkielinen.

4.2 Teknologiat ja arkkitehtuuri

Ohjelmisto ohjelmoidaan Python-ohjelmointikielellä ja siihen rakennetaan komentorivikäyttöliittymä, joka auttaa käyttäjää valitsemaan itselleen sopivat parametrit mallin kouluttamista varten. Tietokantaa ei käytetä, vaan ohjelmisto lukee suoraan tiedostoista ja kirjoittaa niihin. Tämä varmistaa ohjelmiston rakenteen pysymisen yksinkertaisena. Käytännössä käyttäjä tallentaa opintoaineiston muutamana .csv-tiedostona kansioon ja ohjelmisto tallentaa sekä käsittelemänsä aineiston, että ennustukset toisiin kansioihin. Käsitelty aineisto tallennetaan sen vuoksi, että voidaan tarvittaessa varmistaa aineiston eheys, jolla varsinainen koulutus tai ennustus on tehty. Ohjelmiston versionhallinta toteutetaan git:illä ja ohjelmiston varmuuskopiota säilytetään GitHub-palvelussa.

Ohjelmisto pyrkii noudattamaan luokkarakennetta ja kerrosarkkitehtuuria. Rakenne on kuvattu tarkemmin kuvaajassa 4.1. Esimerkiksi ennusteista huolehtii yksi luokka ja kouluttamisesta toinen. Aineiston valmistelee käyttökuntoon kolmas luokka. Sekä ennustaja- että kouluttajaluokat käyttävät samaa aineistonkäsittelijää, koska aineisto käsitellään lähestulkoon identtisellä tavalla kumpaankin käyttötarkoitukseen. Ainoastaan yksi luokka koskee tietokoneen tiedostoihin, ja muut luokat kutsuvat tätä luokkaa, kun jotain tarvitsee lukea tai tallentaa. Lisäksi on toteutettu koulutuksen ja ennustamisen testaamista varten pienet luokat, jotka perivät oikean ennustajan ja kouluttajan toiminnot.

Ohjelmisto käyttää käytännössä samoja aineistonkäsittelyskriptejä, joilla tämän työn aineisto käsitellään. Niitä ei ole erikseen eritelty moduuleihin, jotta niitä olisi helppo verrata tutkimuksessa käytettäviin koodeihin, ja että ne toimivat varmasti samalla tavalla.



Kuva 4.1: Ohjelmiston arkkitehtuuri luokkatasolla. Ohuet nuolet osoittavat, mihin suuntaan käskyt etenevät ohjelmistorakenteessa. Paksut nuolet kuvaavat tiedon liikkumisen suuntaa rajapinnoissa. Edistymisen seuraajaa kutsuvat ne luokat, jotka ovat vihreän alueen päällä. Sanastoa kutsuvat kaikki muut luokat paitsi tiedostonkäsittelijä.

5 Mallit ja kokeet

Tässä osiossa esitellään työssä käytettävät koneoppimismallit ja malleilla suoritettavat koeasetelmat. Kokeiden tavoitteena on määritellä, mikä koneoppimismalli tuottaa parhaiten tavoitetta palvelevat ennusteet käytössä olevalle aineistolle. Tätä mallia käytetään myöhemmin lopullisessa ohjelmassa ennusteen muodostamiseen.

Parhaiten tavoitetta palvelevilla ennusteilla tarkoitetetaan ennusteita, joilla oikeiden positiivisten määrä on mahdollisimman suuri ja väärien negatiivisten määrä mahdollisimman pieni, huolehtien kuitenkin siitä, ettei väärien positiivisten ennusteiden määrä ole liian suuri. Tavoite siis on saada mahdollisimman moni pudokas kinni, antamatta kovin suurta painoarvoa sille, että niin sanottuja ”vääriä pudokashälytyksiä” saattaa esiintyä joukossa. Tavoite on valittu näin sen vuoksi, että kustannus ei-valmistuvasta opiskelijasta on yliopistolle hyvin suuri verrattuna mahdollisiin tukitoimiin, joita ennustettuihin pudokkaisiin tultaisiin mahdollisesti kohdistamaan.

5.1 Mallien mittaaminen

Mallien paremmuutta arvioidaan useammalla eri mittarilla. Koulutusaineistossa on tiedossa jokaisen aineistorivin ”oikea tulos”, eli onko kyseisen rivin opiskelija määritelmän mukaan pudokas vai ei. Mallien paremmuutta vertaillessa koulutusaineisto jaetaan osiin siten, että osalla aineistosta malli koulutetaan ja pienempi loppuosa toimii testiaineistona. Testiaineistosta ”poistetaan” tieto siitä, onko opiskelija pudokas. Tiedot kuitenkin pidetään tallessa erillään aineistosta. Kun testiaineistolla tehdään ennuste, saadaan lopuksi verrattua ennustettua arvoa näihin talteen otettuihin alkuperäisiin pudokastietoihin. Ennuste voi siis tuottaa tulokseksi oikein ennustettuja positiivisia (OP), oikein ennustettuja negatiivisia (ON), väärin ennustettuja positiivisia (VP) ja väärin ennustettuja negatiivisia (VN). Esimerkiksi oikein ennustettu negatiivinen (ON) tarkoittaa sitä, että aineistossa alkuperäinen pudokastieto oli ”ei ole pudokas” ja myös malli ennusti, että tämä opiskelija ei ole pudokas. Oikein ja väärin ennustettujen osuuksista saadaan johdettua useita erilaisia mittareita [8]. Tässä työssä käytetään mittareita: **herkkyys** (engl. recall), **täsmällisyys** (engl. precision), **tarkkuus** (engl. accuracy), **Cohenin kappa** (engl. Cohen’s kappa) ja **f-arvo** (engl. f-score). F-arvo on määritelty erikseen aliluvussa 5.4.

$$\text{herkkyys} = \frac{OP}{(OP + VN)}$$

$$\text{täsmällisyys} = \frac{OP}{(OP + VP)}$$

$$\text{tarkkuus} = \frac{(OP + ON)}{(OP + ON + VP + VN)}$$

Tavoite on siis ennen kaikkea maksimoida herkkyys. Herkkyys on oikeiden positiivisten osuus kaikista aidoista positiivisista. Mallin tarjoama tarkkuus ja täsmällisyys ei todennäköiseti tule olemaan korkein mahdollinen tämän tavoitteen vuoksi. Tarkkuudella tarkoitetaan yhteenlaskettuja oikein menneiden ennustusten (oikea positiivinen tai oikea negatiivinen) osuutta kaikista havainnoista. Täsmällisyys puolestaan on oikein menneiden ennustusten osuus kaikista positiiviseksi ennustetuista havainnoista. Vaikka herkkyys on tärkein, myös tarkkuuden ja täsmällisyyden täytyy olla järkevällä tasolla. Muuten malli voi oppia esimerkiksi luokittelemaan kaikki havainnot pudokkaiksi, jolloin herkkyys olisi 100%.

Cohenin kappa -kertoimella arvoidaan, suoriutuuko malli ennustamisesta yhtään paremmin, kuin jos havainnot jaettaisiin satunnaisesti luokkiin [16]. Cohenin kappa on tilastotieteessä yhteneväisyyden mitta, jota kutsutaan myös kappa-kertoimeksi. Jos kappa on 1, mitattavat ennustajat ovat yhteneväiset. Koneoppimisessa tähän vertailuun valittaisiin mallin tuottamat ennustukset sekä tiedetyt oikeat vastineet kullekin havainnolle. Kappa voi saada arvoja välillä -1 ja 1. Yli 0,6 kappan arvoja pidetään yleensä hyvänä. Tällöin koneoppimismalli on melko yhteneväinen täydellisen ennustajan kanssa. Nolla ja sen alapuolelle jäävät arvot ovat todella huonoja. Tällöin malli ei suoriudu satunnaisuutta paremmin. Alkuperäiset Cohenin ehdottamat raja-arvot kappan tulkitsemiselle on esitetty taulukossa 5.1. Lääketieteen alalla on kuitenkin ehdotettu, että kappan arvoja tulkittaisiin tiukemmin, eikä $0,41 - 0,60$ arvoja pidettäisi edes keskinkertaisena, vaan heikkona

Kappan arvo	Yhteneväisyys
$0 - 0,2$	Mitätön
$0,21 - 0,40$	Heikko
$0,41 - 0,60$	Keskinkertainen
$0,61 - 0,80$	Merkittävä
yli $0,80$	Lähes täydellinen

Taulukko 5.1: Alkuperäinen Cohenin kappan arvojen tulkintaohje [16]

yhteneväisyyden mittana [16]. Tämän vuoksi keskinkertaisiin kappan arvoihin on syytä suhtautua varauksella.

Cohenin kappa κ lasketaan kokeen tuloksista seuraavasti:

$$\kappa = \frac{P_a - P_e}{1 - P_e},$$

jossa P_a on tulosten osuus, jossa molemmat vertailtavat ennustajat olivat samaa mieltä ja P_e on odotettu yhteneväisyys. Käytännössä koneoppimisessa, kun $n = OP + ON + VP + VN$, eli kaikkien havaintojen lukumäärä,

$$P_a = \frac{(OP + ON)}{n}$$

eli sama kuin ”tarkkuus” ja

$$P_e = \frac{\frac{(OP+VN)\cdot(OP+VP)}{n} + \frac{(VP+ON)\cdot(VN+ON)}{n}}{n}.$$

5.2 Mallien valinta

Mallikandidaateiksi valittiin kaksi mallia. Ensiksi valittiin eniten käytetty yksinkertainen malli, päätöspuu, joka on tuottanut muissa tutkimuksissa lupaavia tuloksia pudokkuuden ennustamiseen ja auttanut vaikuttavien tekijöiden määrittämisessä [27, 4, 1, 21, 17, 11, 28]. Myös neuroverkot ovat käytetyimpien mallien joukossa, ja tuottaneet varsin tarkkoja tuloksia, mutta niitä ei valittu mukaan, koska neuroverkkomallista on vaikea saada ulos mallin tärkeimpänä pitämiä muuttujia ja aineisto on suhteellisen pieni [17].

Toiseksi malliksi valittiin XGBoost-malli, joka on vain noin kymmenen vuotta vanhana verrattain uusi malli. Mallin ikä voi olla yksi syy siihen, miksei mallia ole käytetty vielä laajalti, joten sillä voi olla suurempi potentiaali, kuin mitä mallien käyttöaste aihealueen muissa tutkimuksissa tällä hetkellä antaa ymmärtää. XGBoost on kannattava valinta myös siksi, että se käsittelee puuttuvat datakentät automaattisesti ja siitä on mahdollista saada ulos eniten ennusteeseen vaikuttavat muuttujat [2].

Koska on hyödyllistä tietää, mitkä muuttujat vaikuttavat tuloksiin vahvimmin, tähän tutkielmaan valittiin sellaiset mallit, joista tämän tiedon voi saada ulos. Koska päätöspuu on hyvin yksinkertainen malli, se on hyvä verrokkimalli monimutkaisemmalle mallille [4, 17]. Siitä näkee helposti miten päätökset on tehty, joten sitä on käytetty tiedonlouhintatehtäviin [17, 28, 1].

Nämä kaksi mallia ovat keskenään hyvin samanlaiset siinä mielessä, että ne molemmat ovat puumalleja. Päätöspuu on kuitenkin vain yksi puu, kun taas XGBoost yhdistelee useampaa puuta, joista jokainen korjaa edellisen tekemiä virheitä [2, 10]. Yksittäinen puu on usein suuri, kun taas useamman puun mallissa yksittäinen puu saattaa olla hyvinkin pieni, ns. heikko ennustaja [10].

5.3 Suoritettavat kokeet

Kummallakin mallilla suoritetaan kolme koetta, joissa aineisto on käsitelty eri tavoin. Kokeissa käytettävät aineistot on eritelty tarkemmin taulukkoon 5.2. Ensimmäisessä aineistossa yhtä opiskelijaa kohden on vain yksi tietue, jossa opintosuoritukset on koottu jaksoittain omiin kenttiinsä. Aineistossa on 1549 opiskelijaa. Aineistosta on poistettu opiskelijat, joilla ei ole vielä ollut opinto-aikaa kuutta jaksoa. He eivät siis voi vielä tulla määritellyiksi pudokkaiksi, joten he voisivat vääristää mallia. Aineiston muoto on esitetty taulukossa 5.3. Toinen aineisto on muuten sama, mutta siinä on vain vuonna 2020 ja sen jälkeen aloittaneita opiskelioita (863 opiskelijaa). Kolmas aineisto on sama kuin toinen, mutta sisältää lisäksi opiskelijoiden kurssi-ilmoittautumistiedot. Käytännössä tämä tarkoittaa sitä, että jaksokohtaisesti on myös tiedossa sellaisten kurssien määrä, joille on ilmottauduttu, mutta kurssisuoritusta ei ole tullut.

Tämän lisäksi aineistot katkaistaan ajallisesti eri kohdista. Aineistoilla tehdään useampi koe, jossa malli on opetettu joko 1, $1\frac{1}{2}, 2, 2\frac{1}{2}, 3$ ja $3\frac{1}{2}$ vuoden kurssitiedoilla. Eli mukaan pääsee korkeintaan 17 jaksoa vaikka saatavilla olisi enimmillään jopa 38. Tämä johtuu siitä, että lopullista ennustetta tullaan käyttämään opiskelijoiden opintojen alkuvaiheessa niiden opiskelijoiden tavoittamiseen, joilla on riski tulla pudokkaaksi. Vaikka ennusteen tarkkuus ja herkkyys tulee todennäköisesti huomattavasti paranemaan, kun opintoja on takana enemmän, on arvokkaampaa saada kiinni suurin osa mahdollisista pudokkaista aikaisin heikommalla tarkkuudella.

Joskus aineisto on tarpeen tasapainottaa ennen koneoppimismallin koulutusta [11]. Tässä työssä käytettävä aineisto on melko tasapainoinen sekä luokkien jakautumisen, että tietueiden suhteen. Pudokkaita on lähes puolet aineistosta, joten tietueita ei tarvitse tämän takia keinotekoisesti lisätä. Lisäksi tietueita on vain yksi opiskelijaa kohden, joten aineisto on tasapainossa myös tältä osin. Aineistoissa voi olla tyhjiä kenttiä, jos opiskelija ei esimerkiksi ole ollut läsnä jollain lukukaudella. Joidenkin koneoppimiskirjastojen kohdalla tämä aiheuttaisi lisätoimenpiteitä, mutta tässä työssä käytetyt kirjastot käsittelevät puuttuvat

Kokeen tunniste	Aineiston kattamat vuodet	Aineiston koko (opiskelijaa)	Ilmoittautumistiedot mukana
17-22	$2017 - 2022$	1549	Ei
20-22	$2020 - 2022$	863	Ei
20-22_ilm	$2020 - 2022$	863	Kyllä

Taulukko 5.2: Suoritettavissa kokeissa käytettävät aineistot. Lisäksi jokaisen kokeen kohdalla aineisto katkaistaan 1, $1\frac{1}{2}, 2, 2\frac{1}{2}, 3$ ja $3\frac{1}{2}$ vuoden kohdalta.

kentät automaattisesti.

5.4 Optimoinnin tavoite

Malleja kouluttaessa suoritetaan optimointia, kun valitaan parhaita hyperparametrejä mallille ja sopivaa kokoa opetusaineistolle (kuinka monen vuoden ajalta aineistoa otetaan, eli missä vaiheessa opintoja lopullinen ennuste halutaan suorittaa). Kuten on jo useasti todettu, tässä työssä herkkyys on tärkeässä osassa, mutta myös mallin tarkkuudella on väliä. Tämän vuoksi optimoinnissa ei kannata tavoitella suoraan korkeaa tarkkuutta. Yleensä aihealueen tutkimuksissa hyperparametrejä ei ole optimoitu lainkaan tai optimoinnin tavoitemuuttujaa ei ole eritelty [18, 9, 13]. Poikkeuksena on Cheng et al.:n suorittama tutkimus, jossa neuroverkon hyperparametrit on optimoitu tavoitteena mallin tarkkuus [3]. Mallien lopullisen suorituskyvyn arviointiin on käytetty usein joko tarkkuutta tai sellaisia mittareita, jotka tasapainottavat tarkkuutta ja herkkyyttä. Muissa tutkimuksissa lopullinen tavoite on kuitenkin ollut hieman erilainen kuin tässä työssä. Niissä on haluttu minimoida myös se, ettei opiskelijoita luokitella pudokkaiksi turhaan, vaikka se pienentäisi oikein luokiteltujen pudokkaiden osuutta.

Tässsä työssä optimointia varten tulee löytää sopiva mittari, jolla herkkyyttä saadaan painotettua. Yksi tällainen keino on käyttää painotettua f-arvoa [25]. F-arvo yhdistää täsmällisyyden ja herkkyyden halutulla painotuksella β. Kts. yhtälö 5.1. F-arvoa käytetään usein β:n arvolla 1, eli annetaan täsmällisyydelle ja herkkyydelle yhtäläinen painotus. Tällöin puhutaan yleensä F1-arvosta. Tässä työssä halutaan kuitenkin mahdollisuuksien mukaan painottaa herkkyyttä, mikä toteutuu kun $\beta > 1$. Sopiva β:n arvo etsitään mallin koulutusvaiheessa haarukoimalla arvoja jotka toteuttavat $\beta > 1$.

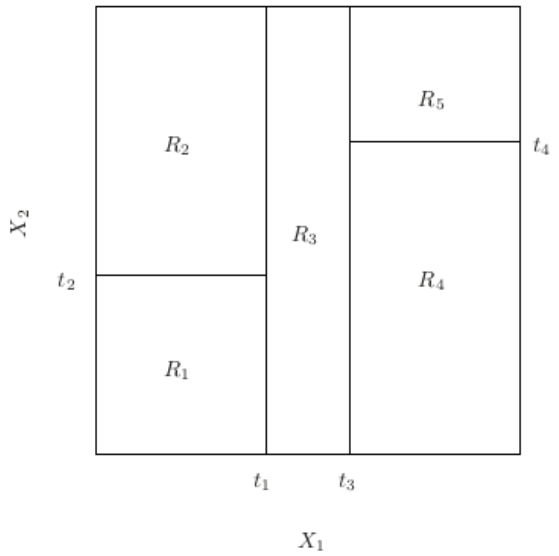
Sarakkeen nimi	Numeroskaala	Selite
pudokas	0,1	Onko pudokas (0=ei, 1=kyllä)
sukupuoli	1,2	Juridinen sukupuoli (1=mies, 2=nainen)
ikä	0-max(int32)	Ikä opintojen alkaessa
lk_ilmo_0	-1-3	1. lukukauden ilmoittautumistieto
lk_ilmo_1	-1-3	2. lukukauden ilmoittautumistieto
lk_ilmo_x	-1-3	seuraavien lukukausien ilmoittautumistiedot
...		(-1=ei op-oikeutta, 0=luvalla p.,
		1=läsnä, 2=poissa, 3=laiminlyöty)
		yht. 14 lk_ilmo saraketta (7 lukuvuotta)
ka_0	0 - max(float32)	aikaisempien opintojen keskiarvo tai NaN
ka_1	0 - max(float32)	1. jakson keskiarvo tai NaN
ka_x	0 - max(float32)	seuraavien jaksojen keskiarvot tai NaN
...		yht. 39 ka_ -saraketta (8,5 lukuvuotta)
op_0	0 - max(float32)	aikaisempien opintojen opintopistemäärä tai NaN
op_1	0 - max(float32)	1. jakson opintopistemäärä tai NaN
op_x	0 - max(float32)	seuraavien jaksojen opintopistemäärät tai NaN
...		yht. 39 op_ -saraketta (8,5 lukuvuotta)
kesk_0	0 - max(float32)	aikaisempien opintojen kesken jätetyt kurssit, lkm*
kesk_1	0 - max(float32)	1. jakson kesken jätetyt kurssit, lkm*
kesk_x	0 - max(float32)	seuraavien jaksojen kesken jätetyt kurssit, lkm*
...		yht. 39 kesk_ -saraketta (8,5 lukuvuotta)*
2_perak_per	0,1	onko 2 peräkkäistä jaksoa ilman suorituksia ilman
		poissaoloilmoitusta (0=ei, 1=kyllä)
3_perak_per	0,1	onko 3 peräkkäistä jaksoa ilman suorituksia ilman
		poissaoloilmoitusta (0=ei, 1=kyllä)
kesk_yht	0 - max(float32)	keskeytettyjen kurssien summa*
hyl_yht	0 - max(float32)	hylättyjen kurssien summa

Taulukko 5.3: Aineisto ennustetta varten. Yksi tietue per opiskelija. Aineisto on koottu opiskelija- ja kurssiaineistosta. ka_ ja op_ -kentissä puuttuva arvo tarkoittaa sitä, että opiskelijalla ei ole suorituksia, mutta on ilmoittanut poissaolostaan. Viimeiset neljä kenttää saavat arvonsa vasta, kun aineisto on katkaistu mallia varten sopivasta kohdasta (esim. 1,5 vuoden kohdalta). *Nämä kentät ovat mukana vain, jos aineisto on koottu opiskelijoista, jotka ovat saaneet opinto-oikeuden vuonna 2020 tai sen jälkeen.

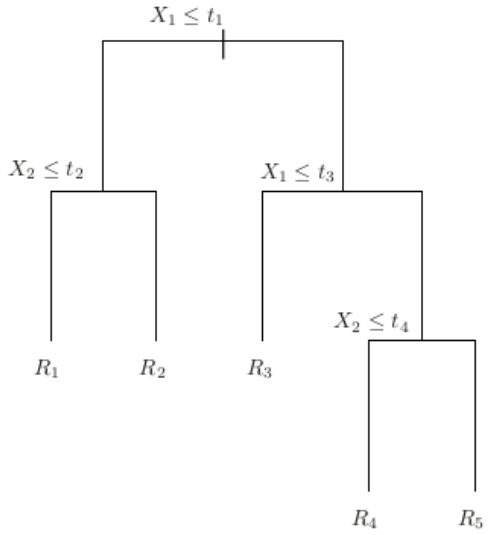
$$F_\beta = \frac{(\beta^2 + 1) \cdot \text{täsmällisyys} \cdot \text{herkkyys}}{\beta^2 \cdot (\text{täsmällisyys} + \text{herkkyys})} \tag{5.1}$$

5.5 Päätöspuut

Päätöspuut etsivät aineistosta rakenteita ja säännönmukaisuuksia luokittelemalla sitä puurakenteen avulla [10]. Puurakenne koostuu toisiinsa linkitetyistä solmuista. Tässä esimerkkinä käytetään alaspäin kasvavaa binääristä puurakennetta. Ylimmästä solmusta (juuresta) puu jakautuu vasempaan ja oikeaan solmuun, jotka jakautuvat molemmat taas kahteen solmuun ja niin edelleen. Jos solmu ei jakaudu, se jää puun lehdeksi, joka määrää lopullisen arvon. Data luokitellaan puun avulla siten, että ensin havaintoa verrataan juureen sijoitettuun sääntöön ja sen perusteella edetään joko vasempaan tai oikeaan solmuun, jossa taas tarkastellaan havainnon arvoja ja edetään joko vasempaan tai oikeaan solmuun. Kun saavutetaan lehti, on kyseinen havainto luokiteltu. Lehdet lopulta määrittävät tietueesta muodostettavan luokan tai muuttujan arvon. Kuvassa 5.1 nähdään aineisto, joka koostuu eri havainnoista, joilla on arvot X_1 ja X_2. Datapisteitä ei ole piirretty kuvaan. Aineisto on jaettu viiteen eri luokkaan. Päätöspuun avulla luokkiin jakaminen tapahtuisi kuvan 5.2 mukaisesti.



Kuva 5.1: Esimerkkiaineisto, jossa esiintyy viisi luokkaa [10].



Kuva 5.2: Esimerkki päätöspuun rakenteesta. Tämä päätöspuu luokittelee aineiston viiteen eri luokkaan [10].

Päätöspuu rakennetaan siten, että yritetään jokaisen solmun kohdalla valita sellainen muuttuja, joka jakaa aineiston parhaiten kahtia [8]. Paremmuuden vertailuun käytetään mallille valittua mittaria. Mittari voi olla esimerkiksi keskineliövirhe tai gini-kerroin. Mittarin valinta riippuu aineistosta ja siitä, onko kyseessä regressio- vai luokittelumalli.

Päätöspuun haasteena on usein ylisovittuminen, eli puu sovittuu liian hyvin harjoitusaineistoon, eikä sitten toimi kovin hyvin ennustettaessa uudella aineistolla [8]. Tämä johtuu osittain mallin rakenteesta: hierarkkinen puurakenne on herkkä aineiston muutoksille. Melko pienikin ero aineistossa voi johtaa siihen, että tietue päätyy niin sanotusti "väärälle polulle" puussa jo lähellä puun juurta aiheuttaen sen, että tulos on hyvin erilainen kuin vierekkäisellä havainnolla oli, koska tälle havainnolle paremmin soveltuvat solmut saattavat sijaita toisella puolella puuta. Ylisovittuminen voidaan havaita esimerkiksi ristiinvalidoinnilla. Ylisovittumista voidaan hallita muun muassa vaatimalla solmulta tiettyä havaintojen minimimäärää jakamista varten, puun "karsimisella" ja puun maksimisyvyyttä rajoittamalla. Näitä keinoja kutsutaan yleisesti koneoppimismallin säännöstelyksi (engl. regularisation).

Tässä työssä käytetään `scikit learn` -kirjaston `DecisionTreeClassifier`-mallia päätöspuun muodostamiseen. Koska kyseessä on luokittelutehtävä, käytetään siihen sopivaa virheen mittaria, ginikerrointa. Ginikerroin mittaa solmun puhtautta, eli käytännössä sitä, kuinka paljon informaatiota säilytetään, jos aineisto jaetaan tietyllä muuttujan arvolla [8].

Ginikerroin voidaan ilmaista seuraavasti:

$$\sum_{k=1}^{K} \hat{p}mk(1 - \hat{p}mk),$$

jossa

$$\hat{p}mk = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k),$$

jossa m on solmu, R_m on "alue" aineistossa ja N_m havainnot, silloin $\hat{p}mk$ on havaintojen osuus, jotka kuuluvat luokkaan k. Näin ollen, jos luokkia on kaksi ja p on havaintojen osuus toisessa luokassa, niin ginikerroin on $2p(1 - p)$.

Kirjasto osaa käsitellä puuttuvat kentät automaattisesti, joten erityistä aineiston valmistelua ei tarvita. Käytännössä malli käsittelee puuttuvan kentän omana kategorisena muuttujanaan. Aineiston normalisointikaan ei ole päätöspuu-mallien osalta tarpeen. Normalisoinnilla tarkoitetaan muuttujien arvojen skaalaamista samaan kokoluokkaan, esimerkiksi välille $0-1$. Ainoa tekninen epävarmuus aineiston kohdalla on sen lievä epätasapainoisuus;

pudokkaita on hiukan vähemmän kuin ei-pudokkaita. Kuitenkin, koska luokkaepätasapaino ei ole huomattava, aineistoa ei tasapainoteta jättämällä osaa ei-pudokkaista pois, tai lisäämällä keinotekoisia tietueita. Mallia pyritään tasapainottamaan sen sijaan parametrien avulla.

Hyperparametreillä mallia saadaan säädettyä aineistoon paremmin sopivaksi. Tällaisilla yksinkertaisilla yksittäisillä päätöspuilla ei kuitenkaan ole kovin runsaasti säätömahdollisuuksia. Kirjastot sallivat yleensä maksimisyvyyden lisäksi ainakin jakamiskohdan paremmuuden arviointiin käytettävän metodin valitsemisen ja vaadittavan havaintomäärän tai painon asettamisen, jotta solmun jako sallitaan [5]. Havainnoille ja luokille voidaan myös antaa painoja, jos aineisto on esimerkiksi epätasapainoinen. Tässä työssä luokat todennäköisesti tulevat tarvitsemaan painotusta, koska positiivinen luokka on tärkeämpi saada ennustettua oikein.

Päätöspuu-mallista saa tulostettua ulos koko puun. Siitä nähdään, miten päätökset tehdään kunkin solmun kohdalla juuresta alkaen kohti lehtiä. Jokaisen solmun kohdalla nähdään, mikä on kunkin solmun ginikerroin, kuinka monta havaintoa solmuun on päätynyt, mikä on luokkien painotettu jakauma ja mikä muuttuja jakaa aineiston seuraavien lehtien välille. Lisäksi mallista saa ulos tiedon, mitkä muuttujat ovat tärkeimpiä ennusteen tuloksen kannalta. Tällaiset muuttujat ovat yleensä hyviä kohdemuuttujan selittäjiä (tässä työssä: pudokas). `scikit learn` -kirjaston `DecisionTreeClassifier`-mallista ulos otetut tärkeimmät muuttujat on määritelty niin sanotun gini-tärkeyden (engl. gini-importance) perusteella [5]. Tärkeys on muuttujan frekvenssi puussa esiintymiselle painotettuna sillä, kuinka hyvin kyseinen jako kulloinkin jakaa havainnot eri luokkiin [8].

5.6 XGBoost

XGBoost-malli kuuluu gradienttitehostettuihin päätöspuihin [2]. Malli tunnetaan myös nimellä ”Extreme gradient boosting”, vapaasti suomennettuna ”äärimmäisen gradienttitehostettu” -malli. Pääpiirteissään malli toimii siten, että se rakentaa useita päätöspuita iteratiivisesti, joista jokainen lisätään yksi kerrallaan malliin. Jokainen puu pyrkii minimoimaan häviöfunktion arvon omalla kohdallaan, korjaten edellisen puun tekemiä virheitä ja lisäten siten mallin tarkkuutta. Optimointiin käytetään sovellettua gradienttimenetelmää (engl. gradient descent). Jokaisen puun kohdalla luokittelija päivitetään ”gradienttiaskeleella”, jonka suuruus riippuu mallin oppimisvauhdista.

XGBoost-luokittelija toteuttaa seuraavat askeleet:

1. **Alustavan mallin luominen.** Luodaan malli, joka antaa saman ennusteen jokaiselle havainnolle. Tässä käytetään yleensä havaintojen kohdemuuttujien keskiarvoa. Binäärisellä luokittelijalla alustavan mallin antama ennuste voi olla esimerkiksi 0.5.

2. **Residuaalien laskeminen mallin ennustuksille.** Jokaiselle havainnolle lasketaan residuaali, joka on käytännössä ennusteen ja aidon havainnon erotus. Jos esimerkiksi havainnon oikea arvo olisi 1 ja alustavan mallin ennuste 0.6, olisi residuaali 0.4.

3. **Häviöfunktion derivaattojen laskeminen.** Lasketaan häviöfunktion ensimmäisen ja toisen asteen derivaatat, yleistäen; gradientti ja Hessen matriisi, joita tarvitaan puun solmuja jakaessa.

4. **Seuraavan puun rakentaminen.** Tässä askeleessa ennustetaan residuaaleja eikä varsinaista kohdemuuttujaa. Nyt pyritään jakamaan aineisto kunkin solmun kohdalla siten, että mahdollisimman samanlaiset residuaalit päätyisivät lopulta samoihin lehtiin. Puu voi näyttää kuten kuvan 5.3 esimerkissä.

5. **Hyödyn laskeminen kullekin tavalle jakaa aineisto kahtia.** Jokaisen solmun kohdalla lasketaan ”hyöty” (engl. gain), joka ilmaisee, kuinka paljon kukin jakamistapa vähentäisi mallin virhettä. Jako voidaan laskea joka ikiselle muuttujalle jokaisten kahden datapisteen välistä, mutta tapa voi käydä aivan liian hitaaksi isommalla aineistolla. Tämän vuoksi voidaan käyttää approksimaatiota hyödyntävää tapaa löytää sopiva jako. Algoritmi tekee kunkin muuttujan arvoista kvantiileita, joista jokaisesta tehdään vain yksi ”hyödyn” laskenta. ”Hyödyn” laskemisessa käytetään aikaisemmin häviöfunktiosta laskettua gradienttia ja Hessen matriisiia. Gradientin hyödyntäminen laskennassa vaikuttaa siihen, että häviötä saadaan systemaattisesti minimoitua. Jokainen askel tuo lähemmäksi häviöfunktion kohtaa, jossa sen derivaatta on nolla – edetään siis oikeaan suuntaan. Hessen matriisi puolestaan auttaa optimoimaan seuraavan askeleen pituutta tarjoamalla tietoa funktion paikallisesta muodosta. Hessen matriisia käytetään ikään kuin sakkona, eli hillitsemään yliampuvia muutoksia ja vähentämään siten ylisovittumista.

6. **Puun kasvattaminen.** Puuhun lisätään uusia solmuja, kunnes jokaisessa lehdessä on enää yksi havainto, tai kunnes puun maksimisyvyys on saavutettu, tai saavutetaan jokin muu rajoite.

7. **Karsiminen.** Puun rakentamisen jälkeen sitä karsitaan tarpeen mukaan. Mallille annettujen parametrien perusteella karsimista voidaan tehdä enemmän tai vähem-

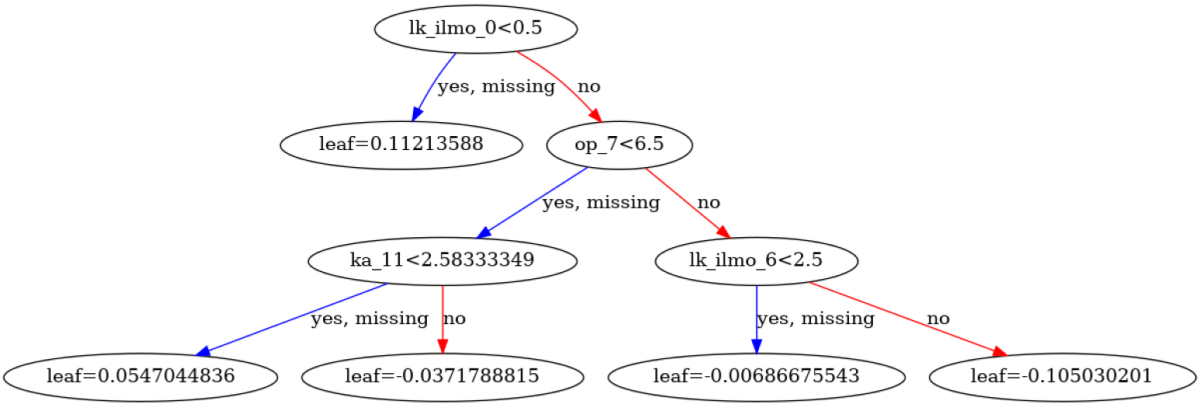
män. Kuitenkin, jos ”hyöty” -arvoksi tulee negatiivinen luku, tullaan lehti aina karsimaan pois, vaikka karsimisparametrejä ei olisi erikseen asetettu mallille.

8. **Mallin päivittäminen.** Uuden puun antamat ennusteet lisätään edelliseen ennusteeseen kerrottuna oppimisvauhdilla, joka on asetettu mallia luodessa. Oletusarvo on usein 0.3.

9. **Seuraavan puun rakentaminen.** Toistetaan kohtia 2-9, mutta alustavan mallin sijaan verrokkina käytetään aina edellisen mallin ennusteita, kunnes lopetuskriteeri täyttyy. Kriteeri voi olla esimerkiksi puiden maksimimäärä, tehostuskierrosten lukumäärä tai jokin muu.

XGBoost-malli toimii monentyyppisellä aineistolla. Puurakenteen vuoksi muuttujat eivät vaadi normalisointia, eivätkä puuttuvat kentät tuota ongelmia. XGBoostia toteuttavissa kirjastoissa puuttuvat kentät käsitellään omana luokkanaan tavalla, jota havainnollistetaan kuvassa 5.3. Tässä työssä käytetään kirjastoa `xgboost` [7]. Hyperparametrejä säätämällä mallin rakenteeseen voidaan vaikuttaa merkittävästi ja saada se näin sopimaan mitä erilaisimpiin aineistoihin sopivaksi. Hyperparametreillä esimerkiksi rajoitetaan puiden määrää, syvyyttä tai solmun jakamista, ellei hyöty ole riittävän suuri.

XGBoost-mallista saadaan päätöspuun tavoin helposti ulos muuttujat, jotka esiintyvät puissa usein, eli jotka monessa tilanteessa jakavat aineistoa hyvin kahteen luokkaan [7]. Samoin jokaisen puun pystyy tulostamaan, niin että nähdään, minkä muuttujan arvojen perusteella solmun jako kahteen uuteen solmuun tehdään. Lisäksi tähän malliin pystyy soveltamaan `shap`-kirjaston tarjoamia analysointityökaluja [24]. Kirjaston **shap-arvo** jäljittelee peliteorian Shapley-arvoa, joka kertoo kunkin muuttujan (peliteoriassa pelaajan) vaikutuksesta lopputulokseen (peliteoriassa voittoon) [15]. Mitä suurempi Shapley-arvon itseisarvo on, sitä suurempi on kyseisen muuttujan vaikutus lopputulokseen. Käytännössä shap-arvon laskemista varten ennuste kokeillaan tehdä kaikilla mahdollisilla syöteyhdistelmillä ja mitataan, miten ennuste muuttuu kullakin yhdistelmällä.



Kuva 5.3: XGBoost-mallin luoma yksittäinen puu.

6 Tulokset

6.1 Päätöspuu

Tässä työssä käytetään `scikit learn` -kirjaston `DecisionTreeClassifier`-mallia Python-kielellä [5]. Kirjastolla suoritetaan kolme koetta, joissa jokaisessa aineisto on katkaistu 1, $1\frac{1}{2}, 2, 2\frac{1}{2}, 3$ ja $3\frac{1}{2}$ opiskeluvuoden kohdalta. Yhteensä malleja luodaan siis 18 kappaletta.

6.1.1 Hyperparametrien valinta

Hyperparametrien optimointiin otettiin säädettäväksi puun maksimisyvyys, positiivisen luokan painotus, lehdiltä vaadittava havaintomäärä ja havaintomäärä, joka vaaditaan solmulta, jotta se voidaan jakaa. Hyperparametrit optimoitiin tavoitteena maksimoida F-arvo (kts. aliluku 5.4). F-arvolle haarukoitiin ensin sopiva β:n arvo, kouluttamalla mallia eri betan arvoilla $1 - 2,5$. Jokaisella eri arvolla hyperparametrit optimoitiin `optuna`-kirjastoa käyttäen. Malli koulutettiin 150 kertaa eri hyperparametriyhdistelmillä ja valittiin 5-kertaisella ristiinvalidoinnilla parhaan f-arvon antanut β:n arvo. Hyvin nopeasti kävi selväksi, ettei mallin taivuttaminen F-arvon avulla tuottanut haluttua jakaumaa herkkyyden ja täsmällisyyden välille. Jo hyvin pieni muutos β:n arvossa yhdestä eteenpäin lisäsi herkkyyttä niin merkittävästi, että jopa neljä viidestä havainnoista saattoi tulla ennustetuksi pudokkaiksi. Tämän vuoksi β:n arvoksi valittiin 1, jossa herkkyys ja täsmällisyys ovat tasapainossa.

Hyperparametrien optimoinnissa käytettävät arvoalueet ovat nähtävillä taulukossa 6.1. Hyperparametrit valittiin jokaiselle mallille erikseen optimoimalla F-arvoa samaan tapaan kuin β-arvoa valittaessa. Optimointi tehtiin rakentamalla 500 eri mallia, ja validoimalla jokainen malli 5-kertaisella ristiinvalidoinnilla. Ristiinvalidoinnista laskettiin F-arvon keskiarvo. Näistä kokeista valittiin parhaiten suoriutuneet hyperparametrit jokaiselle aineistolle.

Parametri	Mahdolliset arvot	Selite
`class_weight`	$\{0:1, 1:(1,10)\}$	Luokkien paino. Positiivinen luokka painotetaan samanarvoiseksi tai tärkeämmäksi.
`min_samples_split`	$2-10$	Minimimäärä havaintoja, jotta solmu jaetaan.
`min_samples_leaf`	$1-10$	Minimimäärä havaintoja, jotka päätyy lehteen.

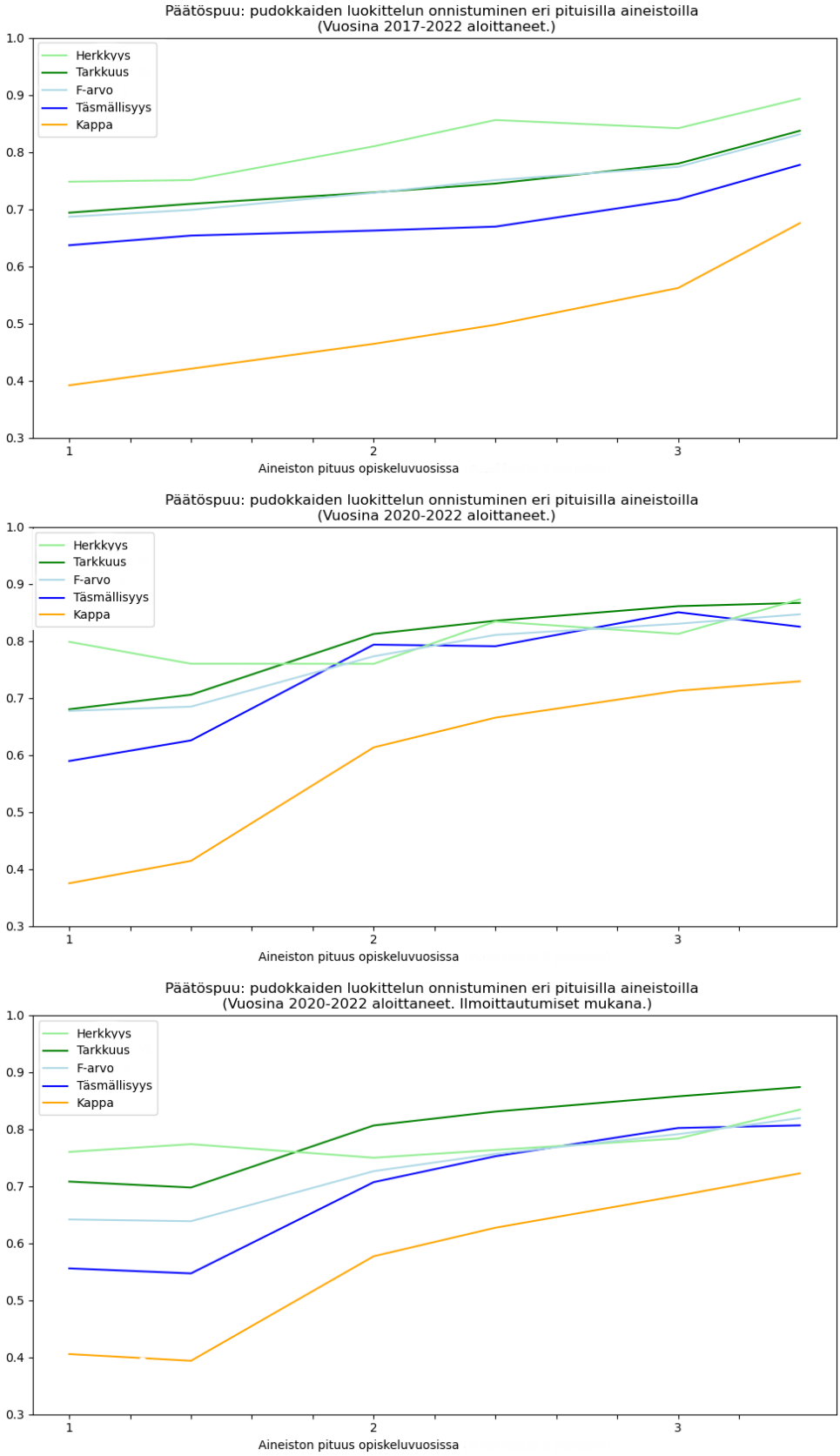
Taulukko 6.1: Hyperparametrit, joista valittiin parhaiten suoriutuvat parametrit päätöspuu-mallin koulutusta varten. Loput parametrit käyttivät kirjaston oletusarvoja.

6.1.2 Mallien suoriutuminen

Päätöspuumallilla ei saatu kovin luotettavia tuloksia aikaan pudokkaiden ennustamisessa, etenkään jos ennuste tehtiin opintojen alussa. Herkkyyttä ei saatu painotettua siten, että muut arvot olisivat myös jääneet järkevälle tasolle. Herkkyyttä ei siis lopulta varsinaisesti painotettu, vaan käytettiin F1-arvoa. Tämän vuoksi malli jäi melko tasapainoiseksi ennustajaksi siinä mielessä, että vääriä positiivisia ja vääriä negatiivisia arvoja tulee suurin piirtein sama osuus. Viivakaavioista 6.1 nähdään, että mallien tarkkuus parani, mitä myöhemmin opintojen aikana ennuste tehtiin.

Kaikilla päätöspuumalleilla tärkeimpiä tekijöitä olivat aina ajallisesti aineiston loppupään opintopistemäärät ja keskiarvot, eli toisin sanoen, mikä on ollut opiskelijan viime aikainen opintomenestys. Myös ensimmäisen lukukauden ilmoittautumistieto nousi osalla kurssi-ilmoittautumistietoja sisältävillä päätöspuumalleilla tärkeäksi tekijäksi. Käytetyllä päätöspuumallilla ”tärkeimmät tekijät” tarkoittavat käytännössä muuttujia, joilla on korkein gini-tärkeys (kts. aliluku 5.5).

Mallit suoriutuivat kuitenkin heikohkosti yleisesti ottaen, joten mallien tärkeimpiä tekijöitä ei voi varauksetta suoraan rinnastaa todellisuuteen. Esimerkiksi siis jos mallin näkökulmasta viimeisin saatu opintopistemäärä on tärkeä tekijä ennustetta tehdessä, se ei välttämättä ole tärkeä tekijä todellisuudessa, koska mallin ennustetarkkuus on matalahko. Tulokset ja mallien tärkeimmät tekijät on kuvattu liitteessä B.



Kuva 6.1: Päätöspuu-mallin suoriutuminen kolmella eri aineistolla, kun malli on koulutettu eri pituisilla aineiston otoksilla.

6.2 XGBoost

XGBoost-mallia käyttäen suoritetaan kolme koetta, joissa jokaisessa aineisto on katkaistu 1, $1\frac{1}{2}, 2, 2\frac{1}{2}, 3$ ja $3\frac{1}{2}$ opiskeluvuoden kohdalta. Yhteensä malleja luodaan siis 18 kappaletta. Tässä työssä XGBoost-mallia käytetään `xgboost`-kirjaston kautta Python-ohjelmointikielellä [7]. Mallina käytetään kirjaston `XGBClassifier`-luokittelijamallia. Kirjasto sallii useiden hyperparamerien asettamisen.

6.2.1 Hyperparametrien valinta

Hyperparametrien optimointia varten ensimäiseksi etsittiin sopiva arvo β:lle F-arvoa (kts. aliluku 5.4) varten. XGBoost-malli ajettiin erikseen jokaiselle arvolle välillä $1,0 - 5,5$, askelvälin ollessa $0,5$. β:n arvo $5,5$ antaa F-arvolle jo hyvin lähelle saman arvon kuin herkkyys on, joten koetta ei tuntunut tarpeelliseksi tehdä suuremmilla β:n arvoilla. Alustavien tulosten jälkeen malli ajettiin vielä β:n arvoille $1,0 - 2,5$ askelvälin ollessa $0,2$. Jokaisen eri β:n arvon kohdalla mallin hyperparametrit optimoitiin F-arvon suhteen. Tämä tapahtui luomalla 150 eri mallia erilaisilla hyperparametreillä. Optimointiin käytettiin `optuna`-kirjaston optimointimetodeja. Lisäksi jokaisen eri hyperparametrikokoelman kohdalla suoritettiin ristiinvalidointi, jossa aineisto jaetaan 5 osaan, joista jokainen toimii vuorollaan testiaineistona, kun malli koulutetaan lopuilla 4 osalla. Lopullinen F-arvo kullekin hyperparametrikokoelmalle oli ristiinvalidoinnista saatujen F-arvojen keskiarvo. Ristiinvalidoinnissa tallennettiin myös herkkyys, täsmällisyys ja tarkkuus. Hyperparametrivaihtoehdoiksi oli asetettu taulukon 6.2 reunaehdot. Hyperparametreissä painotetaan erityisesti positiivisten havaintojen tärkeyttä ja hidasta oppimista, jolla pyritään välttämään ylisovittumista. Koska aineisto on melko pieni, laitteiston suorituskyvystä ei tarvinnut huolehtia, vaan voitiin käyttää hitaita menetelmiä. Esimerkiksi `tree_method:exact` tarkoittaa sitä, että malli harkitsee jokaisessa solmussa muuttujan kohdalla sen jakamista kahteen oksaan joka ikisen aineistossa esiintyvän havainnon arvon väliltä.

β:n arvoksi valikoitui edellä mainittujen testien jälkeen $1,6$. Tällä arvolla F-arvo suosii herkkyyttä, mutta ei jätä täsmällisyyttä (ja sen myötä tarkkuutta) huomiotta. Kun $\beta = 1,6$ ja jos aineiston koko on ennustettaessa 200 opiskelijaa, on ”väärien hälytysten” määrä vielä kohtuullisissa mitoissa: noin 38 opiskelijaa saa väärän hälytyksen (kun aineistoa on 2 vuoden ajalta). Suuremmilla arvoilla saattaa tulla 200 opiskelijasta jopa yli 50 määriteltyä virheellisesti pudokkaaksi, vaikka lähes 90% oikeista pudokkaista saatiinkin

Parametri	Mahdolliset arvot	Selite
`scale_pos_weight`	$0,1 - 5,0$	Positiivisten havaintojen painotus.
`eta`	$0,01 - 0,3$	Oppimisvauhti.
`max_depth`	$3 - 18$	Yksittäisen puun maksimisyvyys.
`min_child_weight`	$0 - 10$	Paino, joka lehdeltä vaaditaan, jotta se voidaan luoda.
`gamma`	$0 - 0,2$	Häviön vähentämisen määrä, joka lehdeltä vaaditaan, jotta se voidaan luoda.
`objective`	`binary:logistic`	Tavoite.
`eval_metric`	`logloss`	Häviöfunktio.
`tree_method`	`exact`	Muuttujan katkaisukohdan arpomistapa lehtiä luodessa.
`base_score`	$0,45$	Alustavan mallin ennuste kaikille havainnoille "onko pudokas".
`n_estimators`	80	Rakennettavien puiden yhteismäärä.

Taulukko 6.2: Hyperparametrit joista valittiin parhaiten suoriutuvat parametrit XGBoost-mallin koulutusta varten. Loput parametrit käyttivät kirjaston oletusarvoja.

kiinni (herkkyys 0,9). $\beta = 1,6$ laskee herkkyyttä hieman, mutta pudottaa jo reilusti ”väärien hälytysten” osuutta. Koska β:n sopiva arvo riippuu täysin mallin käyttäjän tahdosta, jätetään lopulliseen ohjelmaan tämä arvo käyttäjän päätettäväksi. Jos pudokkaalle kohdennettavat tukitoimet ovat matalakustanteisia, voi arvoa halutessaan kasvattaa, ja näin ollen saada useamman pudokkaan kiinni.

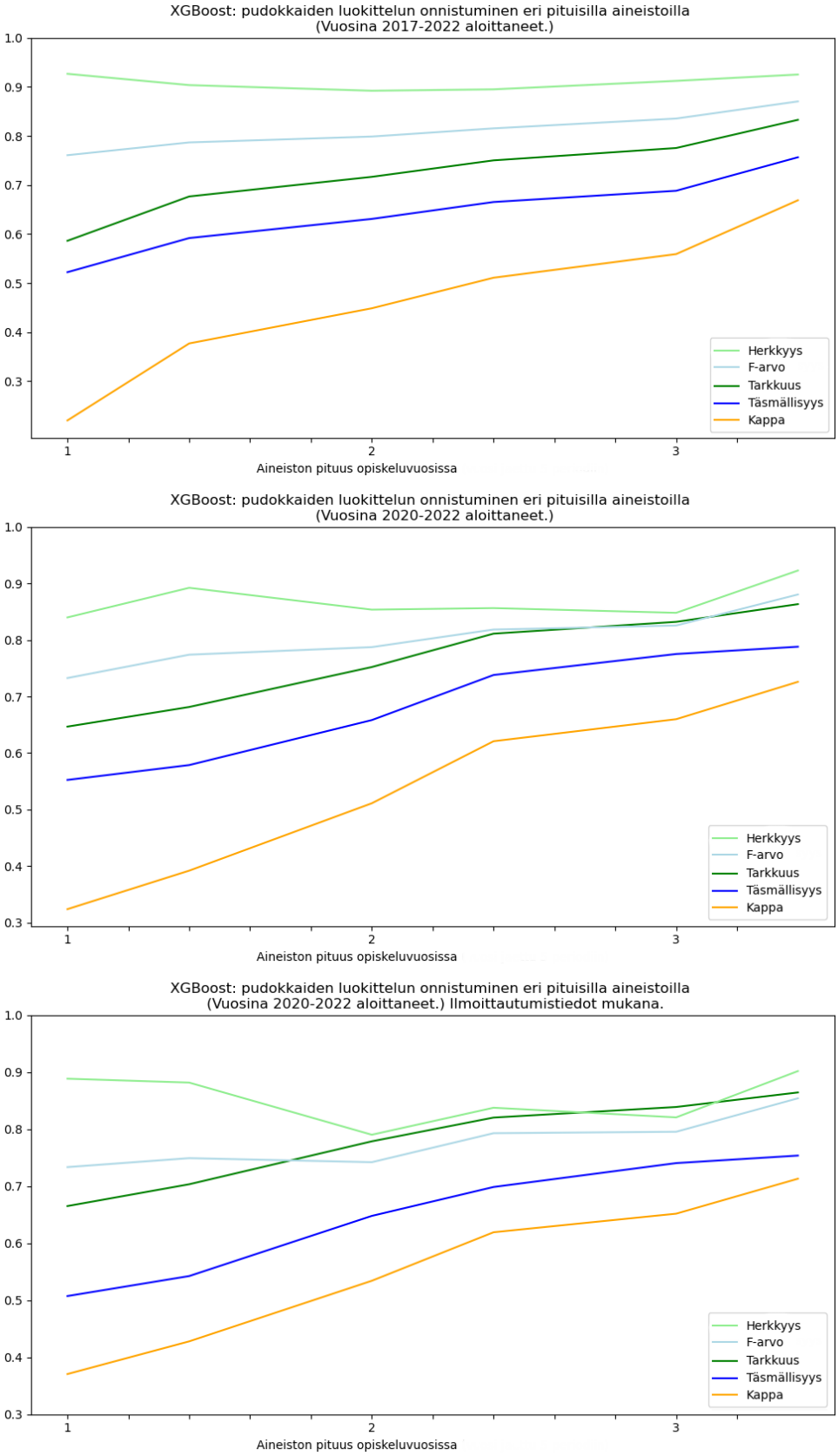
Kun β:lle oli määritelty sopiva arvo, voitiin hyperparametrit optimoida vielä tarkemmin. 150 ajon sijaan tehtiin 500 ajoa per aineisto, joista jokaisella ajolla suoritettiin sama ristiinvalidointi kuin mitä edellä on kuvattu. Tällä tavoin valittiin jokaiselle aineistolle parhaat hyperparametrit. Hyperparametrit ja niiden reunaehdot esitetään taulukossa 6.2.

6.2.2 Mallien suoriutuminen

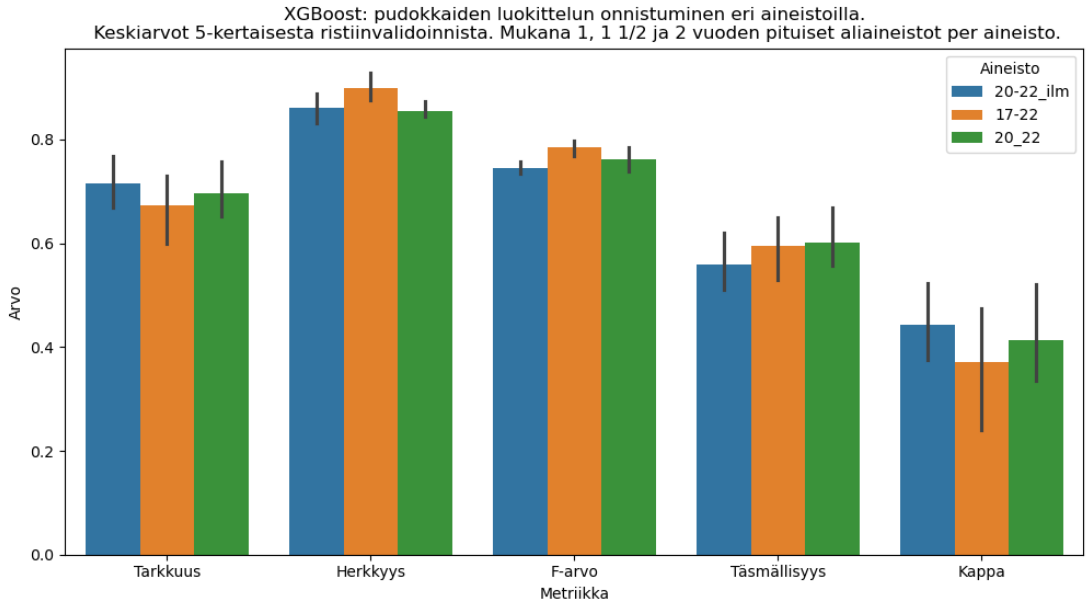
Kaikilla mittareilla mallien suorituskyky oli odotetusti sitä parempi, mitä pidemmältä ajalta opiskelijasta oli aineistoa saatavilla (kts. viivakaaviot 6.2). Ensimmäisen vuoden opiskeluiden jälkeen ennuste oli kaikilla kolmella aineistotyypillä heikko: Cohenin kappa -kerroin oli hyvin matala ($0,22 - 0,37$). Kappa oli kaikista heikoin aineistolla 17-22, kes-

kiarvoisesti vain 0,22, mikä tarkoittaa, että ennuste on vain hiukan parempi kuin luokkien suhteen mukaan tehty satunnainen ennuste. Kappa kuitenkin kohosi nopeasti kun opiskelijasta oli aineistoa pidemmältä ajalta, saavuttaen pisimmällä (3,5 vuoden mittaisella) aineistolla jo hyvän tason, noin $0,67 - 0,73$. Herkkyys on kaikissa kokeissa heti korkealla, mikä tarkoittaa sitä, että pudokkaista on saatu luokiteltua hyvin suuri osa oikein. Kaikissa kokeissa herkkyyden keskiarvo ristiinvalidoidessa oli yli 79%, korkeimmillaan jopa 93%. Kuitenkin, lyhyimmillä aineistoilla varjopuolena on selvästi suurempi määrä "vääriä hälytyksiä", eli pudokasleiman saa moni muukin, kuin vain oikeasti keskeyttämisvaarassa olevat opiskelijat. "Väärien hälytysten" suuresta määrästä kertoo "täsmällisyys"-mittarin matala arvo.

Lyhyimmillä aineistoilla ($1 - 2$ -vuotta opiskelua ennen ennustetta) ilmoittautumisaineiston mukanaolo (aineisto 20-22_ilm) nosti kappa-kerrointa jonkin verran aineistoihin 17-22 ja 20-22 verrattuna, eli ennuste voi siinä mielessä olla tällä aineistolla luotettavampi. Lyhyimpien aineistojen tulokset nähdään tarkemmin pylväskaaviossa 6.3. Toisaalta suurin aineisto (17-22) pärjäsi keskiverrosti parhaiten, jos tarkastellaan, millä aineistoilla tulokset olivat tasaisimpia eri pituisten aineistojen välillä, kuten myös viivakaavion tasaisuudesta voi havaita 6.2. Tätä ehkä selittää aineiston koko; isommalla koulutusaineistolla mallin ylisovittuminen vähenee.



Kuva 6.2: XGBoost-mallin suoriutuminen kolmella eri aineistolla, kun malli on koulutettu eri pituisilla aineiston otoksilla.



Kuva 6.3: XGBoost-mallin suoriutuminen lyhyimmillä aineistoilla (1, $1\frac{1}{2}$ ja 2 vuotta opiskelua ennen ennustamista). Mustat viivat kuvaavat varianssia.

6.3 Tulosten tulkinta

XGBoost tuotti parempia tuloksia ennustettaessa pudokkuutta, kuin pelkkä yksi päätöspuu, mikä ei sinänsä ole yllättävää, koska aineisto on melko moniulotteista. Yksittäinen päätöspuu saattaa ylisovittua liian herkästi opetusaineistoon. XGBoost-mallilla onnistuttiin myös painottamaan mallin herkkyyttä, jolloin väärien negatiivisten tulosten määrä on suhteessa pienempi. XGBoostin ennustuskyky oli myös yleisesti jonkin verran parempi. Esimerkiksi vaikeasti ennustettavalla alueella – kun opintoja oli takana vain 1,5 vuotta – yksittäinen päätöspuu sai tarkkuudeksi noin 0,71, herkkyydeksi 0,75 ja täsmällisyydeksi 0,65. XGBoostilla vastaavat arvot olivat 0,68; 0,90 ja 0,59. Täsmällisyys ja tarkkuus ovat siis lähes samalla tasolla, mutta herkkyys on huomattavasti parempi. Molemmille malleille yhteistä oli se, että suurin aineisto tuotti parhaimpia ja tasaisimpia tuloksia (aineisto koodilla 17-22). Toisaalta ilmoitusaineiston mukana oleminen (aineisto 20-22_ilm) nosti kappa-kerrointa XGBoost-kokeissa noin yhdellä kymmenyksellä vaikeasti ennustettavalla alueella (1-2 vuoden pituisilla aineistoilla).

Johtopäätöksiä pudokkuuteen liittyvistä tekijöistä tehdään pääasiassa XGBoost-mallin tulosten pohjalta, koska se suoriutui yksittäistä päätöspuuta paremmin.

Opintomenestyksellä on merkitystä. Kaikilla malleilla eniten puissa esiintyvä muuttuja oli jonkin tietyn aikajakson opintopistemäärä ja keskiarvo. Tämä ”tietty aikajakso” luonnollisesti vaihteli malleittain, koska malleilla oli eri pituiset opetusaineistot. Useimmiten suurin selittäjä oli aineiston viimeisen jakson opintopistemäärä, eli jos tehtiin ennustetta vaikkapa kaksi vuotta opiskelleille, niin jaksolla, joka osuu toisen vuoden kohdalle, oli eniten merkitystä. Lisäksi tuloksissa toistui se, että noin puoli vuotta ennen viimeistä jaksoa saadulla jaksokeskiarvolla oli myös suuri merkitys. Kertooko tämä siitä, että ensin arvosanat alkavat laskea, sitten opintopisteet vähetä ja lopulta loppuvat kokonaan?

Shap-arvoja (kts. aliluku 5.6) tarkastellessa melko moni muuttuja jakoi opiskelijoita selkeästi. Kaikkien mallien shap-arvot ovat nähtävissä liitteessä A. Kuvaajaa luetaan siten, että positiiviset shap-arvot indikoivat pudokkuutta, kun taas negatiiviset ei-pudokkuutta. Punainen arvo tarkoittaa korkeaa muuttujan arvoa, eli esim. sukupuolen kohdalla punainen on 2 (nainen) ja sininen on 1 (mies). Harmaat ovat puuttuvia arvoja. Mitä enemmän erillään ja kauempana eri väriset arvot ovat toisistaan, sitä selkeämmin muuttuja erottelee eri luokkia. Sukupuoli oli monella mallilla tärkeä jakaja. Shap-arvoista nähdään, että eri väriset havainnot jakautuvat selkeästi ja ovat hieman erillään. Miessukupuoli indikoi shap-arvon perusteella pudokkuutta enemmän kuin naissukupuoli. Puiden rakennetta

tarkastelemalla ero tulee usein tilanteessa, jossa ensin on tarkasteltu opintomenestystä. Korkeampi ikä vaikuttanee myös opintojen jatkumiseen hieman negatiivisesti sen perusteella, että osa malleista piti ikää tärkeänä piirteenä ja myös sen shap-arvo jakautui siten, että vanhemmilla painottui pudokkuus hiukan enemmän. Puun rakenteita tarkastelemalla ikä jakoi opiskelijat kahteen leiriin useimmiten $27 - 33$ ikävuoden kohdalta. Sukupuoli tai ikä eivät kuitenkaan esiintyneet puissa lähellä juurta, vaan jakoivat opiskelijoita vasta tärkeämpien tekijöiden (mallista riippuen tiettyjen ajanjaksojen opintopistemäärät ja keskiarvot) jälkeen.

Aineistossa 20-22_ilm (ilmoittautumisaineisto mukana), keskeytettyjen kurssien määrä nousi useammalla mallilla shap-arvoja tarkastellessa tärkeähköksi tekijäksi, korvaten hylättyjen kurssien määrän tärkeydessä. Aineistoissa 17-22 ja 20-22, joissa tätä tietoa ei ollut saatavilla, hylättyjen kurssien määrä oli tärkeydessä korkeammalla. Tästä voisi tehdä varovaisen johtopäätöksen, että mahdolliset pudokkaat ennemmin jättävät kursseja kesken, kuin odottavat kurssin loppuun asti ja saavat hylätyn.

Hieman yllättävä löydös on, että peräkkäiset poissaolot (3 peräkkäistä jaksoa poissaolevana ilman ilmoitusta) olivat joillain malleilla hyvinkin selkeitä tärkeitä tekijöitä (esimerkiksi aineisto 17-22 ennuste 2,5 ja 3 vuoden opiskeluiden jälkeen), mutta toisilla malleilla tämä muuttuja ei esiintynyt lainkaan tärkeiden tekijöiden joukossa. Esimerkiksi muuttuja ei ollut lainkaan mukana aineistojen 20-22 ja 20-22_ilm tärkeimmissä tekijöissä shap-vertailussa, kun ennustettiin 2,5 vuoden opiskelujen jälkeen. Näin raju vaihtelu saattaa kertoa opetusaineistojen pienehkön koon aiheuttamista ongelmista.

6.4 Rajoitteet

Aineistoon liittyy paljon epävarmuustekijöitä. Sen lisäksi, että aineisto on melko pieni, esimerkiksi ei-valmistumisesta ei ole saatavissa täysin varmaa tietoa, koska opiskelija voi aina anoa opinto-oikeutensa takaisin. Lisäksi ilmoittautumisaineistosta johdetut päätelmät on melko epävarmaa tietoa, sillä joillekin kursseille voi ilmoittautua aikaisessa vaiheessa. Tämän vuoksi tässä työssä käytetty ilmoittautumisaineiston käsittelytapa voi johtaa harhaluuloon, että opiskelija on keskeyttänyt kurssin, vaikka tosiasiassa kurssi ei vaan ollut vielä alkanut lainkaan. Kurssisuoritus voi myös tulla järjestelmään myöhässä. Jos ajatellaan vaikkapa, että opiskelija ilmoittautuu kurssille elokuussa ja kurssi päättyy lokakuussa, mutta arvosana tulee vasta marraskuussa järjestelmään, jää lokakuun lopussa kerättyyn aineistoon käsitys, että opiskelijan kurssi jäi kesken. Ilmoittautumisia pitäisi

mahdollisesti jättää pois aineistosta melko pitkältä ajalta tarkasteluaikavälin loppupäästä. Tämä ei kuitenkaan ole aivan suoraviivaista, sillä toisaalta ilmoittautuminen myös kuvastaa opiskelijan aikeita edetä opinnoissa. Tämän vuoksi tässä työssä karsintaa ei tehty. Ei ole tiedossa, kumpi tapa palvelisi koneoppimismallin suoriutumista paremmin. Ilmoittautumisaineistosta voisi myös johtaa useampia erilaisia tietoja aineistoon kuin mitä tässä työssä tehtiin: esimerkiksi ilmoittautumisten määrä jaksoittain, tai onko joillekin kursseille ilmoittauduttu useamman kerran opintojen aikana.

Opiskelijat saattavat myös suorittaa sellaisia kursseja, jotka näkyvät aineistossa hyvänä opintomenestyksenä, mutta eivät edistä valmistumista. Esimerkiksi tutkintoon mahtuvien vapaavalinnaisten kurssien määrä voi olla jo täynnä ja opiskelija suorittaa ylimääräisiä toisen alan opintoja. Tämän vuoksi voisi kokeilla jättää kurssiaineistosta pois kaikki kurssit, jotka eivät ole koulutusohjelman omia pakollisia kursseja. Tällöin aineisto kertoisi suoraan opiskelijan edistymisestä juuri tietojenkäsittelytieteen kandiohjelmassa, eikä opinnoissa yleisesti. Tässä olisi kuitenkin ongelmana se, että tietojenkäsittelytieteen kandiohjelmaan kuuluu merkittävä määrä valinnaisia ja ”toisen tieteenalan” opintoja, joten paljon sellaisia kursseja jäisi pois, jotka opiskelija lopulta tulee liittämään tutkintoonsa.

6.5 Tulosten vaikutus ohjelmistoon

XGBoost-malli suoriutui keskiverrosti paremmin ennustamisesta kuin yksittäinen päätöspuu. Tämän vuoksi XGBoost malli on valittu ohjelmistossa käytettäväksi koneoppimismalliksi. Tämä työ ei kuitenkaan antanut selkeitä tuloksia siitä, mikä aineisto ja mikä optimoinnin tavoitteessa käytetty β-arvo olisi selkeästi paras ennustamisessa. Se, mitä ominaisuutta haluaa painottaa eniten, ja missä vaiheessa opintoja ennustus ylipäänsä kannattaa tehdä, on enemmän mielipideasia. Tulosten perusteella päätettiin antaa ohjelman käyttäjän tehdä hieman enemmän valintoja koulutettavan mallin suhteen kuin mitä alun perin oli tarkoitus. Käyttäjä saa valita kuinka paljon herkkyyttä painotetaan. Tämä tapahtuu asettamalla β:n arvo. Oletusarvoksi on valittu 1,6 jos käyttäjä ei halua tehdä valintaa itse. Lisäksi käyttäjä saa valita opiskeluiden ajankohdan, jolloin ennuste tehdään, ja onko ilmoittautumisaineisto mukana vai ei. Käytännössä ohjelmistolla pystyy siis suorittamaan kaikki XGBoost-koeyhdistelmät, joita tässä työssä suoritettiin, ja vielä enemmäkin β:n arvoa säätämällä.

Aineisto 17-22 suoriutui hiukan tasaisemmin tuloksin kuin muut aineistot, joten sitä käytetään ohjelmiston oletusmallin kouluttamiseen. Oletusmallin β:n arvoksi valittiin 1,6 ja

ennustamisen ajakohdaksi $1,5$ vuotta opiskeluiden alkamisesta. Tämä valinta perustuu siihen, että se on lyhyin mahdollinen aikaväli, jossa tulokset alkoivat olla mahdollisesti kohtuullisella tasolla. Oletusmalli on kuitenkin helppo yliajaa uudelleen kouluttamalla. Esikoulutettu oletusmalli vain helpottaa käyttöönottoa.

7 Yhteenveto

Tässä työssä tärkeimmiksi pudokkuuden selittäjiksi nousivat keskiarvot ja opintopistemäärät. Tulokset olivat siis saman suuntaisia kuin aikaisemmissa tutkimuksissa. Kuten taulukosta 2.2 nähdään, muissa tutkimuksissa opintojen keskiarvo sekä hyväksyttyjen kurssien lukumäärä ovat olleet yleisimpiä pudokkuuden selittäjiä.

Aineistoa analysoidessa nähtiin selvästi, että opintopistemäärät ja keskiarvot olivat pudokkailla ja ei-pudokkailla erilaiset (kts. aliluku 3.4). Pudokkaat myös herkemmin keskeyttivät joitain kursseja, kuten tietorakenteet ja algoritmit, laskennan mallit ja raja-arvot. Pudokkuus siis näkyi aineistossa ainakin keskiarvoisesti. Kuitenkaan koneoppimismallit eivät pystyneet ennustamaan pudokkuutta korkealla tarkkuudella etenkään opintojen alkuvaiheessa.

Alkuperäisenä toiveena oli, että mallilla voisi ennustaa ”mahdollisimman aikaisessa vaiheessa” opiskelijan todennäköisyyttä tulla pudokkaaksi. Tehtyjen kokeiden perusteella vaikuttaa siltä, että ennustaminen vähemmällä kuin 1,5 vuoden opintoaineistolla ei ole kannattavaa; kappa-kerroin on niin matala, että ennusteen tulos on korkeintaan auttava – ainoastaan hiukan parempi kuin luokkaosuuksiin perustuva satunnainen jako. 1,5 vuoden opintojen jälkeen tai vielä myöhemmin tehdyistä ennusteista sen sijaan saattaa olla konkreettista hyötyä pudokkaiden tunnistamisessa. Ainakin tukitoimet voidaan jättää tekemättä melko luotettavasti niiden opiskelijoiden kohdalla, joita ennuste ei leimaa pudokkaaksi.

Opintomenestys, eli suuri määrä opintopisteitä tai korkea keskiarvo, oli kaikille malleille tärkein tekijä ennustamisessa, kun tarkasteltiin muuttujien esiintyvyyttä puissa, kykyä jakaa aineistoa kahtia tai vertailtiiin shap-arvoja. Huono opintomenestys voi olla opintojen keskeytymisen syy, mutta yhtä hyvin taustalla voi olla jokin toinen syy, joka vaikuttaa opintojen keskeytymiseen ensisijaisesti, mutta heijastuu myös opintomenestykseen.

Tulosten heikkouden perusteella voidaan myös päätellä, että on paljon sellaisia keskeyttämisen syitä, jotka eivät liity suoraan opintoihin. Joukossa täytyy olla jonkin verran sellaisia pudokkaita, joiden opinnot sujuu hyvin, kunnes tapahtuu yhtäkkiä jotain, jonka takia opinnot jäävät kokonaan, eikä tämä ehdi näkyä opintomenestyksessä. Mahdollisesti sosiologian puolelta voisi löytyä vastauksia täydentämään tähän työhön jääviä aukkoja pudokkuuden ymmärtämisessä.

Aineistoa käsitellessä ja koeasetelmia laatiessa tehtiin useita pieniä päätöksiä, jotka yhdessä vaikuttivat lopputuloksiin merkittävästi. Olisi mielenkiintoista tutkia vielä laajemmin, kuinka erilaisia tuloksia voisi saavuttaa erityyppisillä koeasetelmilla. Aineisto on mahdollista käsitellä usealla eri tavalla ennen mallien kouluttamista. Tässä työssä käytetty tapa oli vain yksi monista mahdollisuuksista. Ehkä opintoaineistosta saisi johdettua vielä joitain sellaisia muuttujia, jotka voisivat tuoda lisäarvoa mallin kouluttamiseen. Esimerkiksi olisi voinut luoda lisää koostettuja tietokenttiä ”ensimmäisen lukukauden keskiarvo” tai ”jaksojen määrä, jolla nolla suoritusta” tai jokin muu. Voi myös olla, että jokin muuttujayhdistelmä olisi parempi ja tarkempi määritelmä pudokkaalle kuin luvattoman poissaolon pituus, jota tässä työssä tarkasteltiin. Jaksojen pituutta voisi myös säätää monella eri tavalla.

Olisi aiheellista myös tutkia, parantaisiko aikasarjamuotoinen opetusaineisto ennustustarkuutta. Tällöin aineisto annettaisiin niin sanotusti raakana, eli yhtä opiskelijaa kohden olisi useita rivejä, joista jokainen edustaisi yhtä kurssisuoritusta. Tätä varten pitäisi löytää sopiva malli, joka pystyisi ymmärtämään aineistoa tässä muodossa. Esimerkiksi jonkin sekventiaalisen neuroverkkomallin sovellus voisi sopia tehtävään [3]. Tämä mahdollistaisi kurssikohtaisten erojen esiin tuomisen. Tässäkin aineistossa näkyi jonkin verran eroja kurssikohtaisissa suoriutumisissa pudokkaiden ja ei-pudokkaiden välillä (kts. aliluku 3.4). Eroja näkyi kuitenkin vain joidenkin kurssien kohdalla: useimmilla kursseilla eroa ei ollut lainkaan tai hyvin vähän.

Aineiston käsittelytapojen lisäksi valittavalla optimoinnin kohdemittarilla on tässä työssä valtava merkitys mallin suoriutumiseen, koska hyperparametrit vaikuttavat malliin merkittävästi, ja mittari vaikuttaa hyperparametrien valintaan merkittävästi. Olisi mielenkiintoista tutkia tarkemmin, oliko F-arvo oikea valinta maksimoitavaksi kohdemittariksi hyperparametrejä optimoidessa ja oliko sille annettu β-arvo sopiva? Optimoinnin kohteeksi voisi harkita jotain toista mittaria, joka tasapainottaa halutulla tavalla herkkyyttä, tarkkuutta ja täsmällisyyttä – tai jotain toista arvoa.

Jatkotutkimusaiheeksi jää myös se, minkälaisia toimenpiteitä näihin mallin ennustamiin pudokkaisiin pitäisi kohdentaa. On selvää, että joukossa tulee olemaan myös ”vääriä hälytyksiä”, koska malli ei ole täydellinen. Väärien hälytysten määrä vaikuttanee muun muassa siihen, kuinka suuria kustannuksia yksi toimenpide saa aiheuttaa. Jos tulevaisuudessa mallista saataisiin koulutettua vielä parempi, voisi yhden opiskelijan tukitoimen sallia tulla kalliimmaksi.

Lähteet

[1] G. Abu-Oda ja A. El-Halees. "Data Mining in Higher Education : University Student Dropout Case Study". *International Journal of Data Mining & Knowledge Management Process* 5 (31. tammikuuta 2015), s. 15–27. DOI: 10.5121/ijdkp.2015.5102.

[2] T. Chen ja C. Guestrin. "XGBoost: A Scalable Tree Boosting System". Teoksessa: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM, 13. elokuuta 2016, s. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: https://dl.acm.org/doi/10.1145/2939672.2939785 (viitattu 02. 11. 2024).

[3] Y. Cheng, B. Pereira Nunes ja R. Manrique. "Not Another Hardcoded Solution to the Student Dropout Prediction Problem: A Novel Approach Using Genetic Algorithms for Feature Selection". Teoksessa: *Intelligent Tutoring Systems*. Toim. S. Crossley ja E. Popescu. Vol. 13284. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022, s. 238–251. ISBN: 978-3-031-09679-2 978-3-031-09680-8. DOI: 10.1007/978-3-031-09680-8_23. URL: https://link.springer.com/10.1007/978-3-031-09680-8_23 (viitattu 03. 09. 2024).

[4] A. G. Costa, E. Queiroga, T. T. Primo, J. C. B. Mattos ja C. Cechinel. "Prediction analysis of student dropout in a Computer Science course using Educational Data Mining". Teoksessa: *2020 XV Conferencia Latinoamericana de Tecnologias de Aprendizaje (LACLO)*. 2020 XV Conferencia Latinoamericana de Tecnologias de Aprendizaje (LACLO). Lokakuu 2020, s. 1–6. DOI: 10.1109/LACLO50806.2020.9381166. URL: https://ieeexplore.ieee.org/document/9381166 (viitattu 27. 09. 2024).

[5] *DecisionTreeClassifier*. scikit-learn. URL: https://scikit-learn/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html (viitattu 06. 01. 2025).

[6] M. P. Deisenroth, A. A. Faisal ja C. S. Ong. *Mathematics for Machine Learning*. Cambridge University Press, 2020. URL: https://mml-book.com.

[7] *dmlc/xgboost*. original-date: 2014-02-06T17:28:03Z. 18. tammikuuta 2025. URL: https://github.com/dmlc/xgboost (viitattu 18. 01. 2025).

[8] T. Hastie, R. Tibshirani ja J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[9] H. Huo, J. Cui, S. Hein, Z. Padgett, M. Ossolinski, R. Raim ja J. Zhang. "Predicting Dropout for Nontraditional Undergraduate Students: A Machine Learning Approach". *Journal of college student retention : Research, theory & practice* 24.4 (2023). Place: Los Angeles, CA Publisher: SAGE Publications, s. 1054–1077. ISSN: 1521-0251.

[10] G. James, D. Witten, T. Hastie ja R. Tibshirani. *An Introduction to Statistical Learning*. 2nd. New York, NY: Springer, 2021. ISBN: 978-1-07-161420-4.

[11] L. Kemper, G. Vorhoff ja B. U. Wigger. "Predicting student dropout: A machine learning approach". *European Journal of Higher Education* 10.1 (2. tammikuuta 2020). Publisher: SRHE Website _eprint: https://doi.org/10.1080/21568235.2020.1718520, s. 28–47. ISSN: 2156-8235. DOI: 10.1080/21568235.2020.1718520. URL: https://doi.org/10.1080/21568235.2020.1718520 (viitattu 10. 10. 2024).

[12] M. Kevin P. *Machine Learning : A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series. The MIT Press, 2012. ISBN: 978-0-262-01802-9.

[13] B. Kiss, M. Nagy, R. Molontay ja B. Csabay. "Predicting Dropout Using High School and First-semester Academic Achievement Measures". Teoksessa: *2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA)*. 2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA). Marraskuu 2019, s. 383–389. DOI: 10.1109/ICETA48886.2019.9040158. URL: https://ieeexplore.ieee.org/document/9040158 (viitattu 02. 11. 2024).

[14] C. Lacave, A. I. Molina ja J. A. Cruz-Lemus. "Learning Analytics to identify dropout factors of Computer Science studies through Bayesian networks". *Behaviour & Information Technology* 37.10 (2. marraskuuta 2018). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0144929X.2018.1485053, s. 993–1007. ISSN: 0144-929X. DOI: 10.1080/0144929X.2018.1485053. URL: https://doi.org/10.1080/0144929X.2018.1485053 (viitattu 04. 10. 2024).

[15] S. Mangalathu, S.-H. Hwang ja J.-S. Jeon. ”Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach”. *Engineering Structures* 219 (15. syyskuuta 2020), s. 110927. ISSN: 0141-0296. DOI: 10.1016/j.engstruct.2020.110927. URL: https://www.sciencedirect.com/science/article/pii/S0141029620307513 (viitattu 15. 01. 2025).

[16] M. McHugh. ”Interrater reliability: The kappa statistic”. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB* 22 (lokakuu 2012), s. 276–82. DOI: 10.11613/BM.2012.031.

[17] M. A. Miranda ja J. Guzmán. ”Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos”. *Formación universitaria* 10.3 (2017), s. 61–68. ISSN: 0718-5006. DOI: 10.4067/S0718-50062017000300007. (Viitattu 27. 09. 2024).

[18] D. E. Moreira da Silva, E. J. Solteiro Pires, A. Reis, P. B. de Moura Oliveira ja J. Barroso. ”Forecasting Students Dropout: A UTAD University Study”. *Future internet* 14.3 (2022). Place: Basel Publisher: MDPI AG, s. 76–. ISSN: 1999-5903.

[19] S. Nagrecha, J. Z. Dillon ja N. V. Chawla. ”MOOC Dropout Prediction: Lessons Learned from Making Pipelines Interpretable”. Teoksessa: *Proceedings of the 26th International Conference on World Wide Web Companion - WWW ’17 Companion*. the 26th International Conference. Perth, Australia: ACM Press, 2017, s. 351–359. ISBN: 978-1-4503-4914-7. DOI: 10.1145/3041021.3054162. URL: http://dl.acm.org/citation.cfm?doid=3041021.3054162 (viitattu 21. 09. 2024).

[20] H. Peng, F. Long ja C. Ding. ”Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.8 (elokuu 2005). Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, s. 1226–1238. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2005.159. URL: https://ieeexplore.ieee.org/document/1453511/?arnumber=1453511 (viitattu 26. 10. 2024).

[21] B. Perez, C. Castellanos ja D. Correal. ”Applying Data Mining Techniques to Predict Student Dropout: A Case Study”. Teoksessa: *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)*. 2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI). Toukokuu 2018, s. 1–6. DOI: 10.1109/ColCACI.2018.8484847. URL: https://ieeexplore.ieee.org/document/8484847/?arnumber=8484847 (viitattu 27. 09. 2024).

[22] P. Probst, A.-L. Boulesteix ja B. Bischl. ”Tunability: Importance of Hyperparameters of Machine Learning Algorithms”. *Journal of Machine Learning Research* 20.53 (2019), s. 1–32. URL: http://jmlr.org/papers/v20/18-444.html.

[23] S. M. Ross. ”Chapter 13 - Chi-Squared Goodness-of-Fit Tests”. Teoksessa: *Introductory Statistics (Fourth Edition)*. Toim. S. M. Ross. Oxford: Academic Press, 1. tammikuuta 2017, s. 585–620. ISBN: 978-0-12-804317-2. DOI: 10.1016/B978-0-12-804317-2.00013-8. URL: https://www.sciencedirect.com/science/article/pii/B9780128043172000138 (viitattu 19. 10. 2024).

[24] *shap/shap: A game theoretic approach to explain the output of any machine learning model.* URL: https://github.com/shap/shap (viitattu 15. 01. 2025).

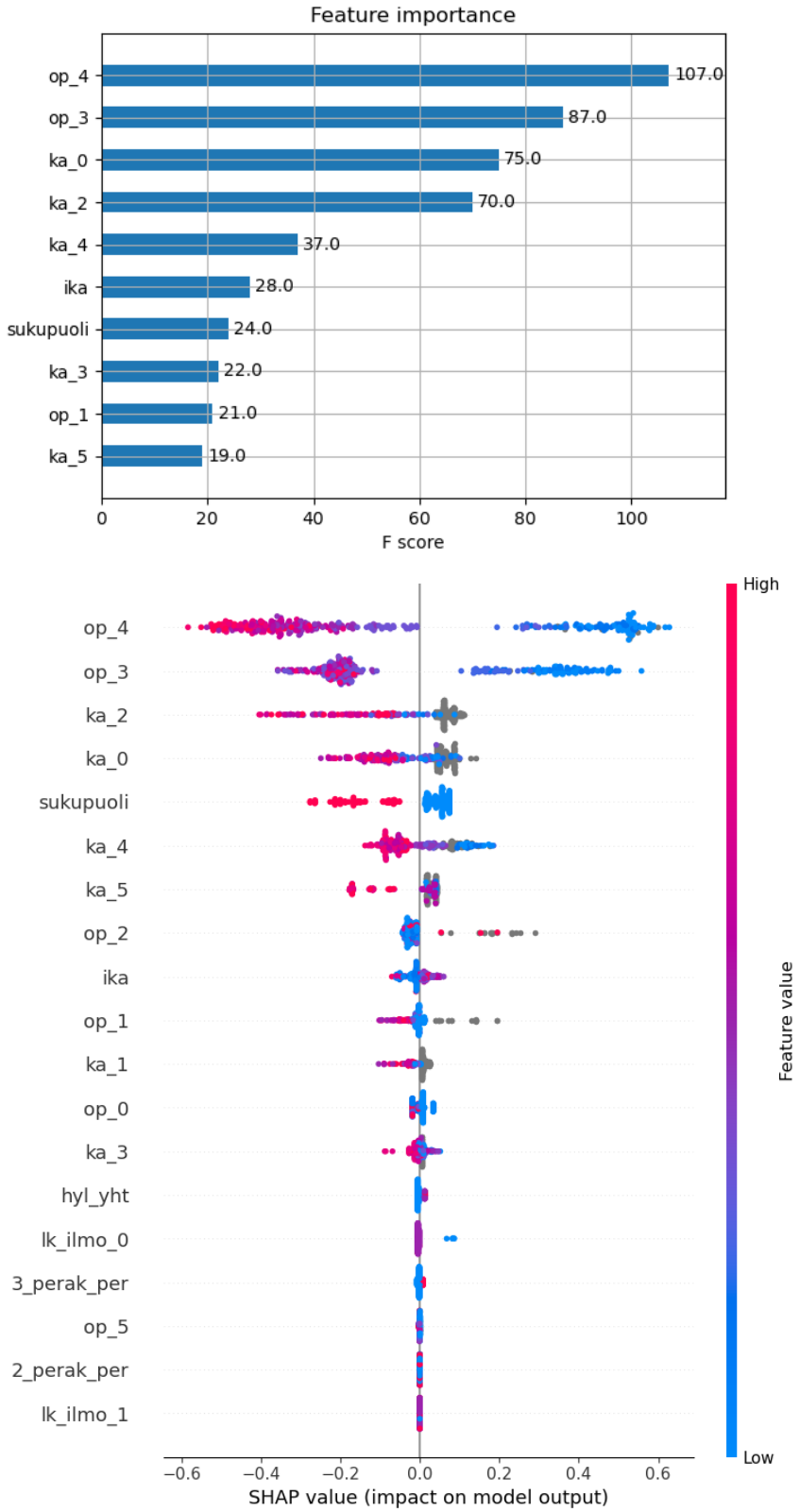
[25] M. Sokolova, N. Japkowicz ja S. Szpakowicz. ”Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation”. Teoksessa: AI 2006: Advances in Artificial Intelligence, Lecture Notes in Computer Science. Vol. Vol. 4304. 1. tammikuuta 2006, s. 1015–1021. ISBN: 978-3-540-49787-5. DOI: 10.1007/11941439_114.

[26] M. Solis, T. Moreira, R. Gonzalez, T. Fernandez ja M. Hernandez. ”Perspectives to Predict Dropout in University Students with Machine Learning”. Teoksessa: *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*. 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI). Heinäkuu 2018, s. 1–6. DOI: 10.1109/IWOBI.2018.8464191. URL: https://ieeexplore.ieee.org/document/8464191 (viitattu 21. 09. 2024).

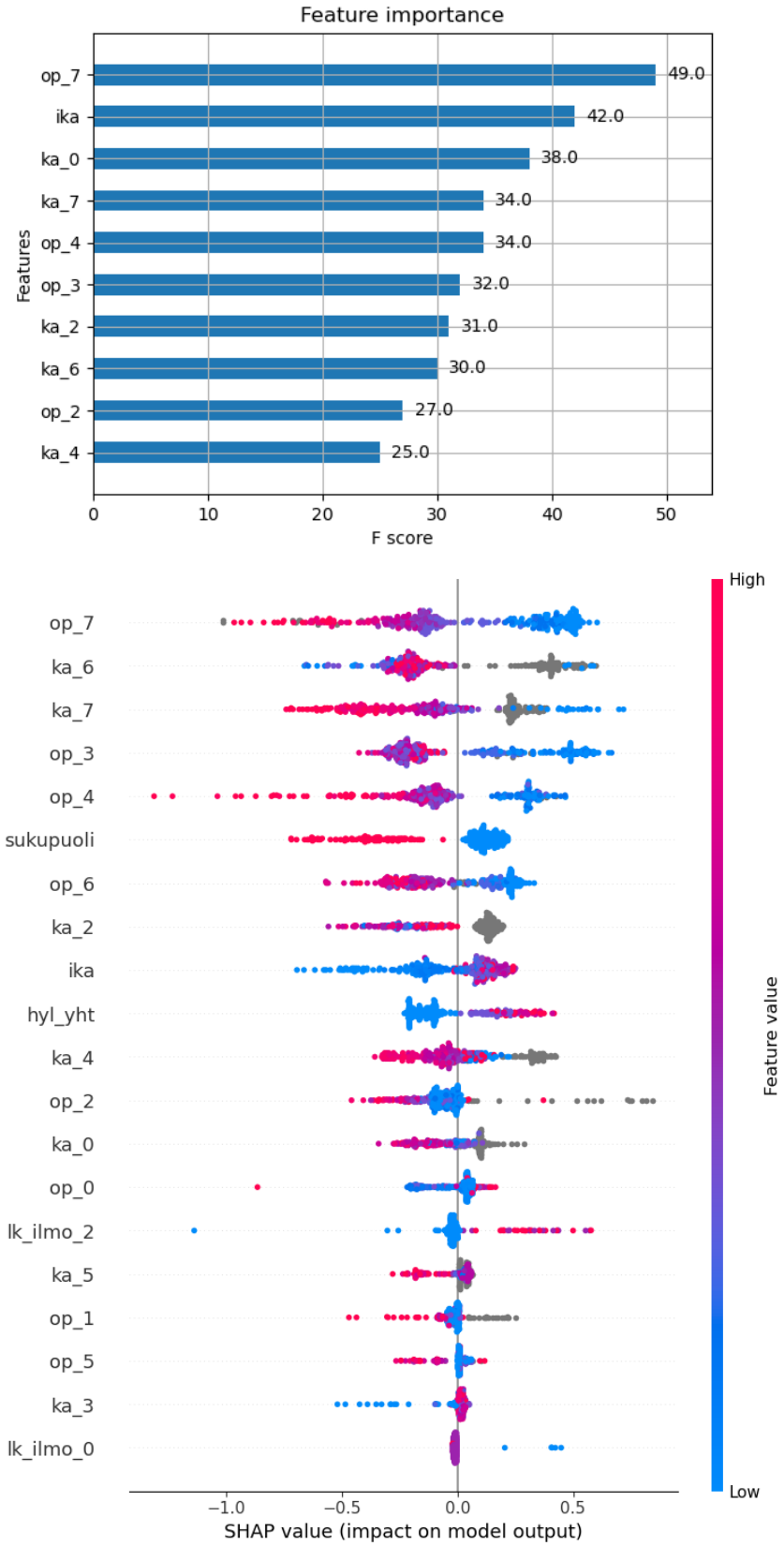
[27] M. Tan ja P. Shao. ”Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method”. *International Journal of Emerging Technologies in Learning (iJET)* 10.1 (21. helmikuuta 2015), s. 11. ISSN: 1863-0383. DOI: 10.3991/ijet.v10i1.4189. URL: https://online-journals.org/index.php/i-jet/article/view/4189 (viitattu 03. 09. 2024).

[28] Y. Zhang, S. Oussena, T. Clark ja H. Kim. ”Use Data Mining to Improve Student Retention in Higher Education - A Case Study.” Teoksessa: *ICEIS - 12th International Conference on Enterprise Information Systems 2010, Portugal, 8-12 June 2010*. Pages: 197. 1. tammikuuta 2010.

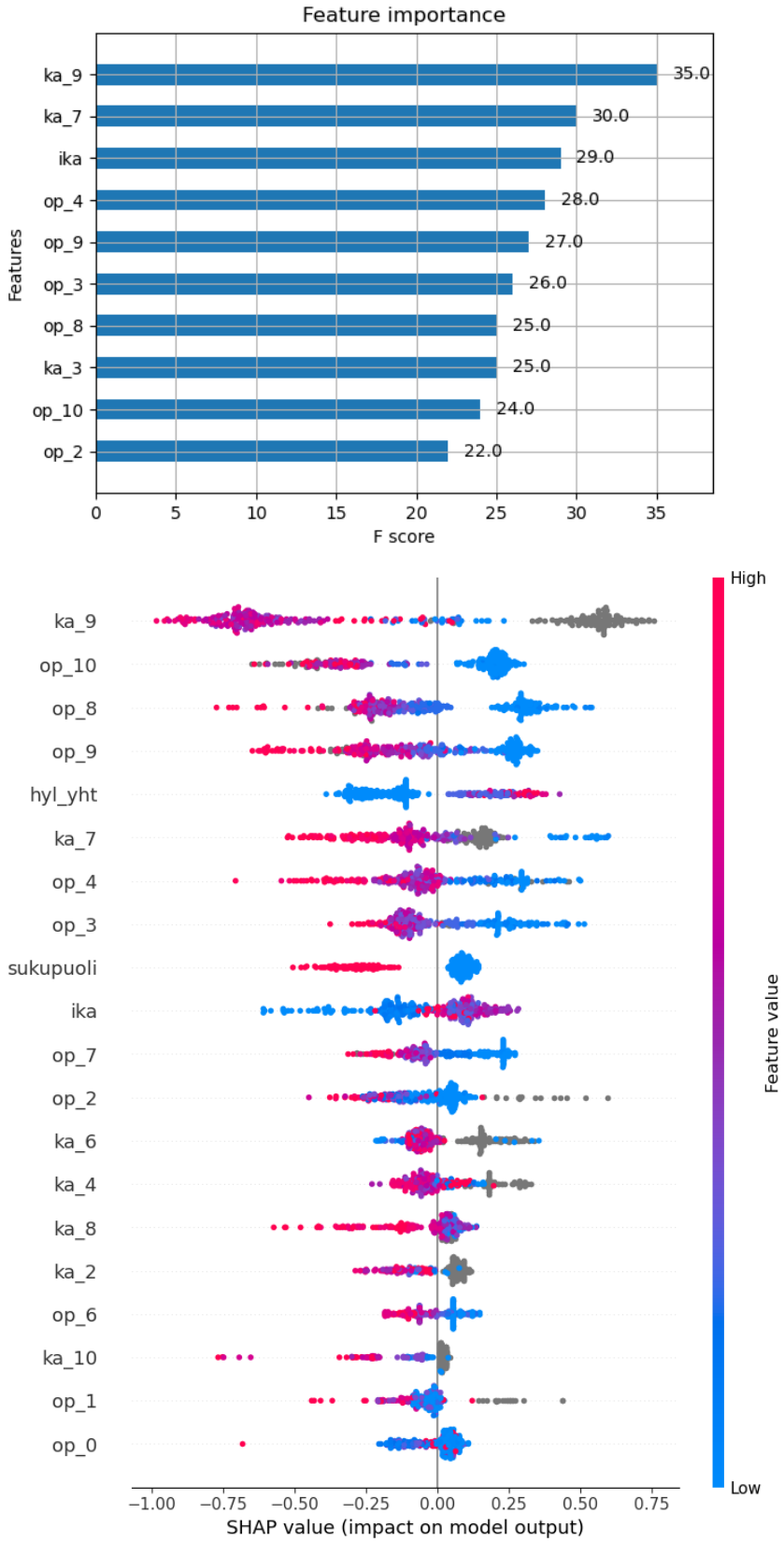
Liite A Tulokset XGBoost-kokeista



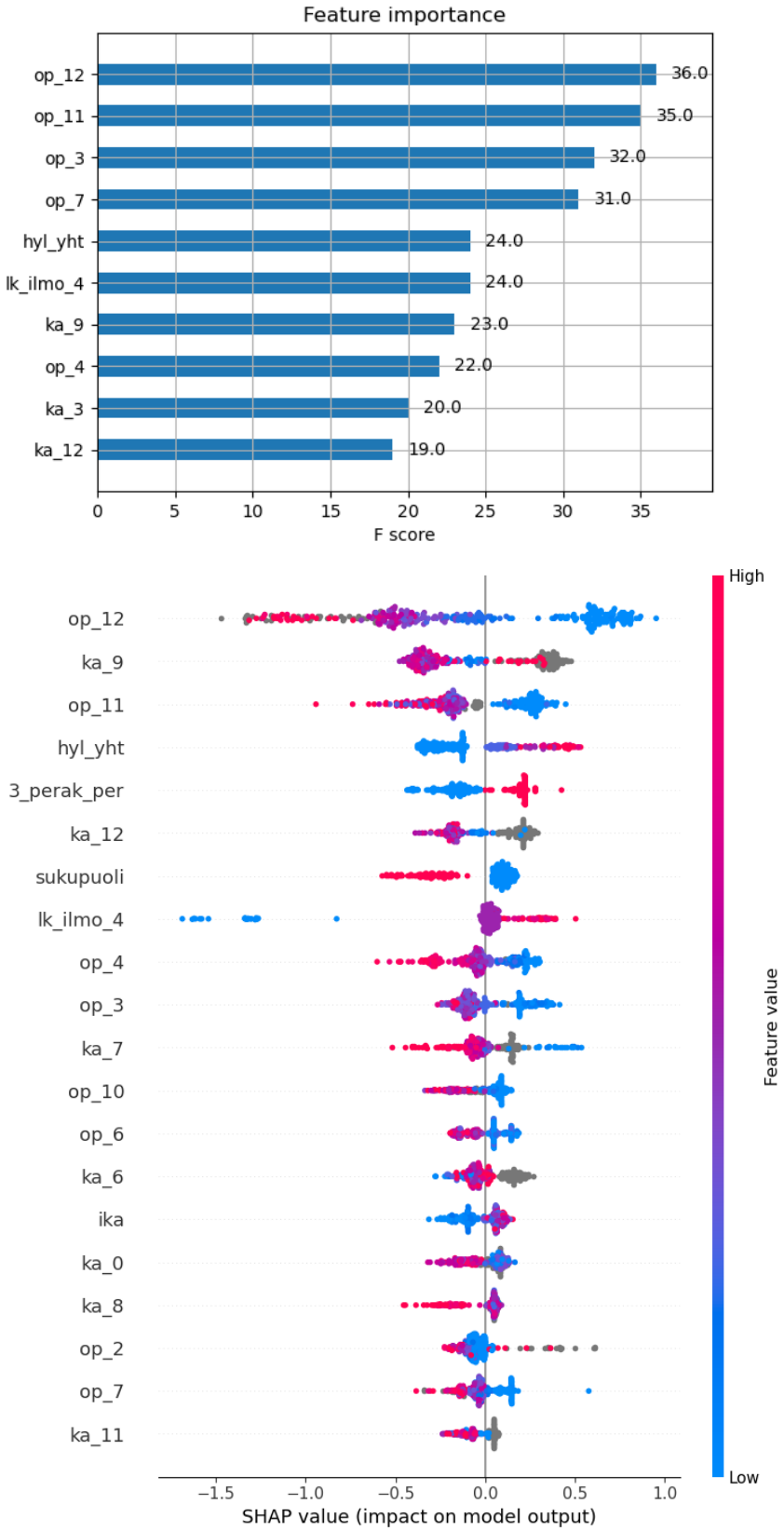
Kuva A.1: Aineisto: 17-22, ennuste 1 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.



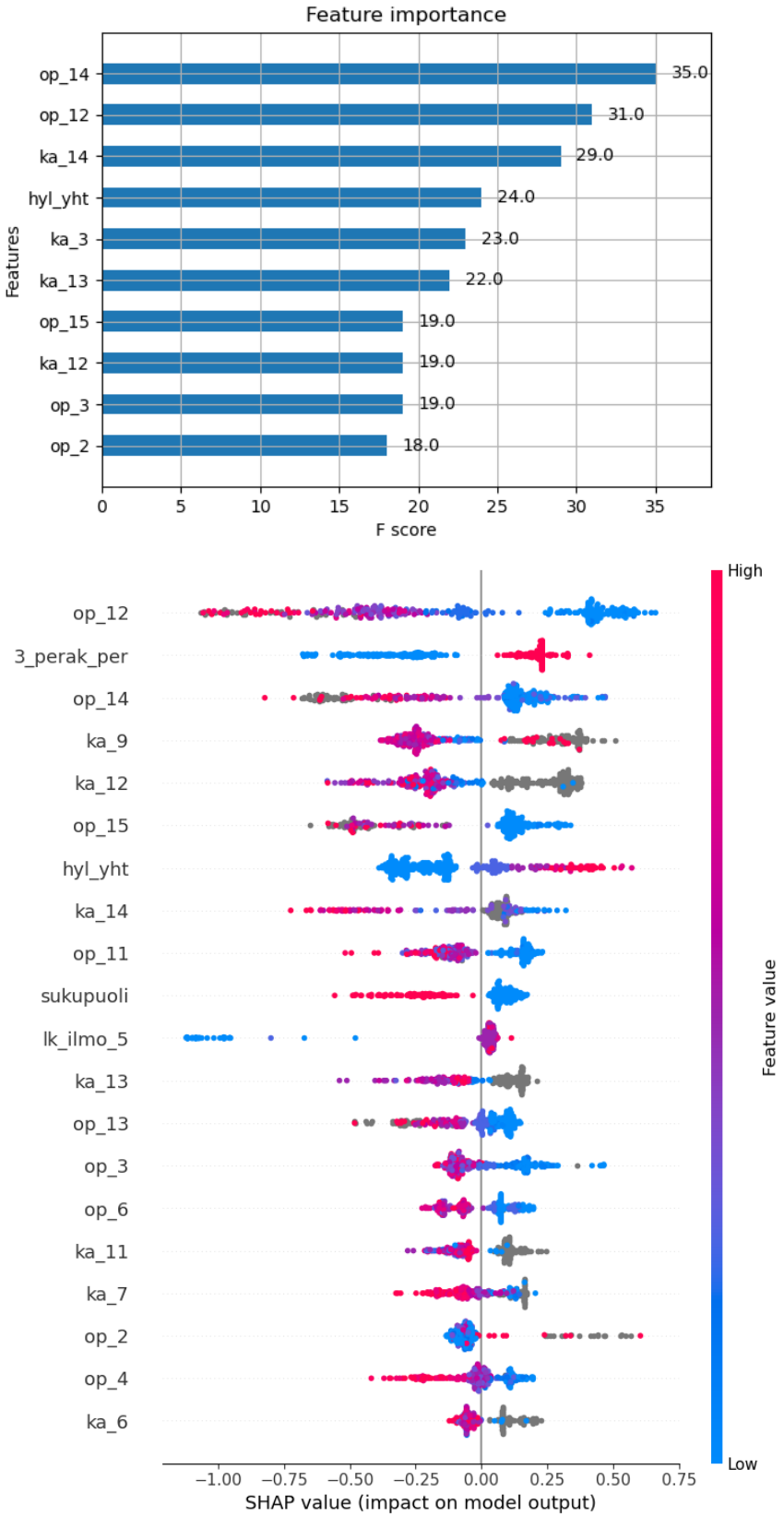
Kuva A.2: Aineisto: 17-22, ennuste 1,5 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.



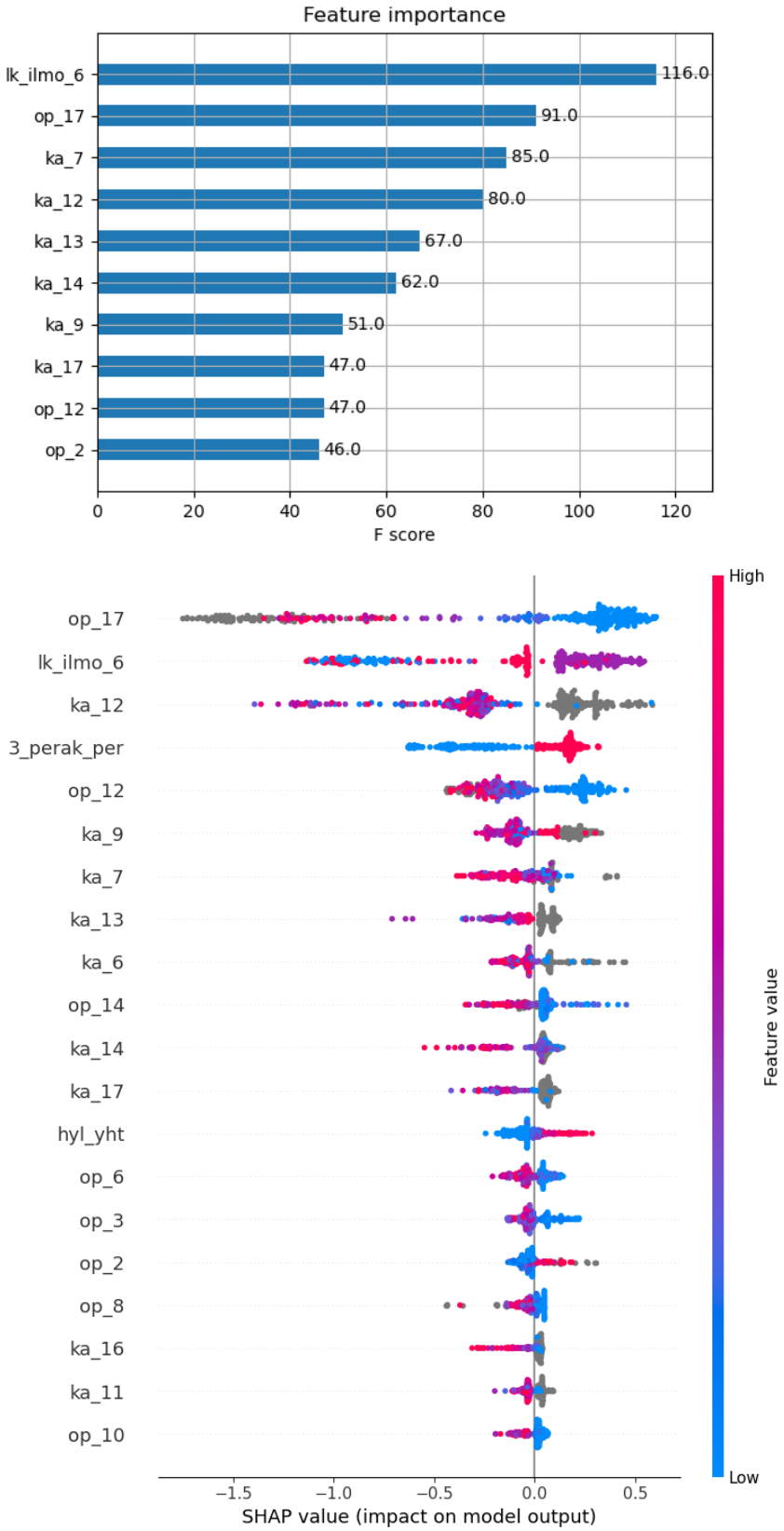
Kuva A.3: Aineisto: 17-22, ennuste 2 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.



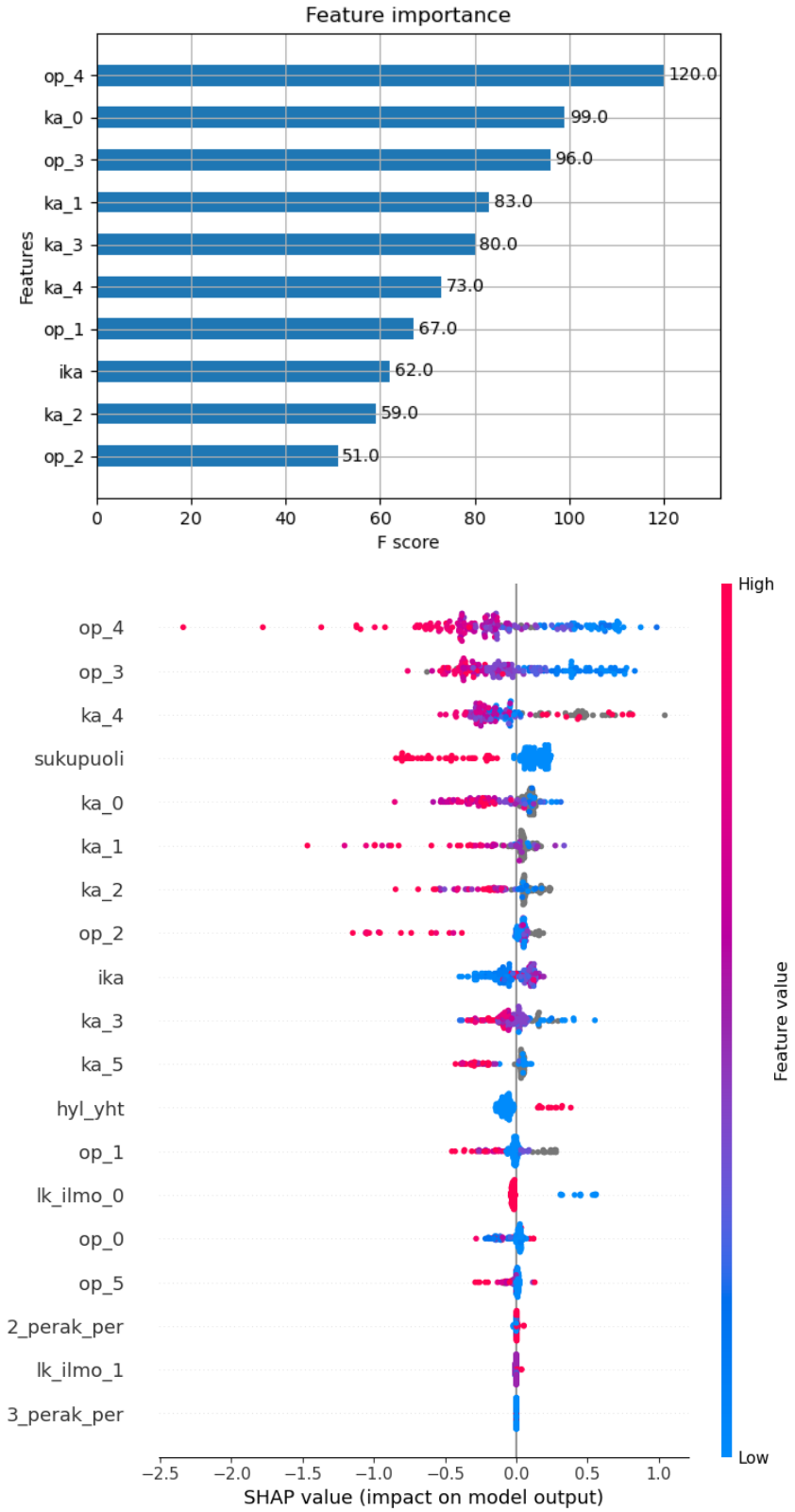
Kuva A.4: Aineisto: 17-22, ennuste 2,5 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.



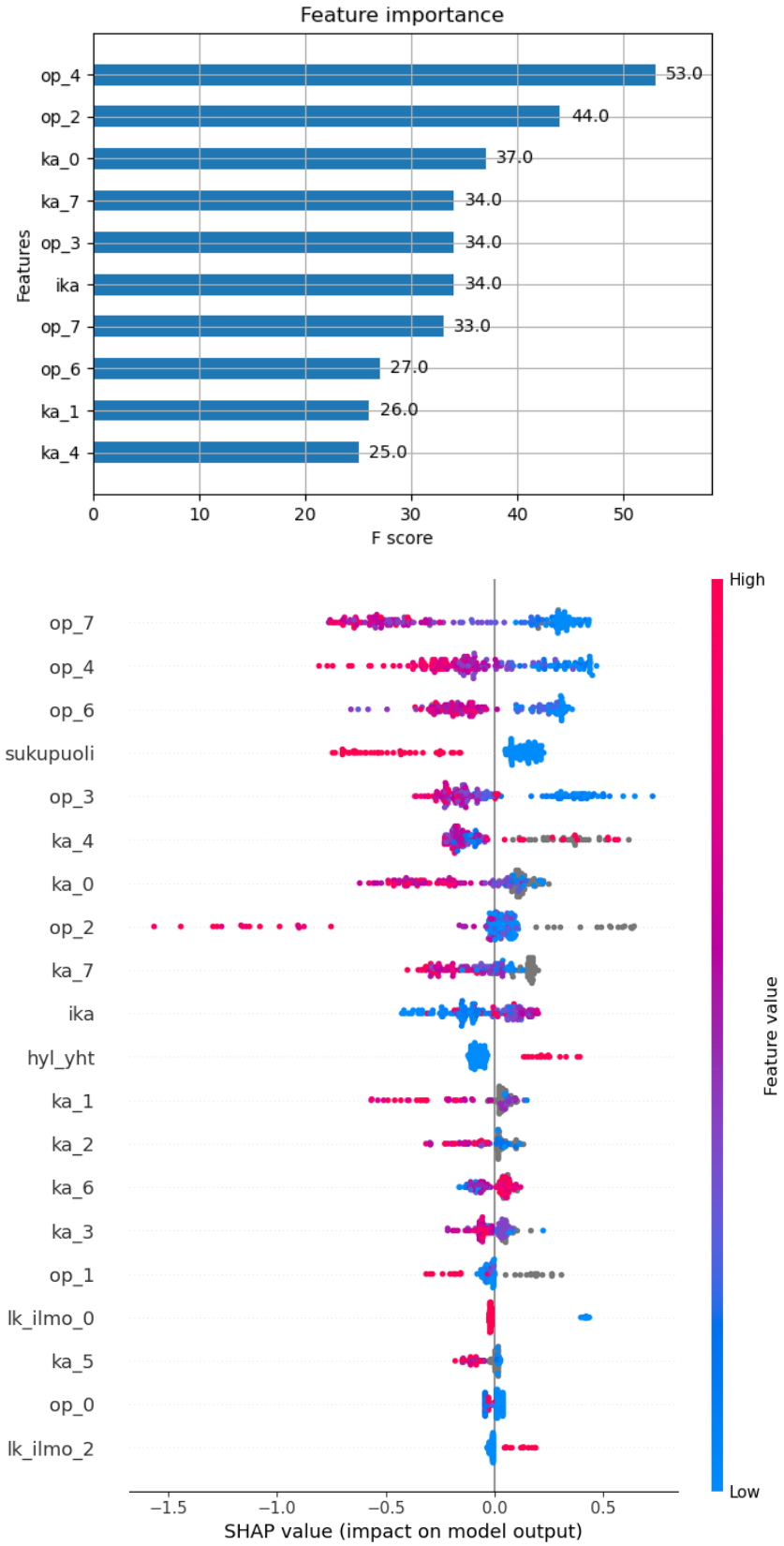
Kuva A.5: Aineisto: 17-22, ennuste 3 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.



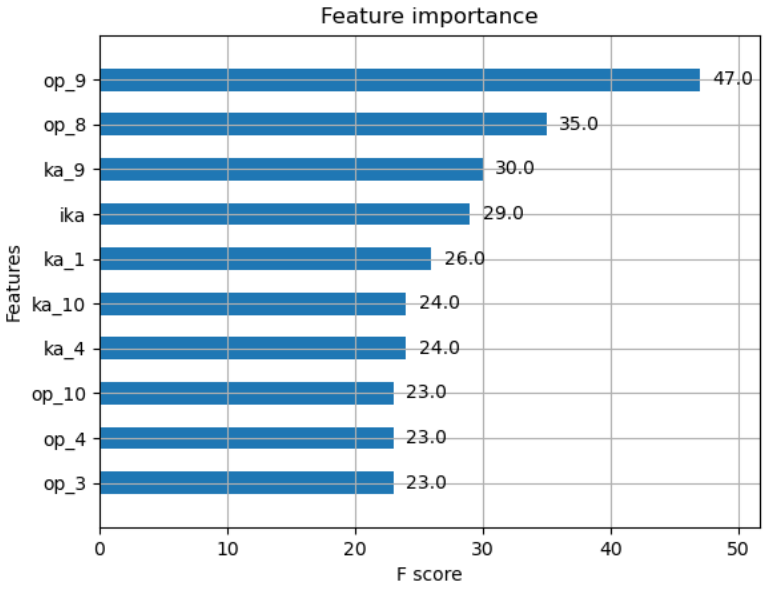
Kuva A.6: Aineisto: 17-22, ennuste 3,5 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.

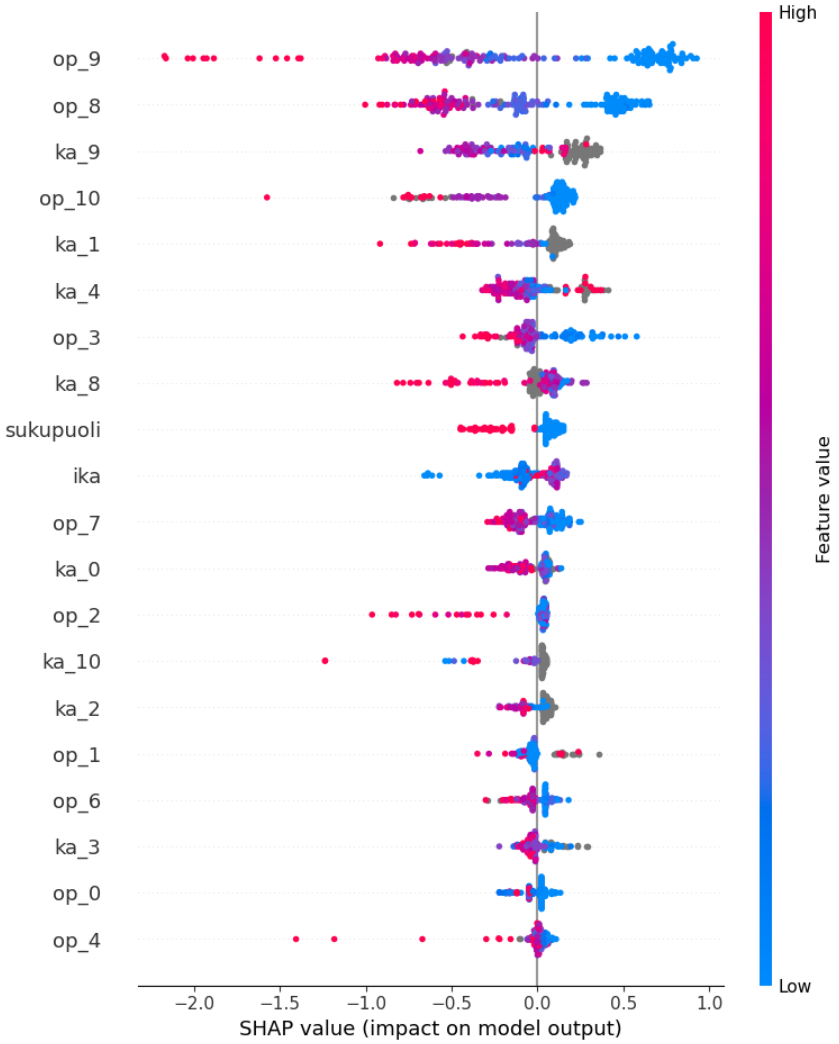


Kuva A.7: Aineisto: 20-22, ennuste 1 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.

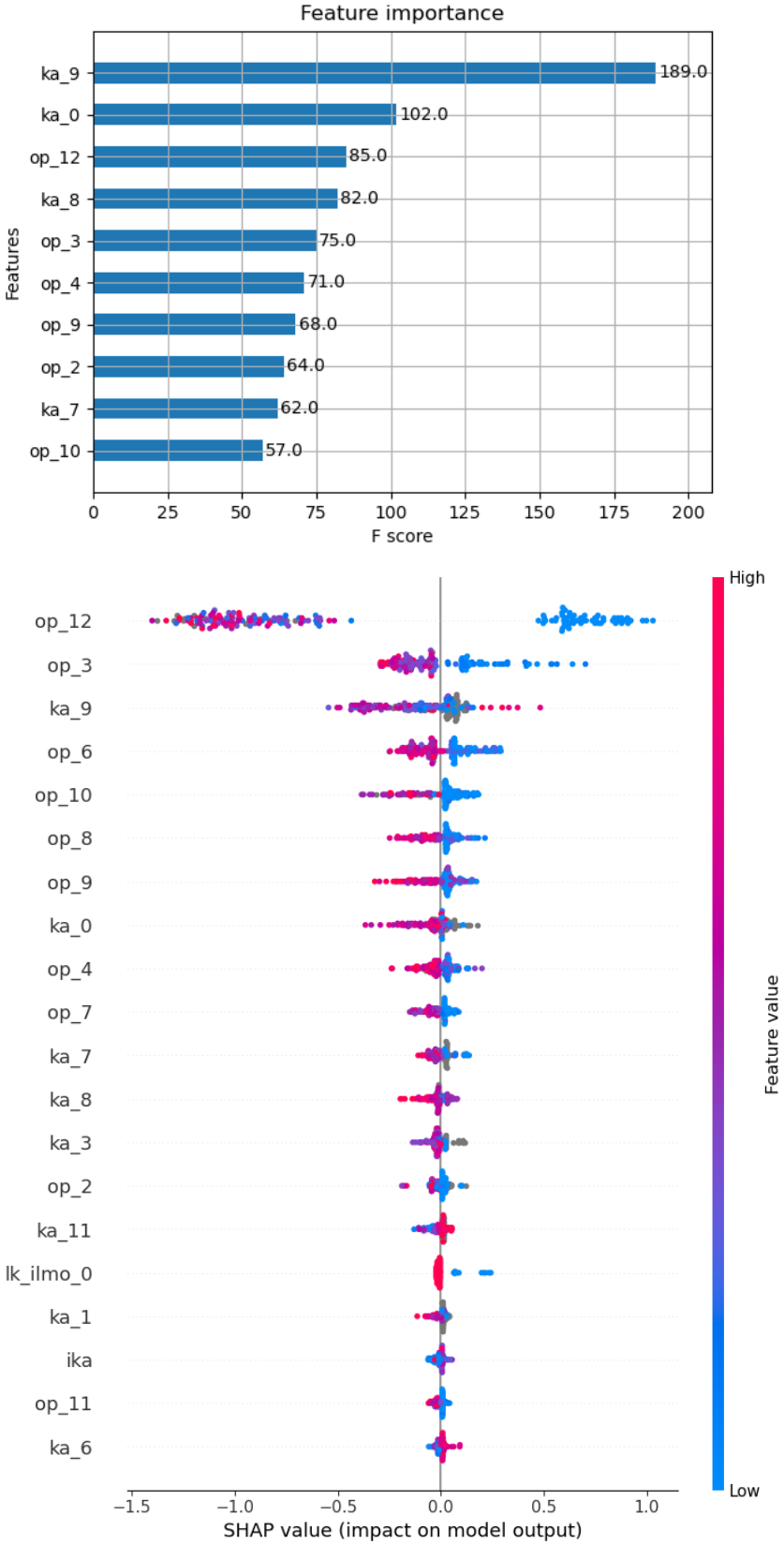


Kuva A.8: Aineisto: 20-22, ennuste 1,5 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.

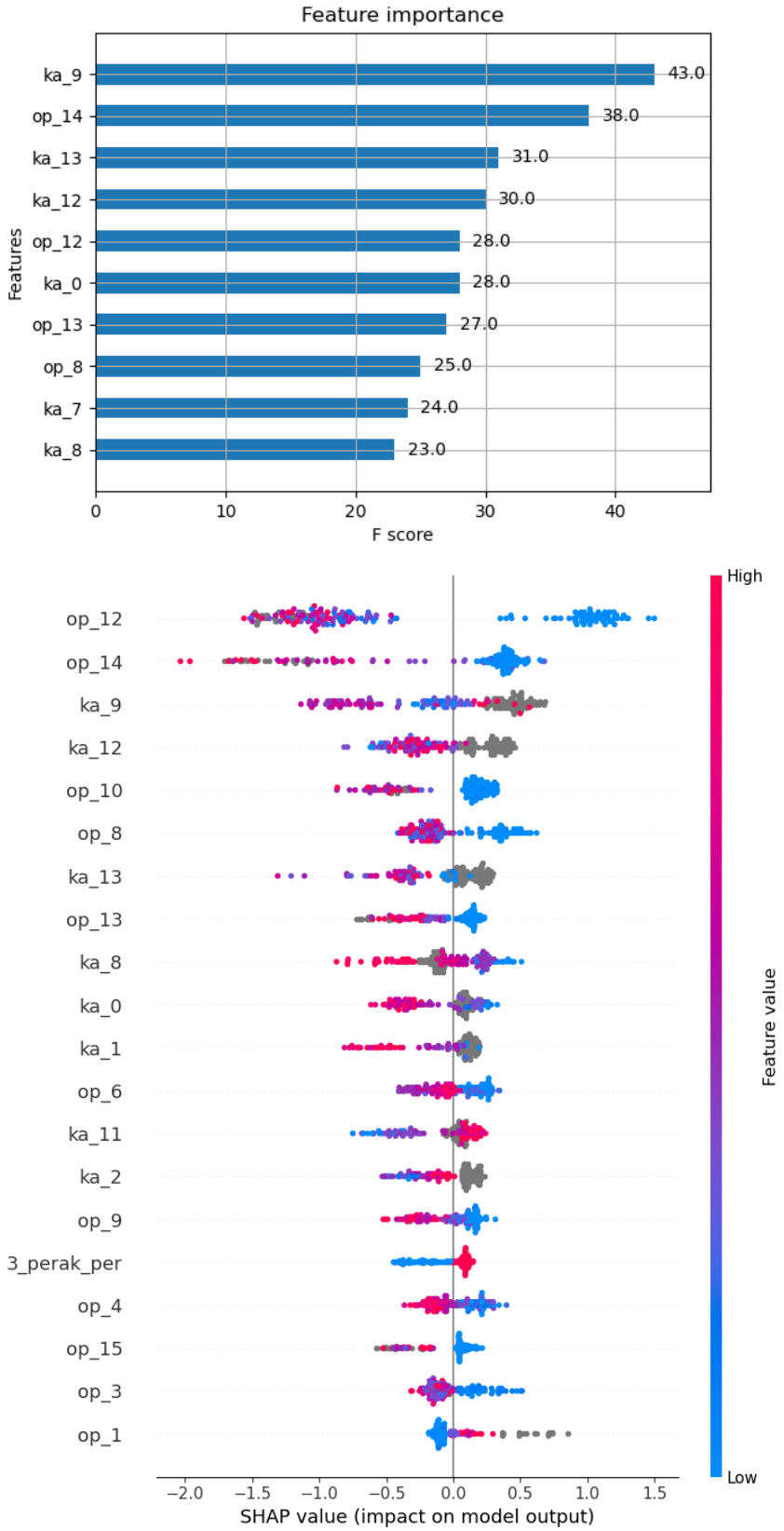




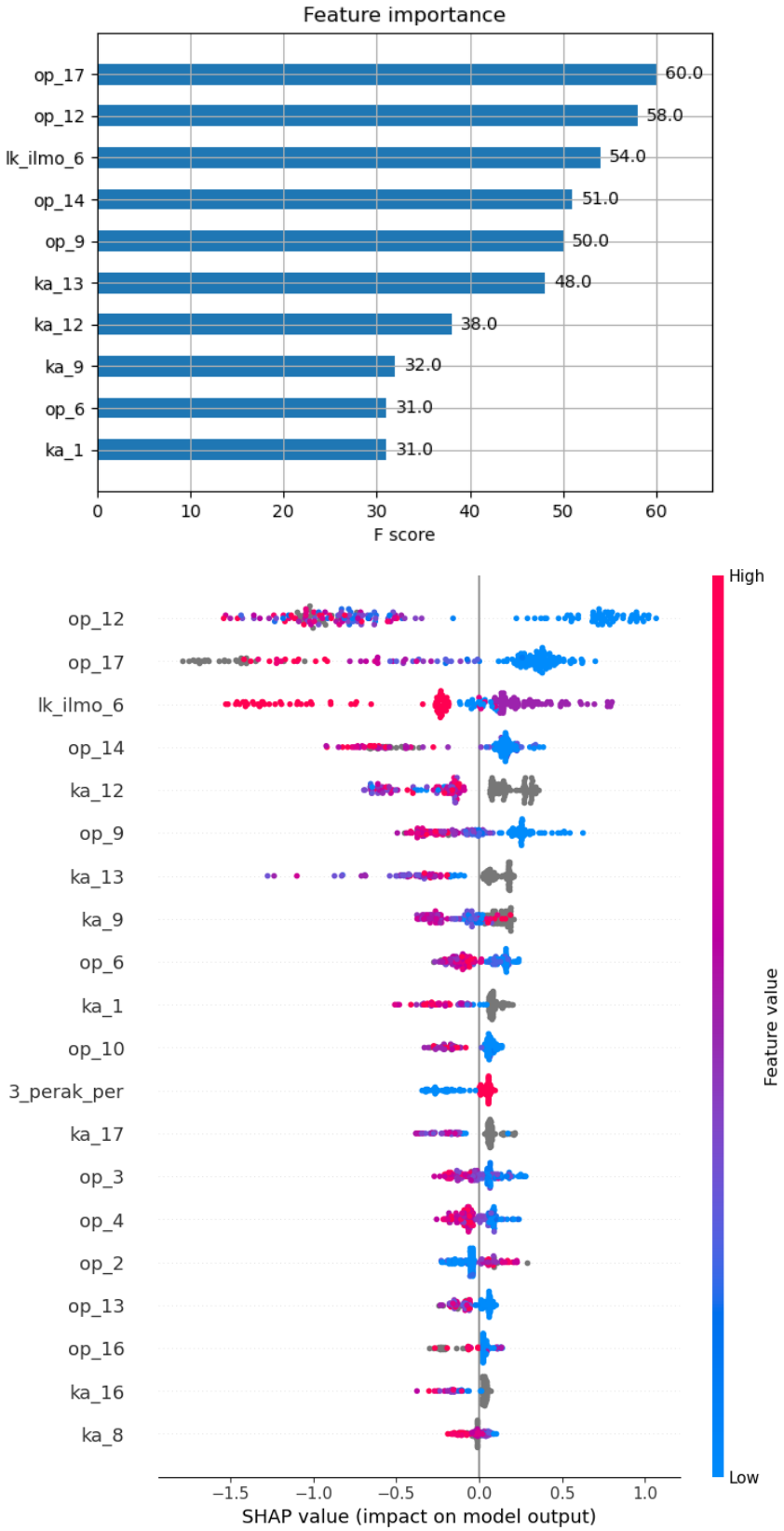
Kuva A.9: Aineisto: 20-22, ennuste 2 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.



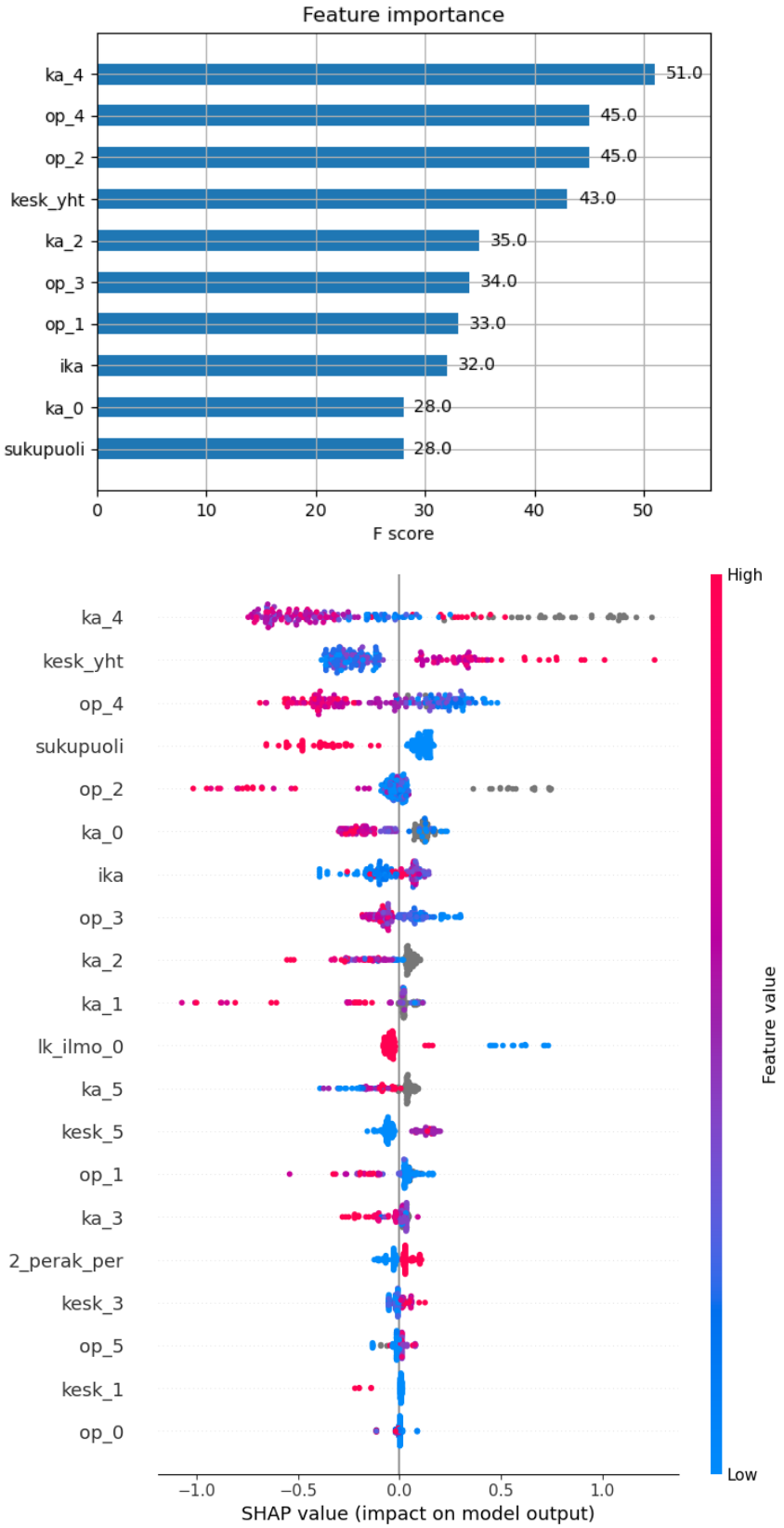
Kuva A.10: Aineisto: 20-22, ennuste 2,5 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.



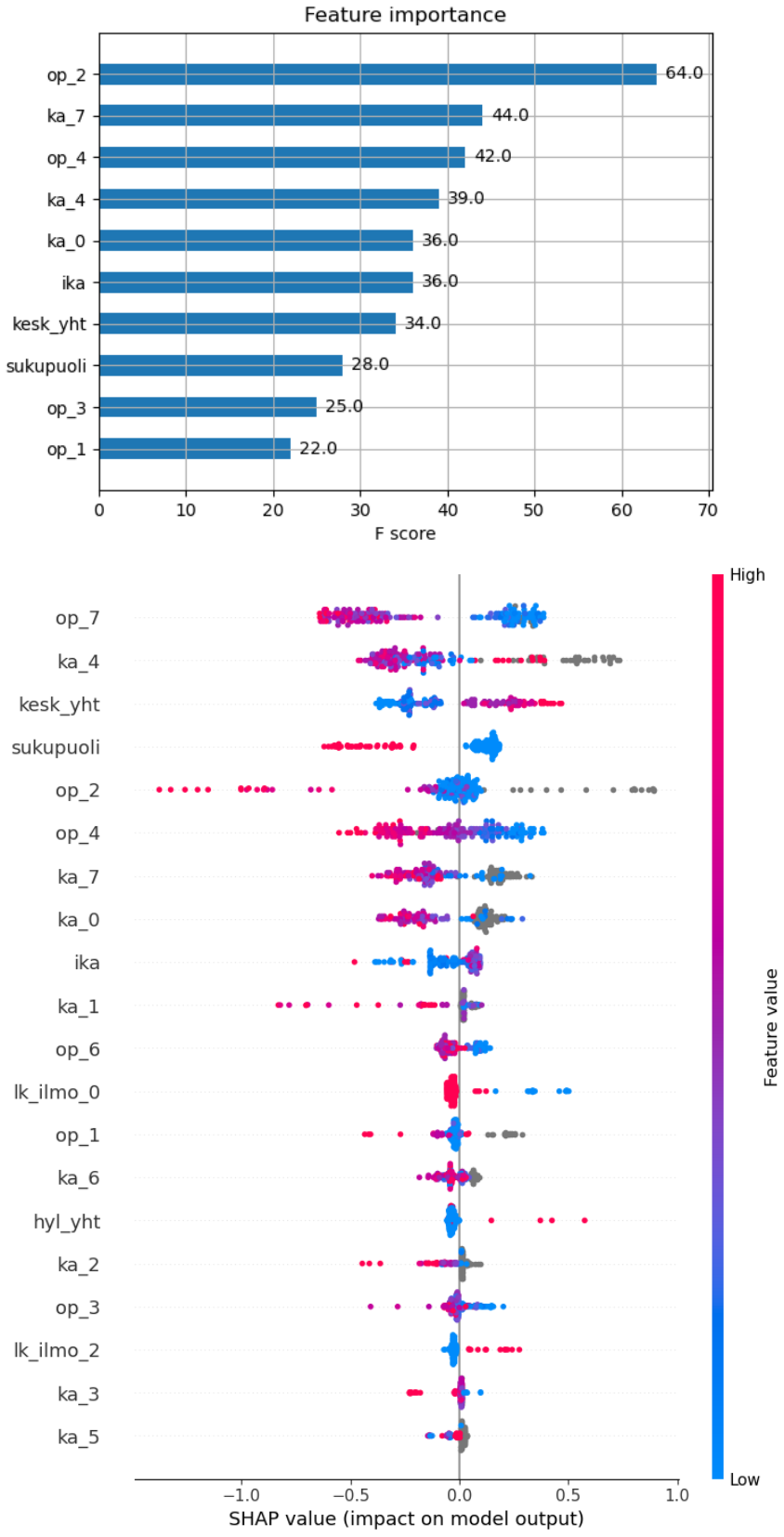
Kuva A.11: Aineisto: 20-22, ennuste 3 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.



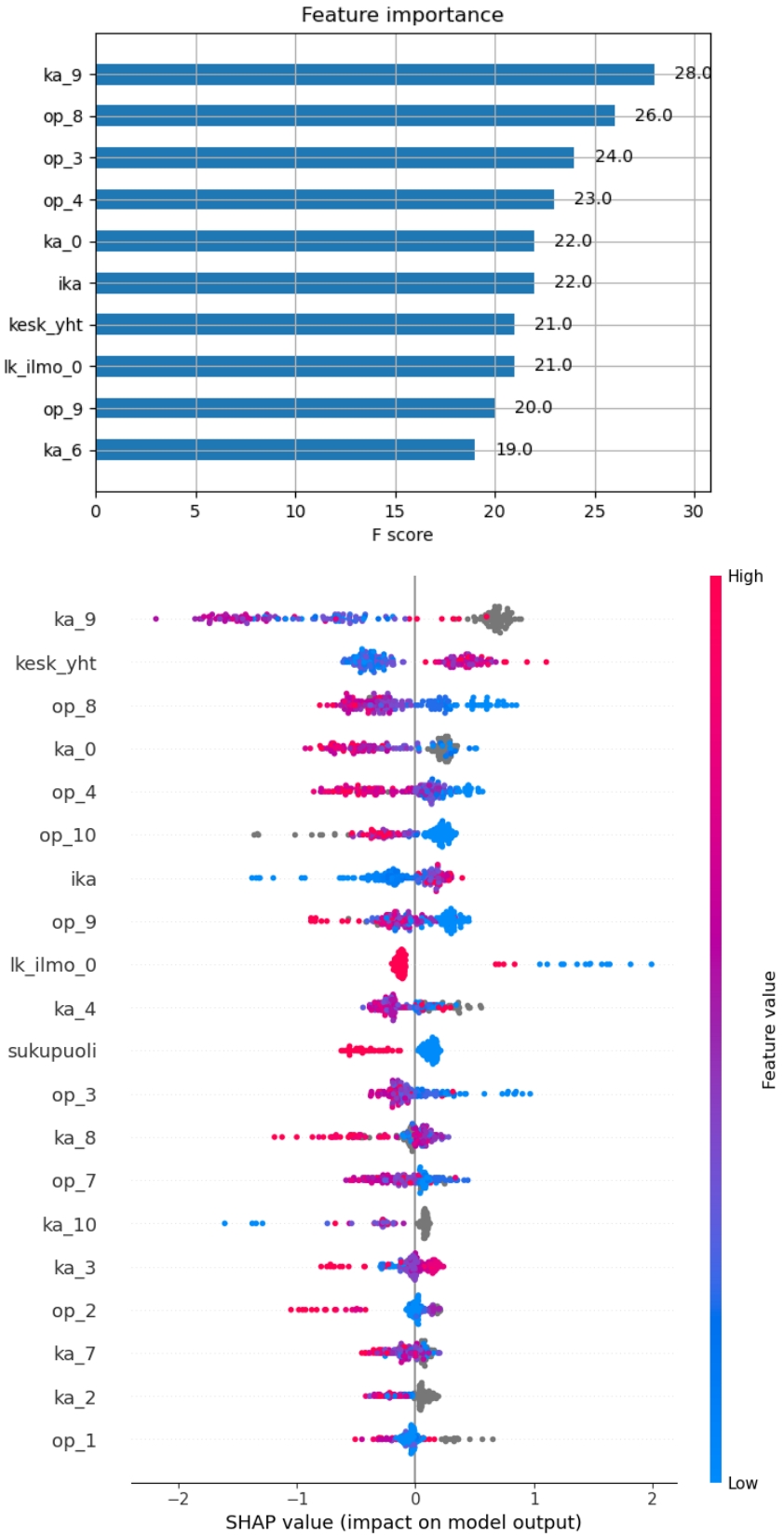
Kuva A.12: Aineisto: 20-22, ennuste 3,5 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.



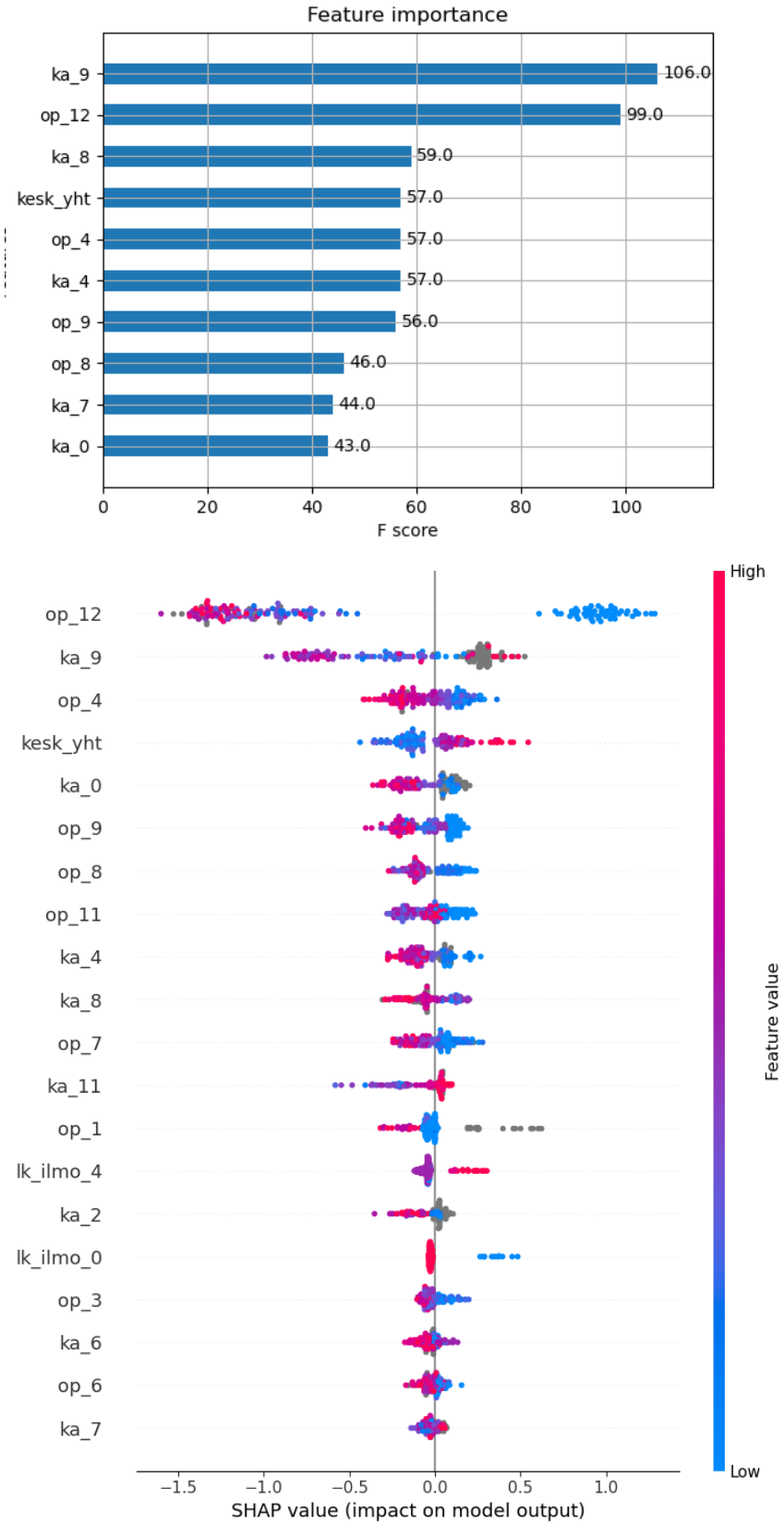
Kuva A.13: Aineisto: 20-22_ilm, ennuste 1 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.



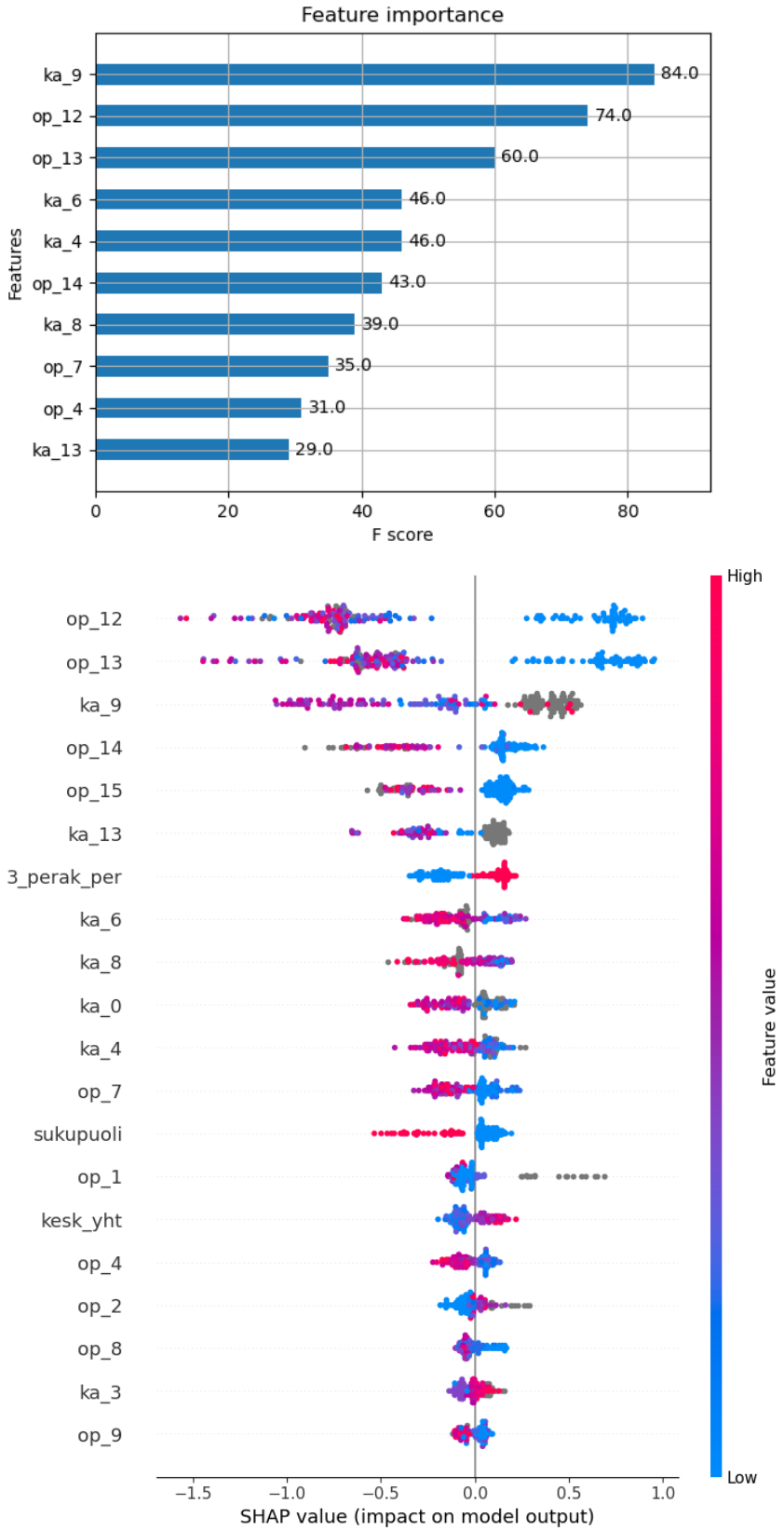
Kuva A.14: Aineisto: 20-22_ilm, ennuste 1,5 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.



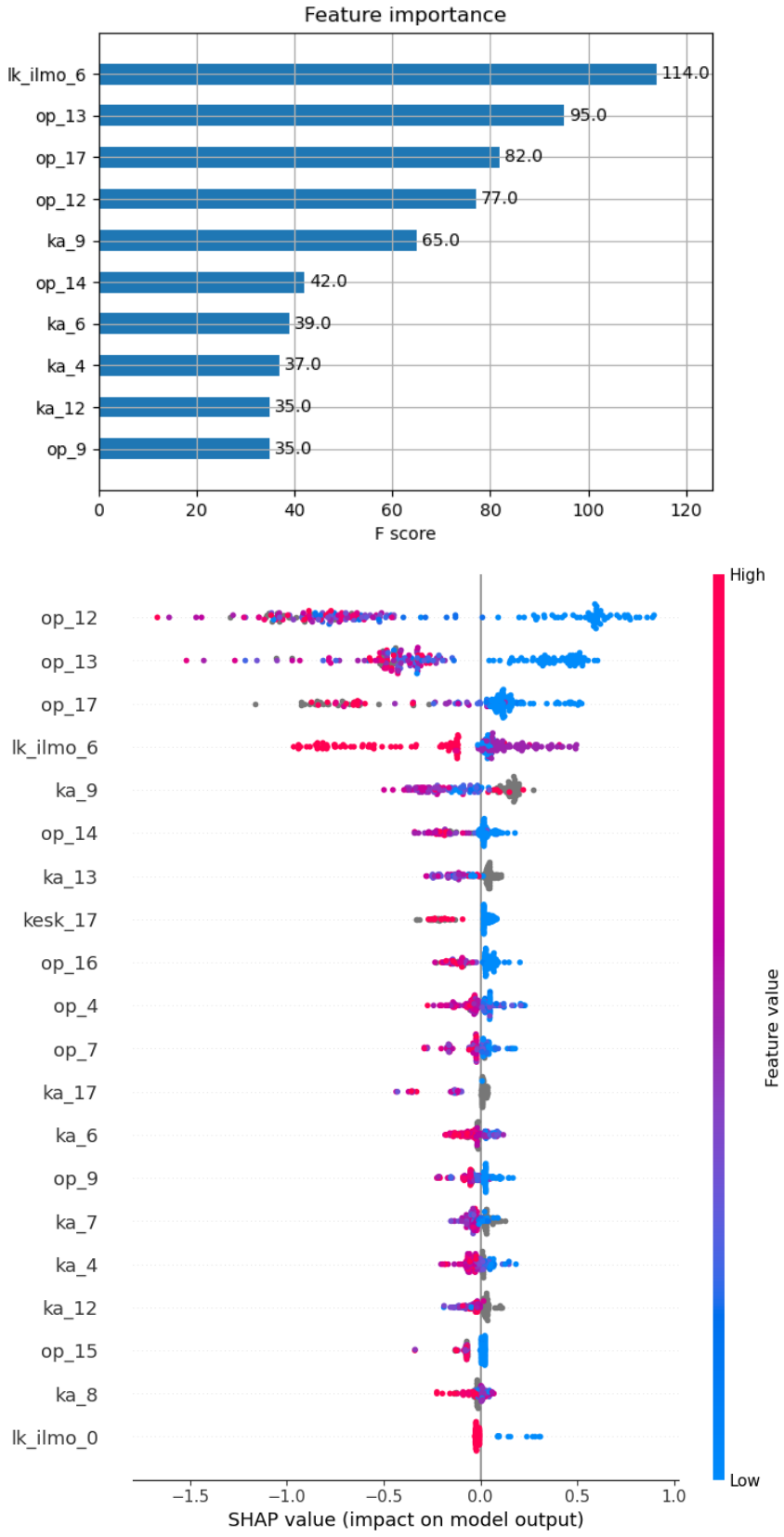
Kuva A.15: Aineisto: 20-22_ilm, ennuste 2 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.



Kuva A.16: Aineisto: 20-22_ilm, ennuste 2,5 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.



Kuva A.17: Aineisto: 20-22_ilm, ennuste 3 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.



Kuva A.18: Aineisto: 20-22_ilm, ennuste 3,5 vuoden opiskeluiden jälkeen. F score tarkoittaa ylemmässä kaaviossa muuttujan esiintymiskertoja puiden solmuissa. Alemman kaavion SHAP value tarkoittaa shap-arvoa, kts. 5.6.

ain. pituus (vuosia)	ristiinvalidointikierros	tarkkuus	herkkyys	f-arvo	täsmällisyys	kappa
1	1	0,565	0,878	0,729	0,508	0,176
1	2	0,626	0,950	0,787	0,548	0,292
1	3	0,574	0,935	0,760	0,514	0,201
1	4	0,594	0,892	0,747	0,528	0,228
1	5	0,573	0,978	0,780	0,513	0,203
1	keskiarvo	0,586	0,927	0,761	0,522	0,220
1,5	1	0,681	0,906	0,790	0,594	0,385
1,5	2	0,700	0,899	0,795	0,613	0,419
1,5	3	0,658	0,928	0,791	0,573	0,347
1,5	4	0,668	0,849	0,756	0,590	0,355
1,5	5	0,676	0,935	0,802	0,588	0,380
1,5	keskiarvo	0,677	0,904	0,787	0,592	0,377
2	1	0,742	0,885	0,807	0,658	0,495
2	2	0,742	0,899	0,814	0,654	0,496
2	3	0,719	0,906	0,807	0,630	0,455
2	4	0,681	0,835	0,754	0,604	0,377
2	5	0,699	0,935	0,812	0,607	0,420
2	keskiarvo	0,717	0,892	0,799	0,631	0,449
2,5	1	0,748	0,914	0,824	0,658	0,509
2,5	2	0,784	0,899	0,834	0,702	0,574
2,5	3	0,758	0,878	0,811	0,678	0,524
2,5	4	0,719	0,842	0,774	0,643	0,449
2,5	5	0,741	0,942	0,835	0,645	0,498
2,5	keskiarvo	0,750	0,895	0,815	0,665	0,511
3	1	0,784	0,914	0,841	0,698	0,575
3	2	0,790	0,935	0,854	0,699	0,589
3	3	0,794	0,964	0,869	0,694	0,597
3	4	0,742	0,820	0,773	0,675	0,489
3	5	0,767	0,928	0,840	0,675	0,545
3	keskiarvo	0,775	0,912	0,835	0,688	0,559
3,5	1	0,845	0,950	0,889	0,763	0,694
3,5	2	0,823	0,914	0,860	0,747	0,649
3,5	3	0,845	0,950	0,889	0,763	0,694
3,5	4	0,813	0,856	0,826	0,758	0,626
3,5	5	0,838	0,957	0,889	0,751	0,681
3,5	keskiarvo	0,833	0,925	0,870	0,756	0,669

Taulukko A.1: Aineisto: 2017-2022 aloittaneet.

ain. pituus (vuosia)	ristiinvalidointikierros	tarkkuus	herkkyys	f-arvo	täsmällisyys	kappa
1	1	0,584	0,792	0,680	0,500	0,210
1	2	0,630	0,836	0,724	0,540	0,294
1	3	0,659	0,822	0,729	0,566	0,341
1	4	0,686	0,875	0,767	0,583	0,397
1	5	0,674	0,875	0,762	0,573	0,377
1	keskiarvo	0,647	0,840	0,733	0,552	0,324
1,5	1	0,665	0,917	0,777	0,559	0,368
1,5	2	0,647	0,890	0,759	0,551	0,333
1,5	3	0,699	0,918	0,795	0,593	0,426
1,5	4	0,686	0,847	0,753	0,587	0,393
1,5	5	0,709	0,889	0,785	0,604	0,440
1,5	keskiarvo	0,681	0,892	0,774	0,579	0,392
2	1	0,757	0,847	0,786	0,663	0,520
2	2	0,740	0,890	0,801	0,637	0,494
2	3	0,734	0,822	0,763	0,645	0,474
2	4	0,767	0,903	0,820	0,663	0,545
2	5	0,762	0,806	0,767	0,682	0,522
2	keskiarvo	0,752	0,854	0,787	0,658	0,511
2,5	1	0,786	0,875	0,815	0,692	0,576
2,5	2	0,821	0,890	0,842	0,739	0,643
2,5	3	0,809	0,822	0,800	0,750	0,614
2,5	4	0,826	0,903	0,850	0,739	0,652
2,5	5	0,814	0,792	0,786	0,770	0,619
2,5	keskiarvo	0,811	0,856	0,818	0,738	0,621
3	1	0,803	0,833	0,802	0,732	0,603
3	2	0,832	0,890	0,848	0,756	0,664
3	3	0,832	0,849	0,827	0,775	0,661
3	4	0,837	0,847	0,828	0,782	0,669
3	5	0,855	0,819	0,823	0,831	0,701
3	keskiarvo	0,832	0,848	0,826	0,775	0,660
3,5	1	0,855	0,917	0,872	0,776	0,710
3,5	2	0,855	0,932	0,881	0,773	0,712
3,5	3	0,844	0,890	0,854	0,774	0,686
3,5	4	0,860	0,917	0,876	0,786	0,720
3,5	5	0,901	0,958	0,919	0,831	0,801
3,5	keskiarvo	0,863	0,923	0,880	0,788	0,726

Taulukko A.2: Aineisto: 2020-2022 aloittaneet.

ain. pituus (vuosia)	ristiinvalidointikierros	tarkkuus	herkkyys	f-arvo	täsmällisyys	kappa
1	1	0,688	0,898	0,749	0,525	0,407
1	2	0,665	0,864	0,720	0,505	0,363
1	3	0,636	0,900	0,727	0,486	0,330
1	4	0,645	0,881	0,720	0,491	0,339
1	5	0,692	0,898	0,752	0,530	0,414
1	keskiarvo	0,665	0,888	0,733	0,507	0,371
1,5	1	0,723	0,814	0,724	0,565	0,442
1,5	2	0,688	0,898	0,749	0,525	0,407
1,5	3	0,688	0,917	0,760	0,529	0,412
1,5	4	0,727	0,915	0,778	0,563	0,473
1,5	5	0,692	0,864	0,735	0,531	0,405
1,5	keskiarvo	0,703	0,882	0,749	0,542	0,428
2	1	0,786	0,729	0,712	0,672	0,534
2	2	0,757	0,864	0,769	0,600	0,512
2	3	0,815	0,883	0,815	0,679	0,619
2	4	0,797	0,729	0,719	0,694	0,554
2	5	0,738	0,746	0,696	0,595	0,453
2	keskiarvo	0,779	0,790	0,742	0,648	0,534
2,5	1	0,803	0,831	0,779	0,671	0,586
2,5	2	0,844	0,864	0,821	0,729	0,668
2,5	3	0,861	0,917	0,860	0,743	0,710
2,5	4	0,802	0,763	0,742	0,692	0,572
2,5	5	0,791	0,814	0,763	0,658	0,561
2,5	keskiarvo	0,820	0,838	0,793	0,699	0,619
3	1	0,850	0,831	0,807	0,754	0,674
3	2	0,838	0,814	0,791	0,738	0,649
3	3	0,873	0,900	0,860	0,771	0,730
3	4	0,837	0,746	0,753	0,772	0,636
3	5	0,797	0,814	0,766	0,667	0,571
3	keskiarvo	0,839	0,821	0,795	0,740	0,652
3,5	1	0,855	0,898	0,846	0,736	0,695
3,5	2	0,890	0,898	0,869	0,803	0,762
3,5	3	0,896	0,967	0,907	0,784	0,782
3,5	4	0,860	0,864	0,833	0,761	0,700
3,5	5	0,820	0,881	0,815	0,684	0,626
3,5	keskiarvo	0,864	0,902	0,854	0,754	0,713

Taulukko A.3: Aineisto: 2020-2022 aloittaneet. Ilmoittautumistiedot mukana.

Liite B Tulokset päätöspuu-kokeista

ain. pituus (vuosia)	ristiinvalidointikierros	tarkkuus	herkkyys	f-arvo	täsmällisyys	kappa
1	1	0,716	0,727	0,697	0,669	0,431
1	2	0,700	0,741	0,689	0,644	0,402
1	3	0,655	0,770	0,667	0,588	0,322
1	4	0,703	0,813	0,711	0,631	0,416
1	5	0,696	0,691	0,671	0,653	0,389
1	keskiarvo	0,694	0,748	0,687	0,637	0,392
1,5	1	0,713	0,748	0,700	0,658	0,427
1,5	2	0,700	0,755	0,693	0,640	0,404
1,5	3	0,723	0,784	0,717	0,661	0,449
1,5	4	0,729	0,734	0,708	0,685	0,456
1,5	5	0,683	0,734	0,675	0,626	0,369
1,5	keskiarvo	0,709	0,751	0,699	0,654	0,421
2	1	0,739	0,835	0,741	0,667	0,484
2	2	0,716	0,813	0,720	0,646	0,440
2	3	0,729	0,791	0,724	0,667	0,462
2	4	0,742	0,799	0,735	0,681	0,487
2	5	0,722	0,813	0,724	0,653	0,450
2	keskiarvo	0,729	0,810	0,729	0,663	0,464
2,5	1	0,713	0,878	0,733	0,629	0,440
2,5	2	0,758	0,842	0,757	0,688	0,521
2,5	3	0,768	0,856	0,768	0,696	0,540
2,5	4	0,758	0,878	0,765	0,678	0,524
2,5	5	0,728	0,827	0,732	0,657	0,463
2,5	keskiarvo	0,745	0,856	0,751	0,670	0,498
3	1	0,806	0,871	0,801	0,742	0,615
3	2	0,771	0,784	0,754	0,727	0,540
3	3	0,771	0,871	0,773	0,695	0,548
3	4	0,794	0,863	0,789	0,727	0,590
3	5	0,757	0,820	0,752	0,695	0,518
3	keskiarvo	0,780	0,842	0,774	0,717	0,562
3,5	1	0,861	0,942	0,859	0,789	0,725
3,5	2	0,829	0,863	0,819	0,779	0,658
3,5	3	0,832	0,878	0,824	0,777	0,665
3,5	4	0,826	0,921	0,826	0,749	0,655
3,5	5	0,838	0,863	0,828	0,795	0,676
3,5	keskiarvo	0,837	0,894	0,831	0,778	0,676

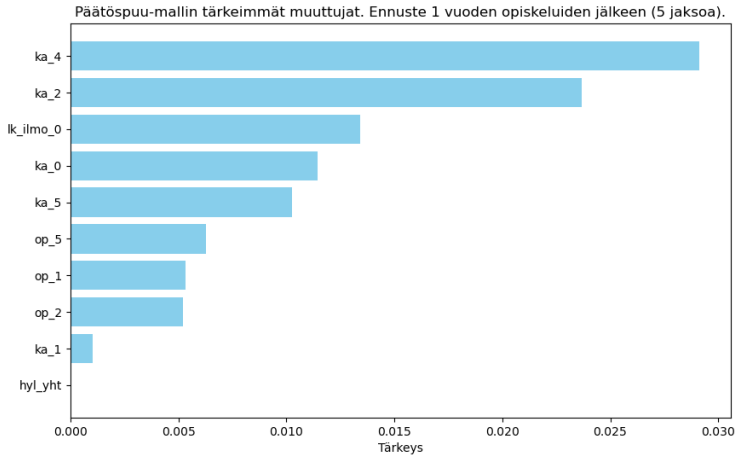
Taulukko B.1: Aineisto: 2017-2022 aloittaneet.

ain. pituus (vuosia)	ristiinvalidointikierros	tarkkuus	herkkyys	f-arvo	täsmällisyys	kappa
1	1	0,688	0,847	0,693	0,587	0,396
1	2	0,676	0,795	0,674	0,586	0,367
1	3	0,705	0,753	0,683	0,625	0,412
1	4	0,709	0,806	0,699	0,617	0,427
1	5	0,622	0,792	0,637	0,533	0,273
1	keskiarvo	0,680	0,798	0,677	0,589	0,375
1,5	1	0,786	0,861	0,770	0,697	0,574
1,5	2	0,636	0,699	0,618	0,554	0,279
1,5	3	0,717	0,685	0,671	0,658	0,423
1,5	4	0,657	0,792	0,659	0,564	0,333
1,5	5	0,733	0,764	0,705	0,655	0,463
1,5	keskiarvo	0,706	0,760	0,685	0,626	0,414
2	1	0,803	0,708	0,750	0,797	0,589
2	2	0,763	0,685	0,709	0,735	0,510
2	3	0,884	0,767	0,848	0,949	0,757
2	4	0,802	0,847	0,782	0,726	0,603
2	5	0,808	0,792	0,776	0,760	0,608
2	keskiarvo	0,812	0,760	0,773	0,794	0,613
2,5	1	0,855	0,861	0,832	0,805	0,706
2,5	2	0,792	0,836	0,772	0,718	0,583
2,5	3	0,896	0,863	0,875	0,887	0,786
2,5	4	0,802	0,833	0,779	0,732	0,602
2,5	5	0,831	0,778	0,794	0,812	0,652
2,5	keskiarvo	0,835	0,834	0,811	0,791	0,665
3	1	0,873	0,847	0,847	0,847	0,738
3	2	0,838	0,740	0,794	0,857	0,662
3	3	0,890	0,836	0,865	0,897	0,773
3	4	0,872	0,847	0,847	0,847	0,737
3	5	0,831	0,792	0,797	0,803	0,653
3	keskiarvo	0,861	0,812	0,830	0,850	0,713
3,5	1	0,873	0,889	0,853	0,821	0,741
3,5	2	0,896	0,836	0,871	0,910	0,784
3,5	3	0,873	0,890	0,855	0,823	0,742
3,5	4	0,826	0,889	0,810	0,744	0,651
3,5	5	0,866	0,861	0,844	0,827	0,727
3,5	keskiarvo	0,867	0,873	0,847	0,825	0,729

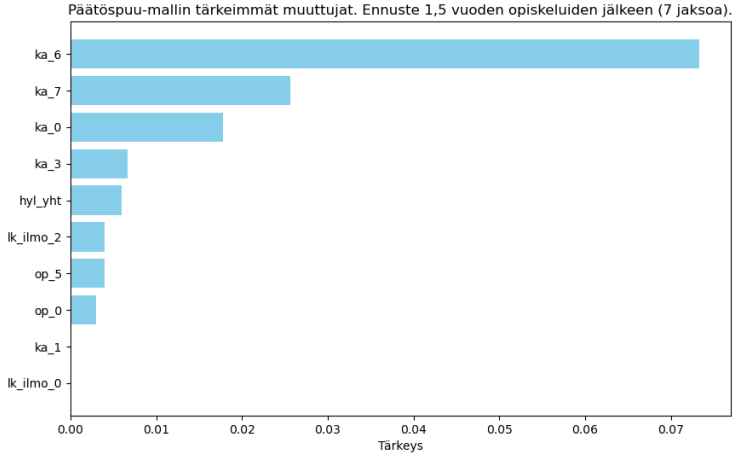
Taulukko B.2: Aineisto: 2020-2022 aloittaneet.

ain. pituus (vuosia)	ristiinvalidointikierros	tarkkuus	herkkyys	f-arvo	täsmällisyys	kappa
1	1	0,717	0,763	0,647	0,563	0,420
1	2	0,694	0,746	0,624	0,537	0,377
1	3	0,671	0,767	0,617	0,517	0,347
1	4	0,703	0,746	0,633	0,550	0,394
1	5	0,756	0,780	0,687	0,613	0,491
1	keskiarvo	0,708	0,760	0,642	0,556	0,406
1,5	1	0,682	0,797	0,631	0,522	0,372
1,5	2	0,630	0,831	0,605	0,476	0,302
1,5	3	0,723	0,750	0,652	0,577	0,428
1,5	4	0,698	0,729	0,623	0,544	0,379
1,5	5	0,756	0,763	0,682	0,616	0,487
1,5	keskiarvo	0,698	0,774	0,639	0,547	0,394
2	1	0,792	0,814	0,727	0,658	0,562
2	2	0,798	0,746	0,715	0,688	0,559
2	3	0,838	0,783	0,770	0,758	0,646
2	4	0,802	0,729	0,717	0,705	0,565
2	5	0,802	0,678	0,702	0,727	0,554
2	keskiarvo	0,806	0,750	0,726	0,707	0,577
2,5	1	0,809	0,729	0,723	0,717	0,577
2,5	2	0,850	0,729	0,768	0,811	0,657
2,5	3	0,780	0,767	0,708	0,657	0,533
2,5	4	0,820	0,729	0,735	0,741	0,598
2,5	5	0,895	0,864	0,850	0,836	0,770
2,5	keskiarvo	0,831	0,764	0,757	0,753	0,627
3	1	0,821	0,763	0,744	0,726	0,606
3	2	0,838	0,814	0,774	0,738	0,649
3	3	0,879	0,817	0,824	0,831	0,731
3	4	0,837	0,712	0,750	0,792	0,630
3	5	0,913	0,814	0,865	0,923	0,801
3	keskiarvo	0,858	0,784	0,791	0,802	0,683
3,5	1	0,821	0,763	0,744	0,726	0,606
3,5	2	0,890	0,881	0,846	0,813	0,761
3,5	3	0,867	0,867	0,819	0,776	0,714
3,5	4	0,884	0,797	0,825	0,855	0,738
3,5	5	0,907	0,864	0,864	0,864	0,794
3,5	keskiarvo	0,874	0,834	0,819	0,807	0,722

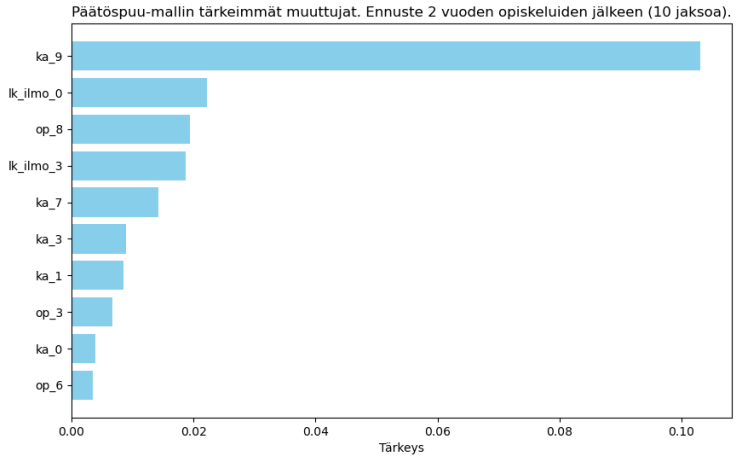
Taulukko B.3: Aineisto: 2020-2022 aloittaneet. Ilmoittautumistiedot mukana.



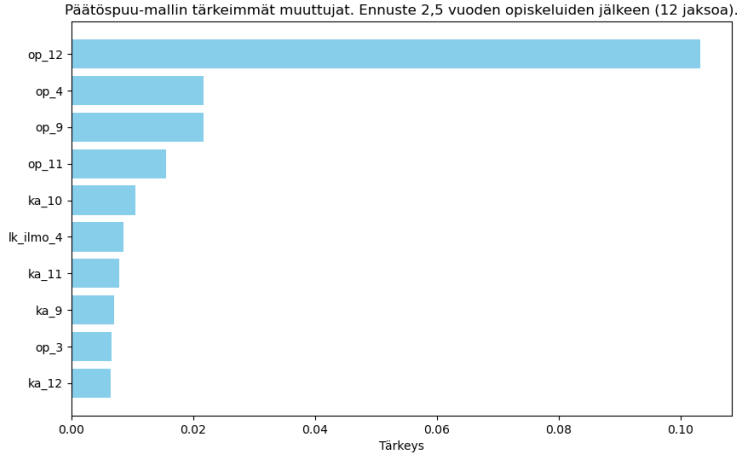
Kuva B.1: Aineisto: 17-22, ennuste 1 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



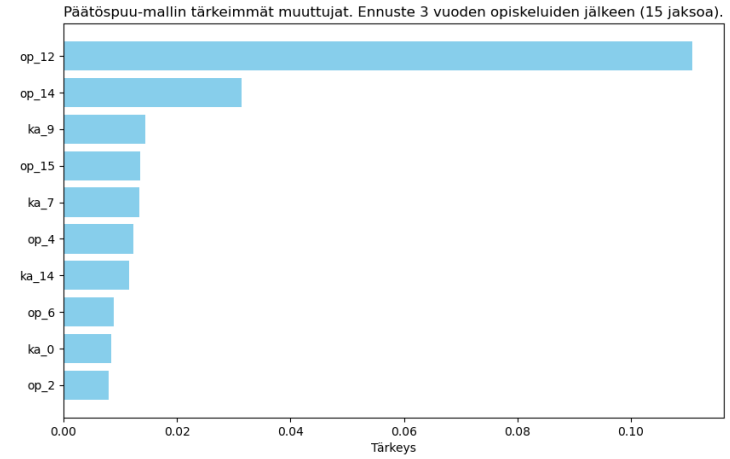
Kuva B.2: Aineisto: 17-22, ennuste 1,5 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



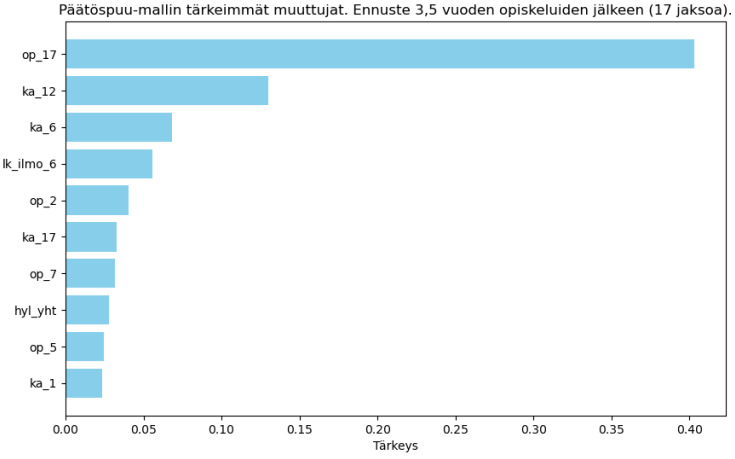
Kuva B.3: Aineisto: 17-22, ennuste 2 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



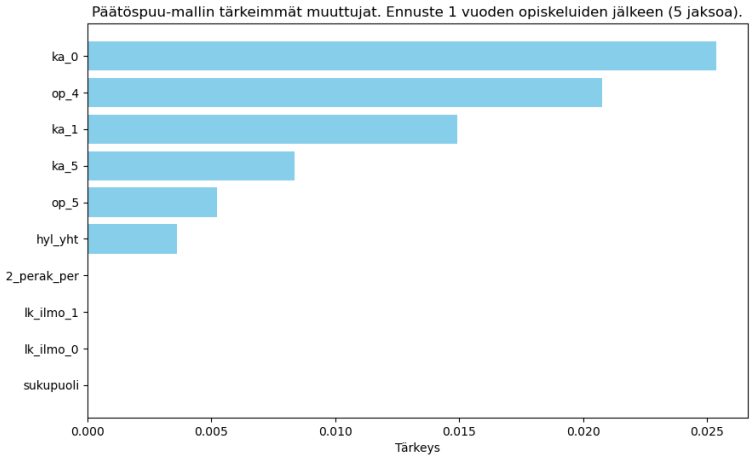
Kuva B.4: Aineisto: 17-22, ennuste 2,5 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



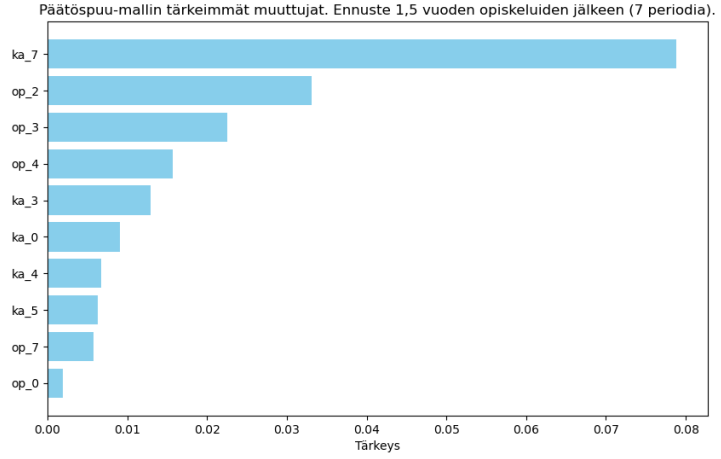
Kuva B.5: Aineisto: 17-22, ennuste 3 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



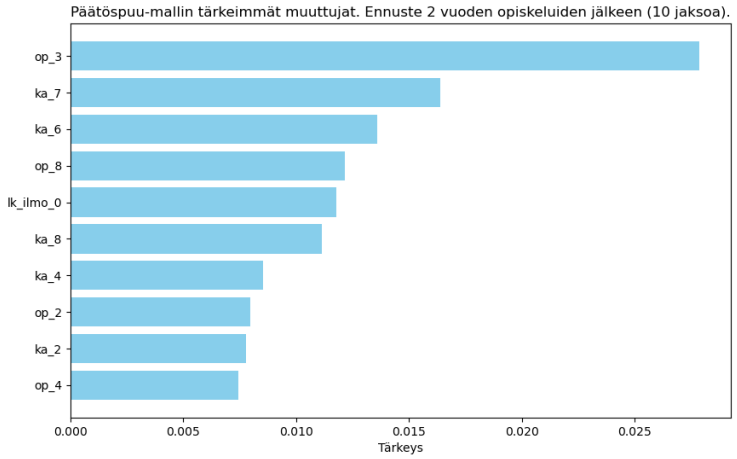
Kuva B.6: Aineisto: 17-22, ennuste 3,5 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



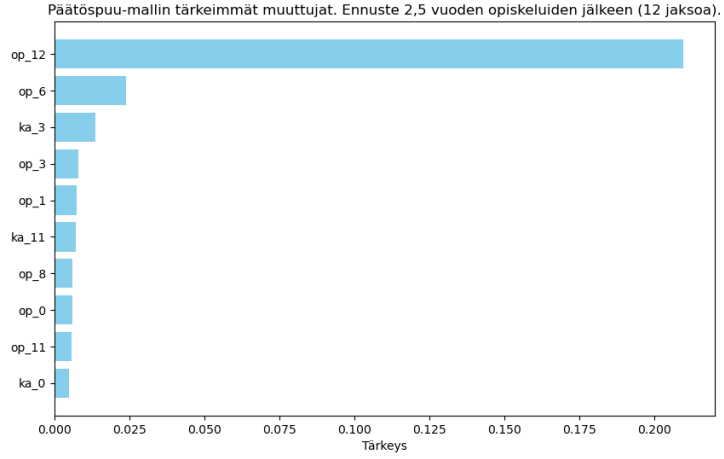
Kuva B.7: Aineisto: 20-22, ennuste 1 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



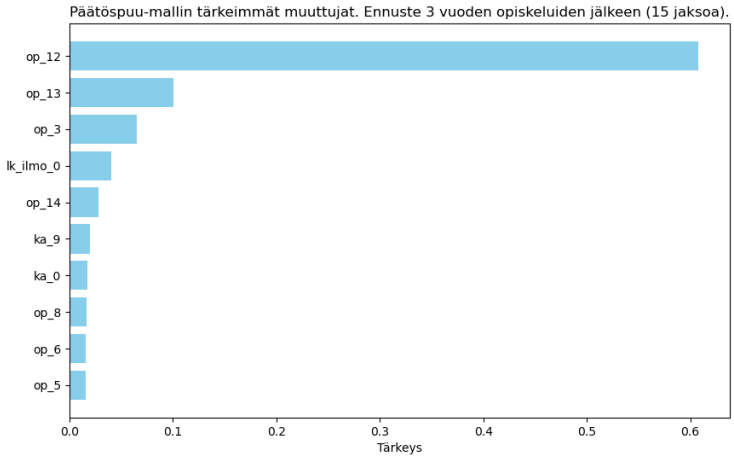
Kuva B.8: Aineisto: 20-22, ennuste 1,5 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



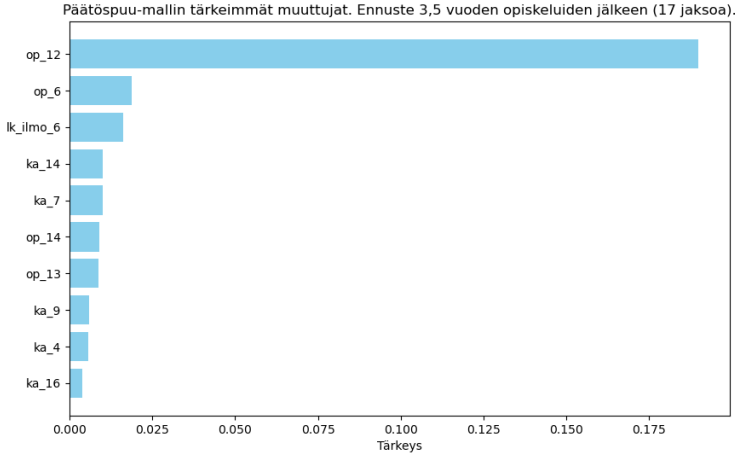
Kuva B.9: Aineisto: 20-22, ennuste 2 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



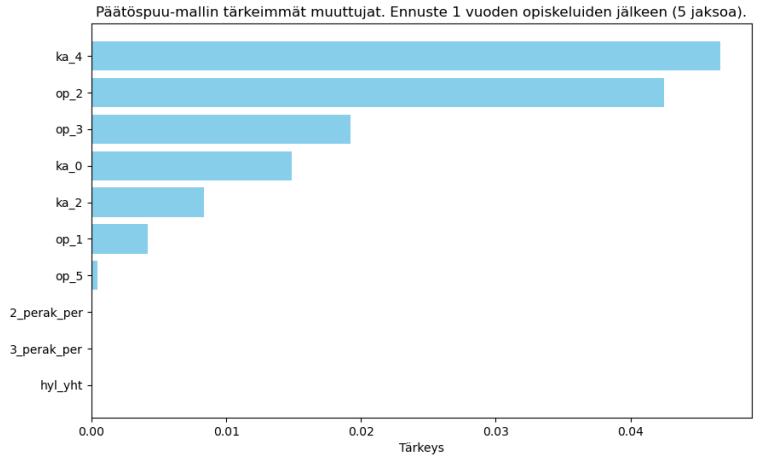
Kuva B.10: Aineisto: 20-22, ennuste 2,5 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



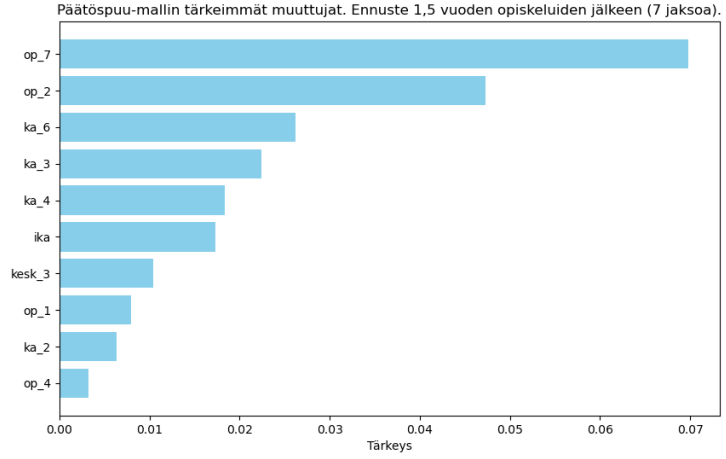
Kuva B.11: Aineisto: 20-22, ennuste 3 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



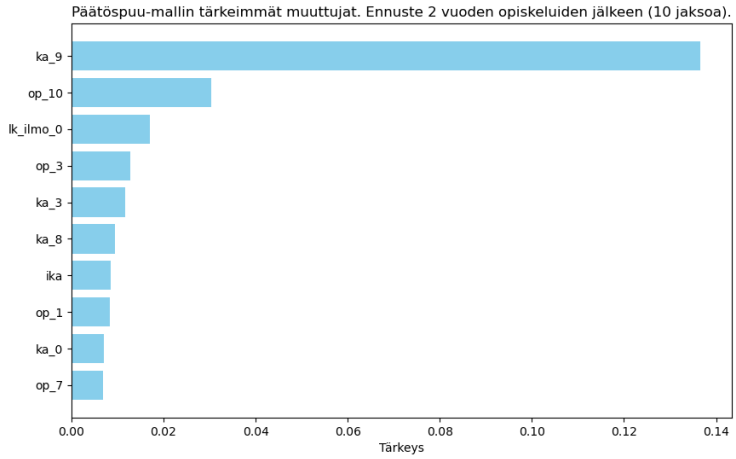
Kuva B.12: Aineisto: 20-22, ennuste 3,5 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



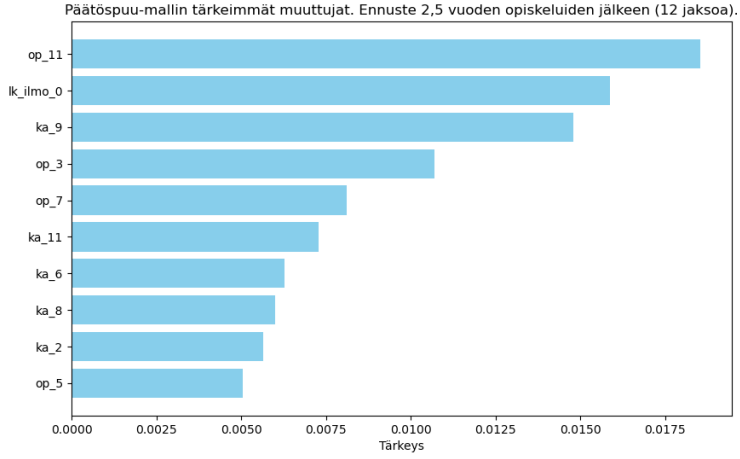
Kuva B.13: Aineisto: 20-22_ilm, ennuste 1 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



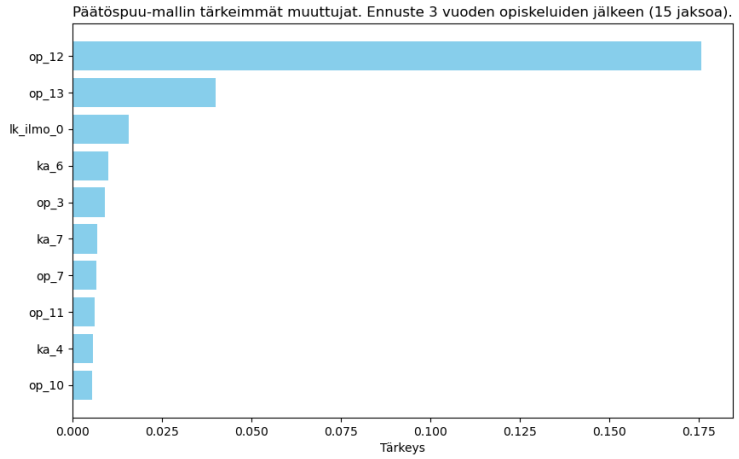
Kuva B.14: Aineisto: 20-22_ilm, ennuste 1,5 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



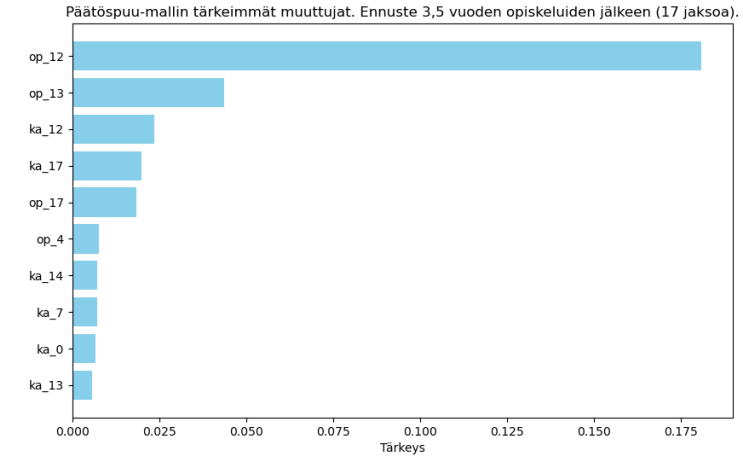
Kuva B.15: Aineisto: 20-22_ilm, ennuste 2 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



Kuva B.16: Aineisto: 20-22_ilm, ennuste 2,5 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



Kuva B.17: Aineisto: 20-22_ilm, ennuste 3 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).



Kuva B.18: Aineisto: 20-22_ilm, ennuste 3,5 vuoden opiskeluiden jälkeen. Tärkeyden mittarina on puun solmujen jakamisessa esiintymisten määrä painotettuna jakamisella saadulla hyödyllä (gini-kertoimella).