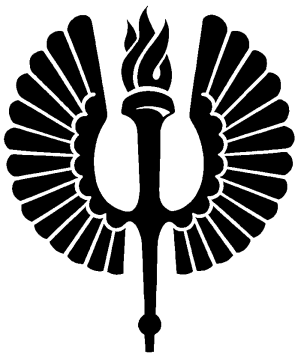


Tulevaisuuden teknologioiden laitoksen ensimmäisen vuoden opiskelijoiden opintopistekertymiin vaikuttavat tekijät

Henri Kajasilta

Pro gradu -tutkielma
Joulukuu 2019

MATEMATIIKAN JA TILASTOTIETEEN LAITOS
TURUN YLIOPISTO

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

KAJASILTA, HENRI: Tulevaisuuden teknologioiden laitoksen ensimmäisen vuoden opiskelijoiden opintopistekertymiin vaikuttavat tekijät
Pro gradu -tutkielma, 40 s.
Sovellettu matematiikka
Joulukuu 2019

Tässä tutkielmassa pyritään selvittämään ensimmäisen lukuvuoden opintopistekertymiin vaikuttavia tekijöitä. Aineistona on Turun yliopiston tulevaisuuden teknologioiden laitoksen ensimmäisen vuoden opiskelijat, jotka ovat vastanneet lukuvuoden alussa taustatietokyselyyn. Taustatietokyselyn perusteella muodostetaan muuttujia, joita käytetään luokittelu- ja regressiomalleissa selittämään opintopistekertymiä. Kaikki tässä työssä käytetyt menetelmät on toteutettu R-ohjelmointikielellä.

Taustatietokyselyn perusteella muodostetaan aihepiirimuuttujia, jotta saadaan paremmin havainnoitua aihealueita verrattuna kysymysten käyttämiseen itsenäisinä muuttujina. Kirjallisuuden perusteella tietyt aihealueet ovat merkityksellisiä korkeakouluopinnoista valmistumisen tai keskeyttämisen selittäjinä, kuten opiskelijoiden motivaatio ja haasteet elämässä. Nämä aihealueet ovat edustettuina myös aihepiirimuuttujissa.

Aihepiirimuuttujat olivat kaikissa käytetyissä menetelmissä merkityksellisimpiä selittäviä muuttujia. Näiden muuttujien perusteella tehtyjen klustereiden klusterikohtaisissa opintokeskiarvoissa oli tilastollisesti merkitseviä eroja. Lineaarisessa regressioanalyysissa aihepiirimuuttujat olivat vaikutukseltaan selvimmät ja satunnaismetsä-menetelmässä opiskelijoiden jakamisessa vaikuttavimmat muuttujat. Käyttökelpoisia ennustemalleja taustatietokyselyn muuttujilla ei saatu rakennettua. Tässä tutkielmassa tehdyt havainnot merkityksellisistä muuttujista ovat kuitenkin aiempien tutkimuksienkien perusteella tärkeitä ennustemallien kehittämisessä.

Asiasanat: opintopistekertymä, luokittelu- ja regressiomallit, vaikuttavat tekijät, ennustemalli, oppimisanalytiikka, taustatietokysely.

Sisällysluettelo

1 Johdanto

Sähköisten oppimisjärjestelmien yleistyessä oppilaitoksille kertyy paljon tietoa opiskelijoista. Sähköiset alustat ovat lisääntyneet opetuskäytössä, ja harjoitustehtävien teosta tallentuu tuloksien lisäksi riveittäin lokitietoja palvelimille. Suurimmassa osassa suomalaisia kouluja on myös käytössä alustoja, joiden avulla voidaan kommunikoida oppilaiden vanhempien kanssa ja joilla voidaan tehdä oppilaita koskevia tuntimerkintöjä. Opettajat tekevät tiedostamattaankin havaintoja opiskelijoiden menestymisestä eri oppiaineissa, ja varsinkin peruskoulussa tämä korostuu samojen opettajien ja oppilaiden ollessa paljon tekemisissä keskenään. Oppimisjärjestelmistä saatavat tiedot hyödyttävät tällä hetkellä pitkälti vain tietoihin liittyvää oppilasta ja hänen opettajiaan, vaikka suuria määriä aineistoa syntyy päivittäin.

Useissa oppimisalustoissa on siirrytty visualisoimaan opiskelijoiden suorituksia ja hyödyntämään niin kutsuttua oppimisanalytiikkaa. Oppimisanalytiikalla tarkoitetaan opiskelijoiden oppimisen mittaamista keräämällä tietoja oppilaiden osaamisesta, kerätyn tiedon analysoimista ja analyysin tulosten hyödyntämistä. Oppimisanalytiikka on käsitteenä suhteellisen uusi ja sitä ei ole vielä tutkittu paljoakaan (Järvinen et al. 2018). Yleispäteviä algoritmeja opiskelijoiden osaamisen ja potentiaalin mittaamiseen ei ole, mutta on kehitetty erilaisia malleja, joita voidaan käyttää oppilaan kykyjen ja kehittymisen arviointiin tämän aiemman suoriutumisen perusteella (Yang ja Li 2018).

Tämä tutkielma keskittyy yliopisto-opiskelijoihin ja heidän yhden lukuvuoden aikana suorittamiinsa kursseihin. Kiinnostuksen kohteena ei ole opiskelijan suoriutuminen yksittäisillä kursseilla tai hänen saamansa arvosanat. Tässä työssä pyritään tilastollisin menetelmin kartoittamaan lukuvuoden aikana saavutettuihin opintopisteisiin vaikuttavia tekijöitä. Aineistoina käytetään opiskelijoista kertyvää opintorekisteriä sekä taustatieto-kyselyitä, joihin opiskelijat ovat saaneet vapaasti vastata. Tavoitteena on tarkastella opiskelijoiden opintopistemääriä ensimmäisen lukuvuoden jälkeen. Tähän liittyvät tulokset ja havainnot toimivat suuntaa antavana informaationa siitä, mitkä aiheet ja kysymykset ovat merkityksellisiä opiskelijoiden opintojen etenemisessä ja minkälaisiin asioihin jatkotutkimuksissa olisi kiinnostavaa painottua. Tämä työ keskittyy nimenomaan yliopisto-opiskelijoihin, joilla on selvästi peruskoululaisia enemmän vapauksia opinnoissaan, mutta tehdyt havainnot voivat olla hyödyllisiä myös muiden asteiden opinnoissa.

Oppimisanalytiikan yleistyessä kehitetään yhä uusia lähestymistapoja rekisteriaineistojen käyttämiseen. Tavoitteena on osaltaan helpottaa opettajien työtä, mutta pääpaino on opiskelijoiden oppimisen seuraamisessa, jotta resursseja osattaisiin keskittää oikeisiin tarpeisiin ja näin ollen oppimistuloksia ja oppilaan kehittymistä saataisiin edistettyä. Tässä työssä keskitytään vain rajalliseen aineistoon, mutta useammista lähteistä saatavat opiskelijoiden tiedot voivat auttaa tulevaisuudessa avaamaan uusia näkökulmia. Aineistojen määrän kasvaessa myös analytiikka muodostuu haastavammaksi ja olennaisiin aiheisiin keskittyminen vaikeutuu. Tämä työ pyrkii myös selvittämään, minkälaiset aihealueet ovat analyyseissä tarpeellisia.

2 Tutkimuksen tarkoitus

2.1 Tutkimusongelma

Tämän tutkielman tarkoituksena on arvioida ja ennustaa, mitkä seikat vaikuttavat opiskelijoiden saavuttamiin opintopistemääriin. Opintopisteitä kertyy suoritetuista kursseista, ja useimmiten kurssi suoritetaan läpäisemällä sen lopussa järjestettävä tentti. Opintopistekertymään vaikuttaa ainakin kaksi asiaa. Ensimmäinen on opiskelijan oma motivaatio, jota tarvitaan, jotta opiskelija ylipäätänsä osallistuisi kursseille ja olisi kiinnostunut tekemään niiden vaatimat tehtävät. Toinen seikka on opiskelijan taitotaso, joka vaikuttaa siihen, miten hyvin opiskelija onnistuu tehtävien ratkaisemisessa. Omaa motivaatiota voidaan pitää näistä kahdesta tärkeämpänä, koska motivoituneelle opiskelijalle on usein tarjolla kurssin aikana apua sekä muilta opiskelijoilta että kurssin vastuuhenkilöiltä.

Motivaation ja taitojen mittaaminen ei ole yksiselitteistä, ja edes näiden käsitteiden rajaaminen on haastavaa. Opiskelijan saavuttamiin opintopistemääriin vaikuttaa moni tekijä, joita ei välttämättä pystytä ottamaan huomioon millään tavalla. Osaa elämänmuutoksista on mahdotonta ennustaa. Suuria muutoksia opiskelijan elämässä voivat aiheuttaa esimerkiksi opiskelijan ihmissuhteiden muutokset tai toimeentuloon liittyvät ongelmat. Tällaiset muutokset, jotka vaikuttavat muun muassa opiskelijan motivaatioon ja tätä kautta opiskeluiden etenemiseen, ovat analyysinkin kannalta haastavia. Yksi mahdollisuus haasteiden välttämiseksi olisi tehdä kyselyitä opiskelun eri vaiheissa, jotta ainakin osa muutoksista havaittaisiin. Vastausprosentit eivät kuitenkaan välttämättä olisi tarpeeksi suuria.

Opiskelua ei voi täysin irrottaa opiskelijan muusta elämästä, joten on tärkeää saada tietoa opiskelijoista myös opintojen ulkopuolelta. Keskittymällä vain siihen informaatioon, joka oppilaitoksille kertyy kurssien suorittamisesta, jätetään väistämättä hyödyntämättä tietoa opiskelijoiden opintojen ulkopuolisesta elämästä. Tässä tutkimuksessa tarkoituksena on käyttää informaatiota, joka kattaa opintoja suuremman osan opiskelijan elämästä. Tällaisen tiedon kriteerinä on, että se on tarpeeksi yleistä kerättäväksi kaikilta, mutta silti sisällöltään hyödyllistä ja kiinnostavaa. Tämänkin jälkeen tuloksia täytyy tulkita varovaisesti ja johtopäätöksiä tehdä harkiten.

Opintopisteiden kertymistä kuvaavien mallien rakentaminen on vaikeaa. Ongelman laajuudeen vuoksi vaikutukseltaan pienetkin tekijät voivat olla arvokkaita, jos ne vaikuttavat kaikkiin tai ainakin suurimpaan osaan opiskelijoista. Tavoitteena on kerätä havaintoja, jotka yleistyvät isoon opiskelijajoukkoon tai ovat ainakin paikkaansa pitäviä vielä vuosienkin kuluttua tutkimuksen oppiaineiden opiskelijoille.

2.2 Aineiston esittely

Tässä työssä käytetään aineistona tietoja Turun yliopiston tulevaisuuden teknologioiden laitoksen ensimmäisen vuoden opiskelijoista lukuvuosilta 2015–16 ja 2016–17. Opiskelijoiden pääaineina ovat lähtökohtaisesti tietojenkäsittelytiede tai tietotekniikka. Aineisto kerättiin kahdesta tietokannasta. Ensimmäinen ja työn kannalta tärkein aineisto saatiin yliopiston opintorekisteristä. Tämä aineisto koostuu suoritetuista kursseista, joiden perusteella saadaan selville myös opiskelijoiden keräämät opintopisteet.

Yliopiston on välttämätöntä pitää tietokantaa opiskelijoiden suorituksista, joten kaikkien opiskelijoiden opintoihin liittyvät tiedot tallennetaan opintorekisteriin.

Toinen aineisto saadaan Turun yliopiston kehittämän ViLLE-järjestelmän kautta, jossa tulevaisuuden teknologioiden laitoksella aloittaneista opiskelijoista on kerätty tietoa vapaaehtoisten kyselyiden pohjalta. Taustatietokyselyn motiivina on kartoittaa tietoa opiskelijoista tämän työn kaltaisten analyysien edistämiseksi sekä antamaan aikaisempaa selkeämpi kuva oppilasmateriaalin ajallisista muutoksista.

2.3 Aiempi tutkimus

Opintojen etenemiseen ja niistä suoriutumiseen vaikuttavat monet tekijät. Opiskelijalla pitää ensinnäkin olla sisäistä kiinnostusta opiskelemaansa alaa kohtaan, tai ainakin muita motivaationa toimivia tekijöitä, kuten mahdollisuus hyväpalkkaiseen työhön, jotta hän jaksaa osallistua oppiaineensa kursseille kuukaudesta toiseen. Opiskelijalla pitää olla aikaa opiskella, joten esimerkiksi työt ja harrastukset eivät saa viedä kaikkea aikaa. Asumisen ja toimeentulon pitää olla turvattuina opiskelujen onnistumiseksi. Aiheesta tehdyt aiemmat tutkimukset painottuvat pitkälti arvosanojen (Cheung 2004), valmistumisen (Ashraf et al. 2018) ja keskeyttämisen ennustamiseen (Rautopuro ja Korhonen 2011).

Tarkastellaan ensiksi opintojen keskeyttämiseen liitettyjä tekijöitä. Keskeyttämiseen myötävaikuttavat selvimmin työssäkäynti, motivaation puute ja ongelmat elämässä (Rautopuro ja Korhonen 2011). Työssäkäynti hidastaa opintojen etenemistä viemällä aikaa opiskelulta ja vaikuttaa luonnollisesti myös jaksamiseen. Tämä yksinään voi johtaa liian suureen kuormitukseen ja mahdolliseen loppuun palamiseen. Työssäkäynnillä opiskelun ohella on havaittu olevan kuitenkin myös ristiriitaisia vaikutuksia (Tuononen et al. 2016). Osa-aikaisena työssäkäyvät voivat esimerkiksi tuntea työssäkäynnin hyvänä vastapainona opiskelulle, mikä auttaa jaksamaan, kun taas toiset opiskelijat kokevat työssäkäynnin liian raskaaksi opiskeluiden ohella. Keskeyttämisriskiä lisäävät myös masennus ja ihmissuhteisiin liittyvät ongelmat. Tekijät voivat olla vahvasti sidoksissa toisiinsa sekä saattavat olla syy-seuraussuhteessa keskenään, joten perimmäistä keskeyttämisen syytä ei ole aina helppo jäljittää. Elämän vastoinkäymiset saattavat johtaa myös tietynlaiseen laiskuuteen, opintoja ei koeta merkityksellisiksi ja tavoitteiden ja opintojen suunta ovat opiskelijalta hukassa. Tämän johdosta kuulumattomuuden ja osaamattomuuden kierre syvenee, mikä lisää keskeyttämisriskiä entisestään.

Keskeyttämiseen vaikuttavia tekijöitä on hyvä verrata valmistumista selittäviin tekijöihin. Ainakin kolmen asian on todettu vaikuttavan periksiantamattomuuteen ja sitä kautta tutkinnon valmiiksi suorittamiseen; itseohjautuvuus, kuuluvuuden tunne ja arvostus opinto-ohjelmaa kohtaan (Ashraf et al. 2018). Opiskelijat, jotka aikaisessa vaiheessa valitsevat pääaineensa ovat sitoutuneempia jatkamaan sen parissa loppuun asti. Päättäväiset opiskelijat, jotka eivät ole valintaa tehneet, vaihtavat myöhemmässäkin opiskelun vaiheessa sopivalta tuntuvampaan pääaineeseen, vaikka se automaattisesti tarkoittaisikin valmistumisen viivästymistä.

Kirjallisuuden perusteella näyttää siis siltä, että nimenomaan motivaatio ja kuuluvuuden tunne ovat tärkeitä valmistumisen osatekijöitä. Motivoituneella opiskelijalla on usein päämäärä, joka auttaa jaksamaan. Kun opiskelija kokee alan omakseen, opiskeluiden eteneminen tuntuu merkitykselliseltä. On erilaisia tapoja hallinnoida motivaatiota ja

nämä liittyvätkin vahvasti edellä mainittuihin teemoihin. Motivaation hallintaan kuuluvat tavoitteiden asettaminen ja positiivisten uskomusten luominen omista kyvyistään. Itseluottavaiset opiskelijat ovat pitkäjänteisempiä toimimaan strategioidensa mukaisesti ja motivoituneempia saavuttamaan itse asettamansa tavoitteet verrattuna ulkopuolelta tulleisiin tavoitteisiin (Cheung 2004). Ylipäätään opiskelijat kokevat tavoitteiden asettamisen johtavan parempiin tuloksiin. Tätä on tutkittu arvosanojen perusteella, mutta samanlainen vaikutus voisi hyvin päteä myös opintopistemääriin.

Kollaboratiivista oppimista voidaan käyttää oman osaamisen vahvistamiseen (Häkkinen ja Arvaja 1999). Kollaboratiivisessa oppimisessa käytetään erityisiä oppimisen menetelmiä, kuten artikulointia. Opiskelijat joutuvat selittämään toisilleen oppimiseen liittyviä asioita ja ehkä ajattelemaan aihetta sen johdosta uudesta näkökulmasta. Myöskin muiden opiskelijoiden kanssa toimimisen sosiaalista puolta ei sovi väheksyä ja tällä on varmasti merkitystä myös kuuluvuuden tunteelle.

3 Tutkimusmenetelmät

Tässä työssä mallinnetaan opiskelijoiden ensimmäisen opiskeluvuotensa aikana saavuttamaa opintopisteiden määrää. Vastemuuttujana on opintopistemäärä joko suoraan tai opintopistemäärästä muodostettu binäärinen muuttuja (eli havainnot luokitellaan arvoltaan joko nollaksi tai ykköseksi). Taustatietokyselyn vastaukset ovat tässä työssä selittäviä muuttujia, joita myös ennustemuuttujiksi kutsutaan. Selittävällä muuttujalla tarkoitetaan muuttujaa, jonka muutokset selittäisivät vastemuuttujan vaihtelua.

Tässä työssä käytetään menetelmiä, jotka perustuvat luokitteluun tai regressioon. Luokitteluongelmissa pyritään tiettyjen ennustemuuttujien avulla luokittelemaan yksilöitä mahdollisimman hyvin vastemuuttujan eri kategorioihin. Luokitteluongelmat jaetaan *ohjattuun oppimiseen* ja *ohjaamattomaan oppimiseen* sen mukaan, onko aineiston luokittelusta tietoa ennalta. Ohjatussa oppimisessa hyödynnetään olemassa olevaa tietoa tilastoyksilöiden oikeasta luokittelusta niin sanotussa opetusaineistossa, ja tämän perusteella opetetaan mallia toimimaan myös tuntemattomilla syötteillä (ennustemuuttujien arvoilla). Myös ohjaamattomassa oppimisessa aineisto jaetaan joukkoihin, mutta luokittelu tapahtuu ennustemuuttujien samankaltaisten ominaisuuksien pohjalta. Tällöin ei ole saatavilla valmista opetusaineistoa, jonka perusteella luokittelusta voitaisiin oppia. Klusterointi on ohjaamatonta oppimista, jossa valittujen muuttujien perusteella pyritään muodostamaan keskenään samankaltaisia ryhmiä eli klustereita. Muodostettujen klustereiden ominaisuuksia voidaan tämän jälkeen vertailla keskenään sekä tarkastella klustereiden eroja niissä huomiotta jätettyjen muuttujien kohdalla.

Regressiossa kiinnostuksen kohteena on vastemuuttujan ja selittävien muuttujien välinen yhteys. Tässä työssä tarkastelu keskittyy siihen, mitkä selittävistä muuttujista ovat ongelmassamme merkityksellisiä ja kuinka paljon ne vaikuttavat vastemuuttujan odotusarvoon.

3.1 Menetelmien esittely

Tämän tutkimuksen aineiston analyysissä käytettävien menetelmien valinta ei ole täysin suoraviivaista. Suurin valintaan vaikuttava tekijä on analyysin laatu eli se, minkälaisia asioita aineistosta halutaan selvittää. Koneoppimisen suosion kasvu on nostanut erilaiset klusterointimetodit ja neuroverkot laajasti sovelletuiksi apuvälineiksi. Myös tässä työssä käytetään eri lähestymistapoja opintopisteisiin vaikuttavien tekijöiden kartoittamiseksi käyttämällä pääkomponenttianalyysia, klusterointia, regressiomalleja sekä päätöspuita.

Pääkomponenttianalyysia käytetään yleisesti pienentämään aineistossa esiintyvien dimensioiden määrää ja sitä kautta multikollineaarisuutta. Myös opiskeluun liittyvissä tutkimuksessa se on usein hyödynnetty apukeino. Gaertner ja McClarty (2015) tarkastelivat opiskelijoiden yliopistovalmiutta kuuden osa-alueen avulla: opintomenestys, motivaatio, käyttäytyminen, sosiaaliset suhteet, perheolosuhteet ja kouluympäristö. Ennustemuuttujat jaettiin ensin osa-alueisiin, minkä jälkeen pääkomponenttianalyysilla saatiin osa-alueiden sisältämistä muuttujista laskettua uudet muuttujat. Ennuste-muuttujien määrä vähennettiin tällä tavoin 140:stä kuuteen.

Cortez ja Silva (2008) puolestaan ennustivat kursseista suoriutumista matematiikassa ja portugalin kielessä (opiskelijoiden äidinkielessä). Opiskelijoista oli kerätty taustatietoja, esimerkiksi perheen taustoista ja opiskelijan vapaa-ajan vietosta, sekä aiemmasta koulumenestyksestä. Tutkimuksessa käytettiin monipuolisesti eri metodeja kuten regressiomalleja, *Naive Bayes* -luokittelua ja neuroverkkoja. Eri menetelmät antoivat samansuuntaisia tuloksia, mutta mikä tärkeintä, hyvin valituilla ennustemuuttujilla ennusteet kurssin suorittamisesta olivat tarkkoja. Binäärisessä luokittelussa (opiskelija läpäisee tai ei läpäise kurssia) opiskelijat osattiin luokitella jopa yli 90% tarkkuudella oikein. Tutkimuksessa oli mukana satoja opiskelijoita, joista matematiikassa lähes kolmannes ei saanut koetta hyväksytysti läpi. Ennusteiden luotettavuus nojasi vahvasti samojen opiskelijoiden aiempien kokeiden tuloksiin. Tässä tutkielmassa ei kuitenkaan ole käytettävissä vastaavanlaista tietoa opintopisteiden ennustamiseksi.

Tässä työssä analyysivaihe jakautuu kolmeen osaan. Ensimmäisessä osassa käytetään klusterointia, jonka avulla opiskelijat pyritään jakamaan ryhmiin ennalta valituin kriteerein niin, että ryhmien sisällä on mahdollisimman samankaltaisia opiskelijoita. Tämän jälkeen vertaillaan näiden ryhmien opintopistekertymiä. Toinen osuus kattaa lineaarisen regressioanalyysin, jolla tutkitaan ennustemuuttujien vaikutusta lukuvuoden aikana kerättyihin opintopisteisiin. Kolmannessa osuudessa pyritään luokittelemaan opiskelijat niihin, jotka saavuttavat tietyn opintopisterajan, ja niihin, jotka eivät tätä opintopisterajaa saavuta. Tähän luokitteluun käytetään *satunnaismetsä*-menetelmää. Samoja muuttujia käytetään ainakin osittain jokaisessa analyysivaiheessa, mutta eri vaiheissa käytetyt lähestymistavat poikkeavat toisistaan. Näin saadaan kattavampi yleiskuva siitä, mitä analyysien avulla voidaan tehdä tai päätellä.

Tässä työssä käytetään myös muita tilastollisia menetelmiä, jotka toimivat edellä mainittujen kolmen keskeisimmän menetelmän apuna. Klusterointia varten hyödynnetään *skaalausta*, jotta käytetyt muuttujat olisivat lähtökohtaisesti vaikutukseltaan yhtä suuria. Toisin sanoen näin ehkäistään, etteivät pienen kokoluokan muuttujat jäisi analyysissa merkityksettömiksi suuremman kokoluokan muuttujien kanssa, ja klusterit olisivat muuttujien suhteen tasavertaisesti muodostettuja. Klusteroinnin, lineaarisen regression ja luokittelun yhteydessä käytetään bootstrap-metodia luottamusvälien laskemisessa. Tällä tavoin saadaan arvioitua klusterikohtaisten keskiarvojen eroja luotettavasti,

lineaarisessa regressioanalyysissa voidaan kuvata kiinnostavien muuttujien merkitystä mallissa ja luokittelussa mallin todellista ennustevoimaa.

Ennustemuuttujista pyritään myös muodostamaan uusia muuttujia, jotka kuvaisivat kattavammin tiettyä aihepiiriä. Näistä uusista muuttujista käytetään jatkossa nimitystä aihepiirimuuttuja tai *aihepiiri*. Taustatietokyselyn muuttujien subjektiivisen tarkastelun lisäksi hyödynnetään pääkomponenttianalyysia, jolloin saadaan objektiivisempi tapa arvioida, mitkä muuttujista erottelevat yksilöitä, ja mitkä antavat samanuuntaista informaatiota. Pääkomponenttianalyysia käytetään muuttujien jakamiseksi *aihepiireihin*, eikä sitä hyödynnetä uudestaan analyysivaiheessa. Pääkomponenttianalyysi tulkitaan tässä työssä luokittelumalliksi, koska sitä käytetään *aihepiirien* muodostamisen apuna.

3.2 Analyysia tukevat metodit

3.2.1 Skaalaus

Yksinkertaisimpana metodina tässä työssä käytetään skaalausta. Taustatietokyselyssä osa vastauksista on annettu Likert-asteikolla 1–5 (*Täysin eri mieltä – Täysin samaa mieltä*), ja ainakin osa kysymyksistä antaa samansuuntaisia vastauksia. Tämän takia kysymyksistä muodostetaan aihepiirejä, joiden voidaan ajatella antavan kokonaisvaltaisesti informaatiota aina yhdestä opiskelijan elämän osa-alueesta.

Skaalaus on tehty suoraviivaisesti kaavalla $y_{ij} = (\overline{x_{ij}} - 1)/4$, jossa muuttuja $\overline{x_{ij}}$ kuvaa aihepiiriin j liitettyjen kysymysten vastausten keskiarvoa oppilaan i kohdalla. Yhteen kysymykseen opiskelija i voi antaa vastauksena korkeintaan arvon 5 ja minimissään arvon 1, joten uusi muuttuja y_{ij} saa arvoja väliltä [0,1].

Poikkeuksena oli kysymys peruskoulun päättötodistuksen keskiarvosta, jossa mahdollinen vaihteluväli on 4-10. Myös tämän vastauksen mahdollinen vaihteluväli skaalattiin välille [0,1], minkä jälkeen se saatiin liitettyää aihepiiriinsä niin, ettei sen vaikutus ole toista kysymystä suurempi.

3.2.2 Bootstrap

Bootstrap on menetelmä, jonka avulla voidaan laskea tilastollisten tunnuslukujen keskivirheitä ja luottamusvälejä (Efron ja Tibshirani 1986). Näin saadaan yksittäisen tunnusluvun lisäksi johdettua tunnusluvun tarkkuus. Tämä auttaa, kun vertaillaan eri populaatioista saatuja tunnuslukuja keskenään. Bootstrap on tilastollinen menetelmä, jossa hyödynnetään aineistoon satunnaista otantaa palauttaen. Ajatuksena on tehdä aineistosta samankokoisia otoksia, ja koska jokainen valittu havainto palautetaan aineistoon, voi yksi havainto olla edustettuna jokaisessa bootstrap-otoksessa useamman kerran. Vastaavasti jotkin havainnot eivät tule valituksi ollenkaan. Keskimäärin noin 2/3 havainnoista on edustettuna yksittäisessä bootstrap-otoksessa. Toistamalla näitä otoksia ja laskemalla jokaiselle esimerkiksi keskiarvo saadaan estimoitua keskiarvoestimaattorin jakaumaa yhden aineiston perusteella. Tässä työssä hyödynnetään tavallisesta bootstrap-metodista paranneltua versiota nimeltään BC_a-bootstrap (*Bias-corrected and accelerated bootstrap*). Suurin hyöty BC_a-bootstrapissa tavalliseen bootstrap-metodiin verrattuna on, että se huomioi harhan ja vinouman bootstrap-estimaattien jakaumassa.

Bootstrap-jakauman ollessa positiivisesti vino luottamusväliä korjataan oikealle ja vastaavasti jakauman ollessa negatiivisesti vino luottamusväliä korjataan vasemmalle. BC_a-bootstrapissa hyödynnettyjen korjausten on havaittu antavan paremmat luottamusvälit tarkasteltavalle tunnusluvulle (Efron 1987).

Bootstrap on kätevä apukeino luottamusvälien ja tätä kautta tilastollisen merkitsevyyden laskemiseen. Verrattuna parametrisiin metodeihin, kuten t-testiin, bootstrap-metodi ei nojaa olettamuksiin havaintojen jakaumasta. Kaikki luottamusvälit ovat tässä työssä laskettu 95%:n luottamustasolla.

3.3 Luokittelumallit

3.3.1 Klusterointi

Klusteroinnin päätavoite on jakaa aineistossa esiintyvät havainnot keskenään samankaltaisiin ryhmiin eli klustereihin. Klusterointimenetelmiä on erityyppisiä. Ehkä tunnetuimpia ovat hierarkkinen klusterointi, jossa klusterit muodostavat puumaisen hierarkiarakenteen, tai osittava klusterointi, jossa ensin muodostetaan klusterikeskuksia, joihin havaintoja sitten sijoitetaan. Näistä menetelmistä on vielä monenlaisia variaatioita, joissa on kuitenkin sama perusajatus. Tässä työssä käytetään *Fuzzy C-Means* -klusterointia (FCM-klusterointi).

FCM-klusterointialgoritmi on hyvin samankaltainen kuin tunnetuimpiin klusterointimenetelmiin kuuluva K-mean-klusterointi, jossa tavoitteena on minimoida klusterien muodostamiseen käytettyjen muuttujien klusterien sisäistä varianssia. FCM-klusteroinnissa havainnot eivät ole tiukasti sidottuja vain yhteen klusteriin, vaan metodi mahdollistaa yksittäisten havaintojen kuulumisen useampaan klusteriin. Havainnoille annetaan jokaisen klusterin kohdalla arvo välillä [0,1], mikä kertoo, kuinka suurella todennäköisyydellä havainto tarkasteltuun klusteriin sijoitetaan. Kunkin havainnon eri klustereihin sijoittumisen todennäköisyydet summautuvat ykköseksi.

FCM-klusteroinnissa (Bezdek 1974; Xie et al. 2010) minimoidaan funktiota

$$J(U, \mathbf{V}) = \sum_{i=1}^{n} \sum_{j=1}^{c} (u_{ij})^m ||x_i - v_j||^2,\ 1 \leq m < \infty,$$

jossa arvo m on (ennalta annettu) sekoitusindeksi, n on havaintojen määrä, c on klusterikeskusten määrä, x_i on aineiston i:nnes havainto ja v_j on j:nnes klusterikeskus. FCM-klusteriointi palautuu K-mean-klusteroinniksi, jos se saa arvokseen $m = 1$. Funktion $J(U, \mathbf{V})$ parametri U on luokitteluosuuksien u_{ij} $(n \times c)$ matriisi ja parametri $\mathbf{V}$ on vektori klusterikeskuksista. Välttämätön ehto funktion $J(U, \mathbf{V})$ minimoimiselle on, että

$u_{ij} = 1/\sum_{k=1}^{c} (d_{ij}/d_{ik})^{2/(m-1)}$ ja

$v_j = (\sum_{i=1}^{n} (u_{ij})^m x_i)/(\sum_{i=1}^{n} (u_{ij})^m),\ \forall j = 1, 2, ..., c.$

Tässä arvo d_{ij} on euklidinen etäisyys i:nnen havainnon ja j:nnen klusterikeskuksen välillä. Klusterikeskukset valitaan satunnaisesti alustetun matriisin U perusteella. Tämän jälkeen klustereihin sijoittumisen todennäköisyyksien arvoja ja klusterikeskuksien arvoja iteroidaan, jotta samankaltaiset havainnot sijoittuvat samoihin klustereihin. Tätä jatketaan, kunnes klusterikohtaisia variansseja ei saada enää tehokkaasti pienennettyä.

Algoritmia kutsutaan useamman kerran eri alkuarvoilla, jotta klusteroinnin tulos ei riippuisi alkuarvauksen onnistumisesta.

Optimaalisen klusterien määrän valitseminen ei ole yksiselitteistä, ja päätöstä tehdessä täytyy käyttää omaa harkintaa. Tässä työssä klusteroinnin tavoitteena on ennustaa mahdollisia eroavaisuuksia lukukauden aikana kerätyissä opintopistekeskiarvoissa. Klusterit muodostetaan aihepiirimuuttujien perusteella. Aihepiirimuuttujat ovat vaikutukseltaan yhtä suuria. Ei ole mielekästä luoda liian montaa klusteria. Ensinnäkin klusteria kohden jää vähemmän opiskelijoita klusterien määrää kasvattaessa ja näin ollen tilastollinen tarkastelu klusterien välillä muodostuu epävarmemmaksi. Toiseksi ollaan kiinnostuneita rajanvedosta opiskelijoiden välillä, jotka keräävät paljon opintopisteitä, ja niiden, jotka keräävät vain vähän opintopisteitä lukuvuoden aikana. Tämä on selkeämpää, kun klustereita on vähän.

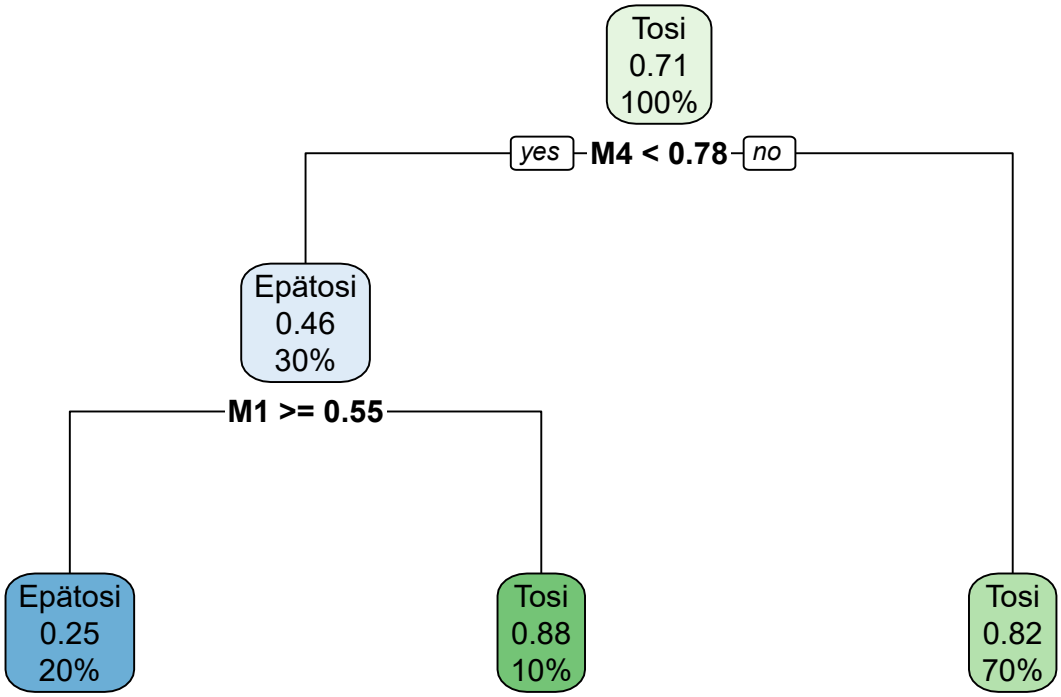
3.3.2 Satunnaismetsä

Satunnaismetsä on ohjatun koneoppimisen menetelmä, jota käytetään sekä luokittelu- että regressio-ongelmiin (Breiman 2001; Liaw ja Wiener 2002). Satunnaismetsä on päätöspuihin perustuva menetelmä. Päätöspuut koostuvat solmukohdista, jotka ovat mallin selittäviin muuttujiin perustuvia ehtoja jakaa aineistoa osajoukkoihin. Solmukohdassa tapahtuvan jaon perusteella aineisto haarautuu osajoukkoihin, jotka edelleen haarautuvat uusissa solmukohdissa pienempiin osajoukkoihin. Päätöspuun viimeisenä vaiheena on luokitella osajoukot, kun kaikki malliin otetut selittävät muuttujat on solmukohtien yhteydessä käytetty.

Puumalleissa on yleensä useita muuttujia, joiden perusteella aineistoa voidaan jakaa. Ensimmäisen solmukohdan lisääminen jakaa alkuperäisen aineiston kahtia muuttujan arvon perusteella. Tämä tarkoittaa, että jaetun aineiston molemmissa haaroissa voidaan havaita luokiteltavan muuttujan jakaumat, jotka eivät välttämättä vastaa luokiteltavan muuttujan koko jakaumaa aineistossa. Kuvassa 1 on esimerkkinä päätöspuu, jossa kahden solmukohdan avulla luokitellaan havaintoja joko *tosiksi* tai *epätosiksi*. Jos jakavat muuttujat ovat hyödyllisiä, jakaantuu aineisto niin, että molempiin haaroihin erottuvat luokittelultaan erilaiset havainnot.

Satunnaismetsässä päätöspuita luodaan useita, jotta yksittäisen päätöspuun painoarvo ei olisi liian suuri. Satunnaismetsässä kuhunkin puuhun valitaan satunnainen otos aineiston yksilöitä ja jokaisessa puun solmukohdassa kokeillaan vain tiettyä määrä satunnaisesti valikoituja muuttujia. Satunnaismetsän ajatusta voidaan kuvata yksinkertaisella esimerkillä: Jos kysyt ystävältäsi arviota tietystä ravintolasta, saat vain tämän kyseisen henkilön mielipiteen ravintolan laadusta. Kysyessäsi mielipidettä samasta ravintolasta kymmeneltä ystävältäsi yhden ystävän mieltymykset eivät saa liian suurta painoa, ja voit tehdä kattavamman päätelmän kyseisen ravintolan laadusta.

Tässä työssä satunnaismetsää käytetään nimenomaan luokitteluongelman ratkaisussa. Ennen satunnaismetsän käyttöä aineisto jaetaan kahteen osaan: *opetus-* ja *testaus*-osuuksiin. Osien suhteelliset koot ovat 0.55 (opetus) ja 0.45 (testaus), jotta myös testaus-osuuteen jää riittävästi aineistoa. Opetusaineiston avulla satunnaismetsä *harjoittelee* jakamaan havainnot oikeisiin luokkiin hyödyntämällä bootstrap-tekniikkaa. Suurinpiirtein kolmannes havainnoista ei tule mukaan yksittäistä puuta harjoitettaessa ja tämä osa havainnoista käytetään siis yksittäisen päätöspuun luokitteluvaiheessa. Lopulta



Kuva 1: Esimerkki päätöspuusta, johon on valittu kaksi solmukohtaa. Algoritmi pyrkii kahden luokiteltavan muuttujan (M1 ja M4) perusteella jakamaan havaintoaineiston yksilöt kahteen luokiteltavan muuttujan luokkaan (tosi/epätosi). Oletetaan, että koko aineistossa oikeasti tosien tapausten osuus on 0.71. Ensimmäisessä solmukohdassa aineisto jaetaan sen mukaan, onko muuttujan M4 arvo alle 0.78 vai ei. Kuvaajan aineistossa 70 prosentilla havaintoyksilöistä arvo on ainakin 0.78, ja näistä oikeasti tosien tapausten osuus on 0.82. Vastaavasti M4:n mukaan epätosiksi luokiteltujen joukossa oikeasti tosien osuus on 0.46. Tämä solmukohta jaetaan vielä muuttujan M1 arvon perusteella, jolloin 20 prosenttia aineistosta luokitellaan epätosiksi (tästä osajoukosta tosien tapauksien osuus on 0.25) ja 10 prosenttia aineistosta luokitellaan tosiksi (tästä osajoukosta tosien tapauksien osuus on 0.88).

testaus-osuutta käytetään arvioimaan mallin yleistyvyyttä. Etuna testausaineistossa on, ettei sitä ole harjoitteluvaiheessa käytetty ja näin ollen mallin yleistyvyyttä pystytään kokeilemaan *uuteen* koskemattomaan aineistoon.

Satunnaismetsä on lähtökohtaisesti robusti aineistoilla, joissa on paljon sekoittavia muuttujia ja vähän päätöksille relevantteja muuttujia. Relevanttien muuttujien pitäisi joka tapauksessa nousta esille päätöksien teossa ja tästä syystä ajatellaankin, ettei satunnaismetsällä voida *ylisovittaa* aineistoa (Breiman 2001).

Satunnaismetsän arviointiin liittyy olennaisesti muutamia käsitteitä, joita esitellään seuraavaksi. Tarkastellaan ensiksi käsitteitä sensitiivisyys (engl. *sensitivity*) ja spesifisyys (engl. *specificity*). Sensitiivisyydellä tarkoitetaan oikein (eli tosiksi) luokiteltujen *tosi*-tapausten osuutta. Esimerkiksi tässä työssä sensitiivisyydellä mitataan, kuinka monta prosenttia määrätyn opintopisterajan saavuttaineista opiskelijoista on ennustemuuttujien avulla luokiteltu oikein. Vastaavasti spesifisyydellä mitataan oikein (eli epätosiksi) luokiteltujen *epätosi*-tapausten osuutta. Tässä työssä tämä tarkoittaa niiden opiskelijoiden osuutta, jotka on oikein luokiteltu jäämään määrätyn opintopisterajan alle. Sensitiivisyyden ja spesifisyyden avulla saadaan parempi käsitys mallin käytettävyydestä verrattuna esimerkiksi pelkän tarkkuuden (engl. *accuracy*) ilmoittamiseen, varsinkin havaintojen ollessa epäsuhtaisesti jakautuneina luokkiin. Tarkkuudella tarkoitetaan tässä oikein luokittelujen havaintojen osuutta. Esimerkiksi tapauksessa, jossa 95% havainnoista kuuluu *tosi*-tapauksiin, saadaan tarkkuudeksi helposti 95%, kun luokitellaan kaikki havainnot *tosiksi*. Tässä tapauksessa kuitenkin sensitiivisyys olisi 100% ja spesifisyys 0%, joka paljastaisi mallin olevan vailla todellista selitysvoimaa.

Kynnysarvo on arvo, joka on valittu jakamaan havainnot binäärisesti kahteen luokkaan: havainnot, jotka saavat kynnysarvoa pienemmän arvon luokitellaan *epätosiksi* ja arvot, jotka ovat vähintään kynnysarvon verran, luokitellaan *tosiksi*. ROC-käyrä (*Receiver Operating Characteristic* -käyrä) kuvaa luokittelun onnistumista kaikilla luokittelun kynnysarvoilla (Fawcett 2006). ROC-käyrää ja AUC-arvoa (*Area Under The Curve* -arvo) käytetään (binäärisissä) luokitteluongelmissa usein arvioimaan mallin hyvyyttä, ja ne perustuvat suoraan sensitiivisyyden ja spesifisyyden arvoihin. Jos luokittelussa arvioidaan pääsääntöisesti enemmän tapahtumia *tosiksi* tuloksena on enemmän oikeita *tosiksi* luokiteltuja tapauksia, mutta myös vastaavasti enemmän virheellisesti *tosiksi* luokiteltuja tapauksia. Kynnysarvoa, jolla tapauksia luokitellaan *tosiksi* voidaan tilanteen mukaan muuttaa. Esimerkiksi, jos luokitellaan kaikki ne havainnot *tosiksi*, jotka ovat mallin mukaan yli 0.1 todennäköisyydellä *tosi*-tapauksia, saadaan sensitiivisyys suuremmaksi, kuin tapauksessa, jossa vaaditaan havainnoilta yli 0.5 todennäköisyyttä *tosi*-luokittelulle. Spesifisyyden arvo olisi puolestaan edellisessä tapauksessa pienempi. Tätä vaihtokauppaa kuvataan ROC-käyrän avulla.

AUC-arvo tarkoittaa ROC-käyrän alle jäävää pinta-alaa. Täydellinen malli saisi AUC-arvon 1, mikä tarkoittaisi, että kaikki havainnot on luokiteltu oikein. Satunnainen luokittelija puolestaan saisi AUC-arvokseen 0.5 ja ROC-käyrässä tämä nähtäisiin diagonaalina viivana. Lähtökohtaisesti mallin kuuluisi saada AUC-arvo väliltä [0.5,1]. AUC-arvo voi olla myös alle 0.5, mutta esimerkiksi AUC-arvo 0 tarkoittaisi, että kaikki luokittelut on tehty väärin ja täydellinen malli saataisiin kääntämällä luokittelut päinvastoin.

Mallin ennustevoiman arvioimisen lisäksi tarkastellaan satunnaismetsälle merkityksellisiä

muuttujia. Vähenevää Gini-arvoa (Han et al. 2016) käytetään arvioimaan, mitkä ovat mallin tärkeimpiä muuttujia havaintojen luokittelussa. Gini-epäpuhtaus (engl. *Gini impurity*) saadaan laskettua kaavalla $G(k) = \sum_{i=1}^{J} P(i)(1 - P(i))$, jossa $P(i)$ on luokan i suhteellinen osuus, $i \in \{1,2,\ldots,J\}$ ja J on luokkien määrä. Gini-epäpuhtauden merkitys voidaan ymmärtää seuraavasti: Valitaan aineistosta satunnaisesti yksi havainto, jonka jälkeen luokitellaan tämä havainto luokkien suhteellisten osuuksien mukaisen jakauman perusteella. Gini-epäpuhtaus on todennäköisyys, että havainto on luokiteltu väärin.

Päätöspuun jokaisen solmukohdan kahdelle haaralle voidaan laskea oikein luokiteltujen havaintojen osuudet. Näiden perusteella laskettujen Gini-epäpuhtauksien painotettu keskiarvo on ko. haaraan kuuluva Gini-epäpuhtaus (painoina haaroihin liittyvien havaintojen lukumäärät). Gini-epäpuhtauden vähenemä saadaan, kun lasketaan erotus Gini-epäpuhtauksista ennen ja jälkeen solmun muodostamisen. Vähenevä Gini-arvo on keskiarvo muuttujan aiheuttamasta epäpuhtauden vähenemisestä satunnaismetsän puiden solmukohdissa. Keskiarvo on painotettu sen mukaan, kuinka monta havaintoa yksittäisessä puussa saavuttaa kyseisen muuttujan solmukohdan. Tarkoituksena on listata muuttujat vähenevän Gini-arvon perusteella, mikä auttaa hahmottamaan muuttujien suhteellista tärkeyttä luokitteluongelmassa.

3.3.3 Pääkomponenttianalyysi

Pääkomponenttianalyysi (*Principal Component Analysis*), josta käytetään tästä eteenpäin myös lyhennettä PCA, on tilastollinen menetelmä, jossa muuttujista luodaan uusia korreloimattomia lineaarikombinaatioita. Näitä uusia muuttujia kutsutaan pääkomponenteiksi. Käytännössä pääkomponenttien muodostaminen tapahtuu niin, että ensimmäisenä muodostetulla pääkomponentilla on kaikista suurin varianssi. Seuraavat pääkomponentit ovat aina ortogonaalisia edellisiin pääkomponentteihin nähden ja jokaisen uuden pääkomponentin osuus jäljelle jääneestä varianssista on suurin mahdollinen (Jolliffe 2002). Pääkomponenttianalyysi on laajasti hyödynnetty menetelmä hahmontunnistuksessa juuri sen ansiosta, että sen avulla pyritään muutamalla komponentilla tunnistamaan suurin osa aineiston vaihtelusta. Tässä työssä pääkomponenttianalyysia käytetään taustatietokyselyn vastausten erottelemiseen selkeimpien erottelevien teemojen havaitsemiseksi.

Määritellään ensin käsitteitä, minkä jälkeen selitetään pääkomponenttianalyysin toimintaperiaate.

- Vektori $\mathbf{x} = [x_1, x_2, ..., x_p]^T$, jossa merkintä ^{T} on transpoosi. Tämä vektori sisältää selittävät muuttujat.
- Vektorilla $\mathbf{x}$ on tunnettu kovarianssimatriisi $\sum$. Kun kovarianssimatriisia ei tiedetä, käytetään otoskovarianssimatriisia.
- Lineaarikombinaatio on vektorin $\mathbf{x}$ ja skalaarikertoimien $\boldsymbol{\alpha}$ pistetulo: $\boldsymbol{\alpha} \cdot \mathbf{x} = \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_p x_p = \sum_{i=1}^{p} \alpha_i x_i$.

Pääkomponenttianalyysin vaiheet:

1) Etsitään vektorille $\mathbf{x}$ sellainen lineaarikombinaatio $\boldsymbol{\alpha_1} \cdot \mathbf{x}$, joka maksimoi lausekkeen $Var(\boldsymbol{\alpha} \cdot \mathbf{x}) = \boldsymbol{\alpha} \cdot \sum \boldsymbol{\alpha}$. Yksikäsitteisen ratkaisun takaamiseksi vaaditaan lisäksi, että $\boldsymbol{\alpha}$ on rajoitettu ja tämän vuoksi määrätään rajoite $\boldsymbol{\alpha} \cdot \boldsymbol{\alpha} = 1$. Varianssin maksimoimiseksi voidaan näin ollen maksimoida funktiota $\boldsymbol{\alpha} \cdot \sum \boldsymbol{\alpha} - \lambda(\boldsymbol{\alpha} \cdot \boldsymbol{\alpha} - 1)$, jossa λ on Lagrangen kerroin.

   Differentiointi $\boldsymbol{\alpha}$:n suhteen johtaa lopulta ehtoon $(\sum - \lambda \mathbf{I}_p)\boldsymbol{\alpha} = 0$, jossa matriisi $\mathbf{I}_p$ on $p \times p$ identiteettimatriisi. Parametria λ kutsutaan korrelaatiomatriisin $\sum$ ominaisarvoksi ja vektoria $\boldsymbol{\alpha}$ vastaavaksi ominaisvektoriksi. Edellisestä kaavasta huomataan rajoite-ehdon $(\boldsymbol{\alpha} \cdot \boldsymbol{\alpha} = 1)$ avulla, että $\boldsymbol{\alpha} \cdot \sum \boldsymbol{\alpha} = \boldsymbol{\alpha} \cdot \lambda \boldsymbol{\alpha} = \lambda$, joten ominaisarvon λ täytyy olla niin suuri kuin mahdollista. Valitaan parametri λ_1 suurimmaksi kovarianssimatriisin ominaisarvoksi ja $\boldsymbol{\alpha}_1$ vastaavaksi ominaisvektoriksi. Tästä seuraa, että yhtälön $(\sum - \lambda_1 \mathbf{I}_p)\boldsymbol{\alpha}_1 = 0$ ratkaisuna ovat ensimmäisen pääkomponentin skalaarikertoimet.

2) Selitetään seuraavaksi vielä toisen pääkomponentin laskeminen. Etsitään vektorin $\mathbf{x}$ lineaarikombinaatio $\boldsymbol{\alpha_2} \cdot \mathbf{x}$, joka maksimoi ensimmäisen vaiheen tapaan lausekkeen $Var(\boldsymbol{\alpha} \cdot \mathbf{x}) = \boldsymbol{\alpha} \cdot \sum \boldsymbol{\alpha}$. Edellisessä vaiheessa esitetyn rajoituksen lisäksi vaaditaan, että uusi lineaarinen funktio on korreloimaton edellisen funktion kanssa, eli $cov(\boldsymbol{\alpha}_1 \cdot \mathbf{x}, \boldsymbol{\alpha} \cdot \mathbf{x}) = \boldsymbol{\alpha} \cdot \sum \boldsymbol{\alpha}_1 = \lambda_1 \boldsymbol{\alpha}_1 \cdot \boldsymbol{\alpha} = 0$. Näin saadaan lisäehto, jonka mukaan $\boldsymbol{\alpha}_1 \cdot \boldsymbol{\alpha} = 0$. Ominaisvektori $\boldsymbol{\alpha}_2$ löydetään yhtälöstä $\boldsymbol{\alpha} \cdot \sum \boldsymbol{\alpha} - \lambda(\boldsymbol{\alpha} \cdot \boldsymbol{\alpha} - 1) - \theta \boldsymbol{\alpha}_1 \cdot \boldsymbol{\alpha}$.

   Samaan tapaan kuin edellisessä vaiheessa, otetaan edellisestä yhtälöstä differentiaali $\boldsymbol{\alpha}$ suhteen ja merkitään se nollaksi. Tämä johtaa yhtälöön $\sum \boldsymbol{\alpha} - \lambda \boldsymbol{\alpha} - \theta \boldsymbol{\alpha}_1 = 0$. Kerrotaan tämä yhtälö vielä vasemmalta puolelta vektorilla $\boldsymbol{\alpha}_1$, jolloin saadaan $\boldsymbol{\alpha}_1 \cdot \sum \boldsymbol{\alpha} - \lambda \boldsymbol{\alpha}_1 \cdot \boldsymbol{\alpha} - \theta \boldsymbol{\alpha}_1 \cdot \boldsymbol{\alpha}_1 = 0$. Ensimmäiset termit ovat kovarianssiehdon mukaan nollia, joten saadaan, että $\theta = 0$. Tämän mukaan ratkaistavaksi yhtälöksi jää $\sum \boldsymbol{\alpha} - \lambda \boldsymbol{\alpha} = (\sum - \lambda \mathbf{I}_p)\boldsymbol{\alpha} = 0$, joka on toisen pääkomponentin ratkaistavaksi tarvittava ominaisarvoyhtälö. Samalla tavalla kuin aiemmin saadaan valittua toiseksi suurin ominaisarvo ja sitä kautta toisen pääkomponentin skalaarikertoimet. Ominaisarvo λ_2 ei voi olla sama kuin λ_1, koska muutoin ehto $\boldsymbol{\alpha}_1 \cdot \boldsymbol{\alpha}_2 = 0$ ei pitäisi paikkaansa.

3) Toistetaan edellinen kohta lopuillekin pääkomponenteille, niin että laskettavat funktiot ovat aina korreloimattomia edellisten funktioiden kanssa. Tämä voidaan toistaa aina p:hen pääkomponenttiin asti. Eli $Var(\boldsymbol{\alpha}_i \cdot \mathbf{x}) = \lambda_i$, kun $i = 1, 2, ..., p$. Ensimmäinen pääkomponentti selittää eniten aineiston vaihtelusta ja kaikki seuraavat pääkomponentit selittävät aina edellistä pääkomponenttia vähemmän. Tarkoituksena on, että mahdollisimman vähillä pääkomponenteilla saadaan selitettyä mahdollisimman paljon aineiston vaihtelusta.

3.4 Regressiomallit

3.4.1 Lineaarinen regressioanalyysi

Helpoin ja ehkä tunnetuin regressiomalli on lineaarinen regressio, jossa estimoidaan jatkuvan (vähintään välimatka-asteikollisen) vastemuuttujan ja selittävien muuttujien

lineaarista yhteyttä (Hastie et al. 2009). Analyysilla voidaan tarkastella selittävien muuttujien voimakkuutta ja merkitystä suhteessa vastemuuttujaan. Ensisijaisesti tässä työssä keskitytään havaitsemaan, mitkä selittävistä muuttujista vaikuttavat vastemuuttujan arvoihin. Lineaarista regressiomallia voidaan käyttää myös suorana ennusteena, jolloin tietyillä selittävien muuttujien arvoilla pyritään ennustamaan vastemuuttujan odotusarvoa.

Lineaarisen regressiomallin yhtälö on

$$y_i = \alpha_1 x_{i1} + \alpha_2 x_{i2} + ... + \alpha_p x_{ip} + \beta + e_i,$$

jossa yhtälön oikealla puolella ovat i:nnen havainnon selittävät muuttujat $\mathbf{x}_i$, niiden regressiokertoimet $\boldsymbol{\alpha}_i$, jäännöstermi (residuaali) e_i ja mallin vakiotermi β. Parametri p on selittävien muuttujien määrä mallissa. Yhtälön vasemmalla puolella on i:nnen vastemuuttujan havainto y_i, $i = 1, ..., n$, jossa n on otoksen koko.

Lineaarisen regressiomallin yhtälön osa eli prediktori $\alpha_1 x_{i1} + \alpha_2 x_{i2} + ... + \alpha_p x_{ip} + \beta$ on vastemuuttujan ennustettu arvo $\widehat{y}_i$. Näin ollen i:nnen jäännöstermin kaava voidaan kirjoittaa muodossa $e_i = y_i - \widehat{y}_i$ (Weisberg 2005). Toisin sanoen jäännöstermit ovat regressiomallissa havaittujen arvojen ja ennustettujen arvojen erotuksia.

Keskineliövirheen (MSE, *Mean Squared Error*) kaava on $MSE = (1/n) \cdot \sum_{i=1}^{n} e_i^2$. Lineaarisessa regressioanalyysissa halutaan minimoida keskineliövirhettä, minkä perusteella myös selittävien muuttujien estimaatit valitaan.

Estimoidut jäännöstermit vaikuttavat siihen, kuinka käyttökelpoinen malli lopulta on. Havaittujen arvojen ja ennustettujen arvojen erotuksen olessa pieni voidaan ajatella mallin sopivan hyvin kuvaamaan ennustettavaa ilmiötä. Yksi tapa arvioida mallin hyvyyttä on selityskerroin R^2. Selityskertoimella kuvataan, kuinka täydellisesti havainnot asettuvat estimoidulle regressiosuoralle. Selityskerroin lasketaan kaavalla

$$R^2 = \frac{Mallin\ selittämä\ varianssi}{Aineiston\ yhteenlaskettu\ varianssi} = \frac{n \cdot MSE}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

jossa $\bar{y}$ on havaintojen y_i $(i = 1, ..., n)$ keskiarvo.

Ennen regressiomallin soveltamista esitellään muutamia oletuksia, joiden pitävyys pitää tarkastaa analyysia tehtäessä, sekä näihin liittyviä määritelmiä. Ainakin seuraaviin asioihin on kiinnitettävä huomiota lineaarisessa regressioanalyysissä:

1) **Lineaarinen yhteys vastemuuttujan ja selittävän muuttujan välillä.** Analyysin tekemisestä ei ole hyötyä, jos valittujen selittävien muuttujien ja vastemuuttujan välillä ei ole lineaarista yhteyttä.

2) **Homoskedastisuusoletus täyttyy.** Tämä tarkoittaa, että jäännöstermeillä on sama äärellinen varianssi. Tämä näkyy jäännöstermejä tarkastellessa esimerkiksi sirontakuviossa, jossa jäännöstermien tulisi olla tasaisesti jakaantuneita suhteessa selittävien muuttujien $\mathbf{x}$ arvoihin. Samalla voidaan havaita, etteivät jäännöstermit muodosta mitään kuvionomaista muotoa, mikä voisi viitata esimerkiksi jonkin muuttujan puuttumiseen mallista.

3) **Selittävien muuttujien keskinäinen korreloimattomuus.** Multikollineaarisuudesta puhutaan, kun regressiomallissa olevat muuttujat korreloivat keskenään. Kollineaaristen muuttujien tulkinta on vaikeaa, koska eri muuttujien vaikutusta ilmiöön on hankala arvioida. Voidaan myös ajatella, että muuttuja on tullut useamman kerran huomioiduksi mallissa, jos ne korreloivat vahvasti keskenään ja antavat samansuuntaista informaatiota.

VIF-arvoja (*Variance Inflation Factor* -arvoja) käytetään havaitsemaan multikollineaarisuutta regressioanalyysissa. VIF-arvot lasketaan ns. toleranssin käänteislukuna, kun toleranssin määritellään olevan 1 – R_i^2. Termi R_i^2 on selityskerroin, jossa i:s riippumaton muuttuja estimoidaan muiden riippumattomien muuttujien perusteella. Käytännössä estimoidaan lineaarisen regressiomallin kaavassa esitettyjä selittävien muuttujien arvoja vastemuuttujan sijaan. Usein VIF-arvojen halutaan olevan ainakin alle 5 ja arvoa 10 pidetään jo kriittisenä (Franke 2010).

Regressiomallia arvioidaan oletusten täyttymisen lisäksi muillakin tavoin. Tavoitteena on mallintaa reaalimaailman tapahtumia yhtälön muodossa, mutta poikkeavat havainnot tuottavat mallintamiseen ongelmia. Regressiomallin havaintopisteet, joilla on suuri jäännöstermi tai jotka eroavat selvästi muista havaintopisteistä voivat vääristää mallin käytettävyyttä. Näitä havaintoja pyritään tunnistamaan laskemalla ns. Cookin etäisyys (Chatterjee ja Hadi 1986):

$$D_i = \frac{\sum_{j=1}^{n}(\widehat{y}_j - \widehat{y}_{j(i)})^2}{pMSE},$$

jossa

- $\widehat{y}_j$ on j:s ennustettu vastemuuttujan arvo
- $\widehat{y}_{j(i)}$ on j:s sovitettu vastemuuttujan arvo, kun mallista on poistettu i:s havainto.
- p on selittävien muuttujien määrä mallissa
- MSE on keskineliövirhe

Käytännössä Cookin etäisyys lasketaan poistamalla i:s havainto mallista ja laskemalla regressio uudestaan. Havainnoista jokainen täytyy testata, joten regressiomalli lasketaan uudestaan havaintojen lukumäärän verran niin, että jokainen havainto on kertaalleen poistettu mallista. Cookin etäisyys havainnolle i on kaikkien niiden (vastemuuttujien) muutosten summa, jotka regressiomallissa tapahtuvat poistettaessa i:s havainto.

3.4.2 LASSO

Lasso (*Least absolute shrinkage and selection operator*) on lineaarinen regressio, joka perustuu ns. *kutistamiseen* (engl. *shrinkage*). Lasso-menetelmän mukaisessa kutistamisessa aineiston muuttujien kertoimia *sakotetaan* pienemmiksi, mikä pakottaa osan kertoimista nolliksi. Vain ne muuttujat, jotka eivät kutistu nollaksi, pidetään mallissa. Tämä voi auttaa tuottamaan malleja, joilla on pienempi ennustevirhe

kuin alkuperäisellä mallilla. Käytännössä ennustevirhettä saadaan siis pienennettyä kehittämällä harhainen (pienempivarianssinen) malli kuin hyödyntämällä harhatonta (suurempivarianssista) mallia (Fonti ja Belitser 2017).

Lassossa tarkoituksena on minimoida lauseke (Tibshirani 2011)

$$\sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=1}^{p} |\alpha_j|,$$

jossa parametri $\lambda(\geq 0)$ on sakkokerroin, joka säätelee, kuinka paljon aineiston muuttujien kertoimia kutistetaan. Parametrin λ kasvaessa suurempi osa muuttujien kertoimista saa arvon nolla ja muuttujat näin ollen poistetaan mallista. Ääritapauksia ovat, kun $\lambda = 0$, jolloin yhtään muuttujaa ei poisteta mallista (eli saadaan tavallinen lineaarinen regressio), sekä (teoreettisesti), kun parametri $\lambda \to \infty$ ja kaikki kertoimet saavat arvokseen nollan. Kaavan ensimmäinen termi on jäännöstermien neliöiden summa ja toinen on virhetermi.

Tässä työssä Lassolla tuetaan päätöksiä, jotka tehdään muuttujien poistamiseksi regressiomallista. Mallin havainnot on skaalattu vähentämällä havainnoista tarkasteltavan selittävän muuttujan keskiarvo ja jakamalla tämä tarkasteltavan selittävän muuttujan keskihajonnalla. Tarkoituksena on, että kutistaminen vaikuttaa tasapuolisesti selittävien muuttujien regressiokertoimiin. Mielenkiinnon kohteena on, mitkä muuttujien kertoimista menetelmä asettaa nollaksi, vaikka lopullinen päätös malliin jätettävien muuttujien suhteen tehdään subjektiivisesti arvioiden.

3.5 Analyysiohjelmat

Menetelmien soveltamisessa aineistoon käytettiin R-ohjelmointikieltä (R versio 3.5.3). Tämän tutkielman tekstit ja ohjelmakoodit on kirjoitettu RStudio-ohjelmointiympäristössä. Myös kaikki tutkielman kuvaajat ja taulukot on tuotettu RStudio-ohjelmointiympäristössä.

4 Aineiston kuvailu ja visualisointi

Tässä luvussa tarkastellaan aineistoa kuvaajien ja taulukoiden avulla. Tarkastelut keskittyvät aiheisiin, joista on mahdollisesti hyötyä analyysia tehtäessä ja joissa kuvaajat voivat auttaa tekemään perusteltuja ratkaisuja analyysien suhteen. Taulukoiden avulla esitellään aineiston rakennetta ja opintopistemääräaineiston tunnuslukuja. Kuvaajien ja taulukoiden avulla on helpompi tarkastella, minkälaisen aineiston kanssa työskennellään, ja usein visuaalinen esitystapa auttaa hahmottamaan kokonaiskuvan paremmin kuin pelkkä sanallinen selitys.

4.1 Aineiston esittely

Aiemmissa luvuissa esiteltiin lähteet, joista tutkimuksen aineisto on kerätty. Esitellään seuraavaksi taulukoiden avulla muuttujia, joista aineisto koostuu ja mihin näitä muuttujia käytetään. Sivutaan myös aihetta, mihin kysymyksiin aineiston avulla pystytään vastaamaan.

Taulukko 1: Esimerkki opintorekisteristä. Aineisto on pseudonymisoitu niin, että se voidaan yhdistää taustatietokyselyiden kanssa. Opintorekisterin tiedoista käytetään tässä tutkimuksessa pääasiassa opintopisteiden lukumäärää.

Opiskelijanumero	Oppiaine	Suorituspaivamaara	Laajuus	Oppiaineen_laitos	Lasketaan_kokonaislaajuuteen
9a8a6e36	Tietotekniikka	2016-07-23	4	Tulevaisuuden teknologioiden laitos	1
c5a9670b	Tietotekniikka	2012-05-14	5	Tulevaisuuden teknologioiden laitos	1
3b2a907a	Tietojärjestelmätiede	2014-12-15	3	Johtamisen ja yrittäjyyden laitos	1
400c0db6	Matematiikka	2001-12-17	5	Matematiikan ja tilastotieteen laitos	1
c40c4752	Biotekniikka	2012-12-12	3	Biokemian laitos	1

4.1.1 Opintorekisteri

Kaikista opiskelijoiden suorittamista kursseista on merkinnät opintorekisterissä. Opintorekisteriin on kirjattu kurssin nimi, arvosana ja suorituspäivämäärä (usein tämä on päivämäärä, jolloin tentti on suoritettu hyväksytysti). Näiden tietojen lisäksi rekisteristä saadaan opiskelijan tunnistetiedot ja päivämäärä, milloin opiskelija on opintonsa aloittanut. Opintorekisterin kautta saadussa aineistossa ovat kaikki opiskelijoiden suoritukset tarkastelluilta vuosilta, joten myös sivuaineopintoja on näin ollen helppo seurata.

Opintorekisteriin kerättävien tietojen syy on ilmeinen: opiskelijan suoritettua kurssin on tarpeen dokumentoida kyseessä oleva suoritus. Tässä tutkimuksessa kaikista tärkein dokumentoitu tieto on opiskelijan opintopisteiden määrä, jota tässä työssä pyritään selittämään. Päivämäärä ja opiskelijan tunnistetiedot ovat tarpeellisia, jotta saadaan oikean vuoden opinnot yhdistettyä oikeaan opiskelijaan.

Taulukossa 1 on muutaman rivin otos opintorekisterin tiedoista. Taulukosta on jätetty tilanpuutteen vuoksi ja selkeyden takaamiseksi pois muutama sarake. Opintorekisteriin tallennetuista tiedoista pois jätettyjä sarakkeita ovat *Kirjoihintulopäivämäärä*, joka kertoo opiskelijan ensimmäisestä opiskeluiden aloituspäivästä tässä yliopistossa, *opintojen aloituspäivämäärä*, josta selviää opiskeluiden alkaminen tutkinnossa, ja *opintojakso*, joka kertoo tarkemmin kurssin nimen. Tiedot ovat hyödyllisiä eikä niitä ole kadotettu. Opiskelijanumerot on pseudonymisoitu merkkijonoiksi. Muualta saatuihin opiskelijoiden tietoihin on opiskelijanumero pseudonymisoitu vastaavasti samalla tavalla, joten taulukoiden yhdistäminen ei tuota jatkossa ongelmia vaan onnistuu suoraan pseudonymisoitujen opiskelijanumeroiden avulla.

4.1.2 Kyselyt

Taustatietokyselyt (TTK) on suoritettu ViLLE-järjestelmässä ja tutkimuksessa käytetyt kysymykset on suunnattu ensimmäisen vuoden tietojenkäsittelytieteiden sekä tietotekniikan opiskelijoille. Kyselyajankohta on ollut pian opintojen aloittamisen jälkeen. Kyselyyn on päässyt vastaamaan ViLLE-järjestelmässä *opinto-ohjaus* -kurssilla, joka on lähtökohtaisesti kaikille tulevaisuuden teknologioiden laitoksen uusille opiskelijoille pakollinen kurssi. Kysymykset koskevat muun muassa opiskelijoiden aiempaa koulumenestystä, vapaa-ajan viettoa sekä yleistä vireystilaa.

Taustatietokyselystä on muodostettu tämän työn selittävät muuttujat, joten analyysit perustuvat opiskelijoiden kyselyissä antamiin vastauksiin. Havainnot opiskelijoiden ajatuksista ja suunnitelmista tarjoavat mahdollisuuksia arvioida, mitkä tekijät vaikuttavat lukuvuoden aikana kerättyyn opintopistemäärään. Vastausten avulla on

Taulukko 2: Esimerkki taustatietokyselystä. Osa vastauksista on annettu avoimiin vastauskenttiin, jolloin esimerkiksi vaihteluvälille annetuista vastauksista on otettu keskiarvo.

ID	Sukupuoli	Ika	Peruskoulun_KA	Aika_Kotitehtaviin	Ohjelmointi_Kielet	Teen_Tehtavat_Heti
cf2f96a3	Mies	24	7.7	10	Java, Python	4
b7a3cc57	Nainen	20	8.6	15		5
7aa543d2	Mies	19	8,5	10	Lua	2
ce403fe8	Nainen	26	8.6	15	Java	4
4ee13e43	Nainen	19	9.8	10		4

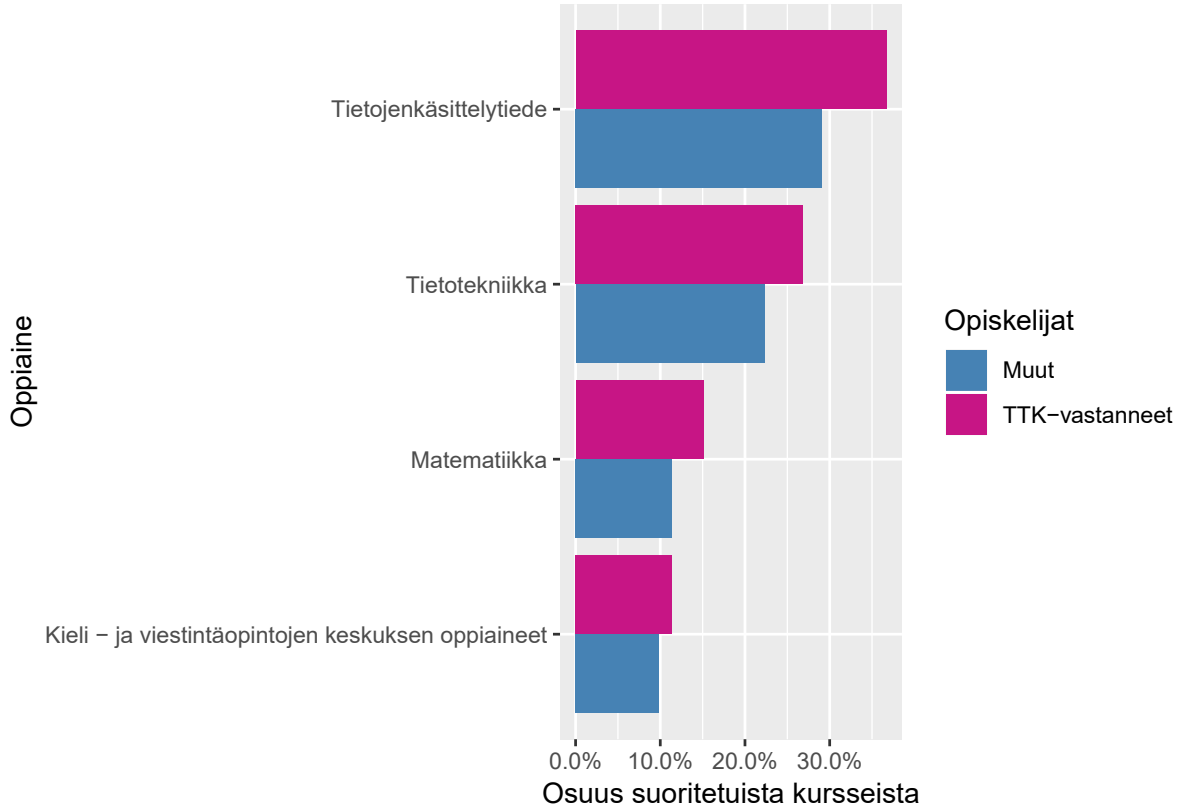
esimerkiksi mahdollista saada selville opiskelijoiden kiinnostuksen taso opiskelua kohtaan, mikä auttaa vastaamaan paremmin aiemmin esille nostettuun motivaationäkökulmaan. Kysely on ollut vapaaehtoinen jo senkin takia, että moniin kysymyksiin opiskelijoilla ei ole velvoittavaa tarvetta vastata.

Kuvaajissa ja analyyseissa keskitytään vain osaan kysymyksistä. Selkeyden vuoksi kysymyksiä pyritään myös jakamaan aiemmin mainittuihin aihepiireihin, jotta analyyseihin ei tulisi monta samanlaista kysymystä vaan samanlaisista kysymyksistä muodostettaisiin laajempia aihepiirimuuttuja. Kaikkien kysymysten käyttäminen yksittäin samanaikaisesti esimerkiksi regressiomalleissa voi johtaa helposti mallin ylisovittamiseen, joten lähtökohtana on pyritty pitämään malliin otettavien muuttujien määrä kohtalaisena. Vielä tärkeämpänä syynä aihepiirien muodostamiselle on yksittäisten kysymysten epävarmuus. Yksittäiset kysymykset kattavat vain pienen osan opiskelijan ajatuksista, mutta jakamalla kysymyksiä saadaan kattavampi yleiskuva. Esimerkiksi vastaukset väitteisiin *Haen muualle opiskelemaan* ja *Ala kiinnostaa* kertovat yhdessä paremmin opiskeluiden oletetusta etenemisestä lukuvuoden aikana kuin vastauksien tarkastelu yksittäin.

Taustatietokyselyistä kerätty aineisto ei ole yhtä täydellistä kuin opintorekisterin aineisto. Muutamat avoimet vastauskentät kuten harjoitus- tai kotitehtäviin käytetty aika ovat useammallakin tavalla epävarmoja, koska arvio perustuu opiskelijan omaan kokemukseen ja kirjaamisessa on opiskelijoiden välillä selkeitä eroja. Tällaiset kysymykset eivät ole kuitenkaan pääosassa analyyseja, vaan on keskitytty kysymyksiin, joissa vastausvaihtoehdot ovat Likert-asteikolla 1–5 (*Täysin eri mieltä – Täysin samaa mieltä*). Taulukossa 2 on esimerkkejä siitä, mitä opiskelijoilta kysyttiin, ja taulukosta näkyy myös, missä muodossa vastaukset on annettu. Jos vastauksiin on annettu vaihteluväli, esimerkiksi kotitehtäviin käytetystä ajasta, on paremman tiedon puutteessa käytetty keskiarvoa opiskelijan antamasta arviosta.

4.2 Aineiston alustava tarkastelu

Aiemmin todettiin, etteivät taustatietokyselyihin vastanneet opiskelijat kata kaikkia aloittaneita opiskelijoita. Vaikka kyselyihin on vastannut valtaosa aloittaneista opiskelijoista, tarkastellaan seuraavaksi ryhmien (vastanneet ja vastaamatta jättäneet) eroja opintopisteissä. Ryhmien vertailuun ei ole opintorekisterin tietojen lisäksi toista mittaria tässä työssä saatavilla. Taustatietokyselyyn vastanneita opiskelijoita on kahdelta lukuvuodelta yhteensä 142 ja vastaamatta jättäneitä opiskelijoita 48. Nämä kaksi ryhmää kattavat oletettavasti kaikki laitoksella aloittaneet opiskelijat lukuvuosilta 2015—16 ja 2016—17. Poikkeuksen voivat tehdä opiskelijat, jotka ovat myöhemmässä



Kuva 2: Yleisimpien oppiaineiden jakauma taustatietokyselyyn (TTK) vastanneissa ja vastaamatta jättäneissä. Pylväät kuvaavat taustatietokyselyyn vastanneita (violetti) ja kaikkia muita (sininen) laitoksella alottaineita opiskelijoita. Yleisimmät oppiaineet ovat samat, vaikka oppiaineiden osuudet eivät olekaan täysin samoin jakautuneet taustatietokyselyyn vastanneiden ja vastaamatta jättäneiden opiskelijoiden kesken.

opiskelujen vaiheessa osallistuneet opinto-ohjauskurssille ja vastanneet kyselyyn, mutta lähtökohtaisesti kaikki kyselyyn vastanneet olivat suorittamassa ensimmäistä opiskeluvuottaan tulevaisuuden teknologioiden laitoksella. Kaikki tässä työssä mukaan otetut opiskelijat ovat osallistuneet opinto-ohjauskurssille ja opiskelijat on jaoteltu kahteen ryhmään ainoastaan sen perusteella ovatko he vastanneet kyselyyn vai eivät.

Kuvaan 2 on piirretty yleisimmät oppiaineet taustatietokyselyyn vastaamatta jättäneiden opiskelijoiden ja vastanneiden opiskelijoiden suhteen. Kyselyyn vastanneet opiskelijat ovat suorittaneet enemmän tulevaisuuden teknologioiden laitokselle tyypillisiä oppiaineita kuin vastaamatta jättäneet. Kummassakin ryhmässä suosituimmat oppiaineet ovat kuitenkin samat ja jakaumat hyvin samankaltaiset.

Taulukon 3 mukaan varsinkin ryhmien lukuvuoden opintopistekeskiarvot eroavat toisistaan huomattavasti. Taustatietokyselyyn vastanneet opiskelijat suorittivat keskimäärin yli 10 opintopistettä enemmän kuin muut laitoksella aloittaneet opiskelijat. Aineiston tarkastelua jatketaan ottamalla mukaan vain taustatietokyselyyn vastanneet opiskelijat, koska vastaamatta jättäneistä opiskelijoista on saatavilla vain opintorekisterin tietoja.

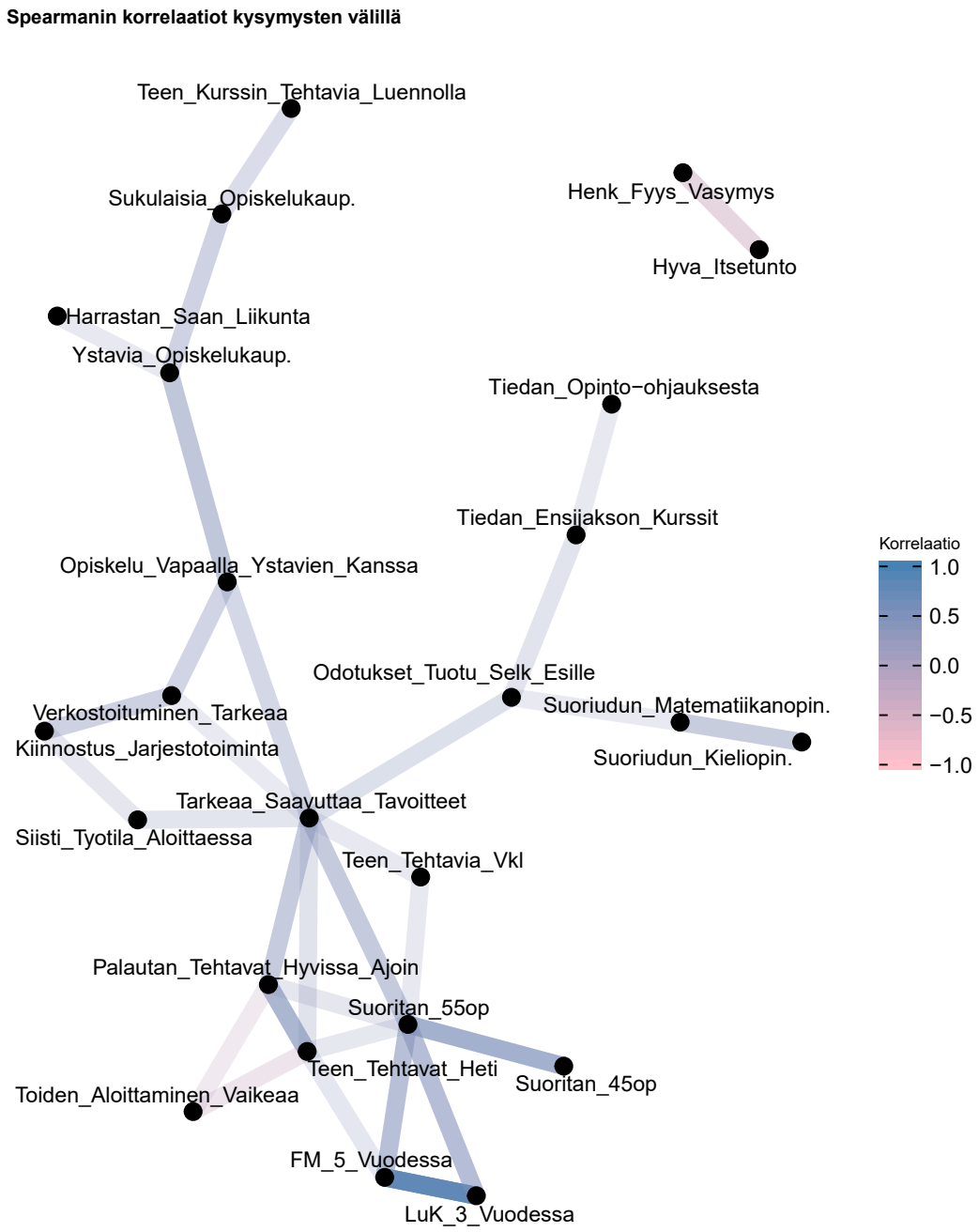
Seuraavissa kuvaajissa keskitytään tarkastelemaan taustatietokyselyiden vastauksien yhteyksiä. Kuva 3 esittää opiskelijoiden vastausten perusteella tehtyä verkostoa, joka

Taulukko 3: Opintopistemäärien keskiarvot ja hajonnat taustatietokyselyyn (TTK) vastanneissa ja vastaamatta jättäneissä.

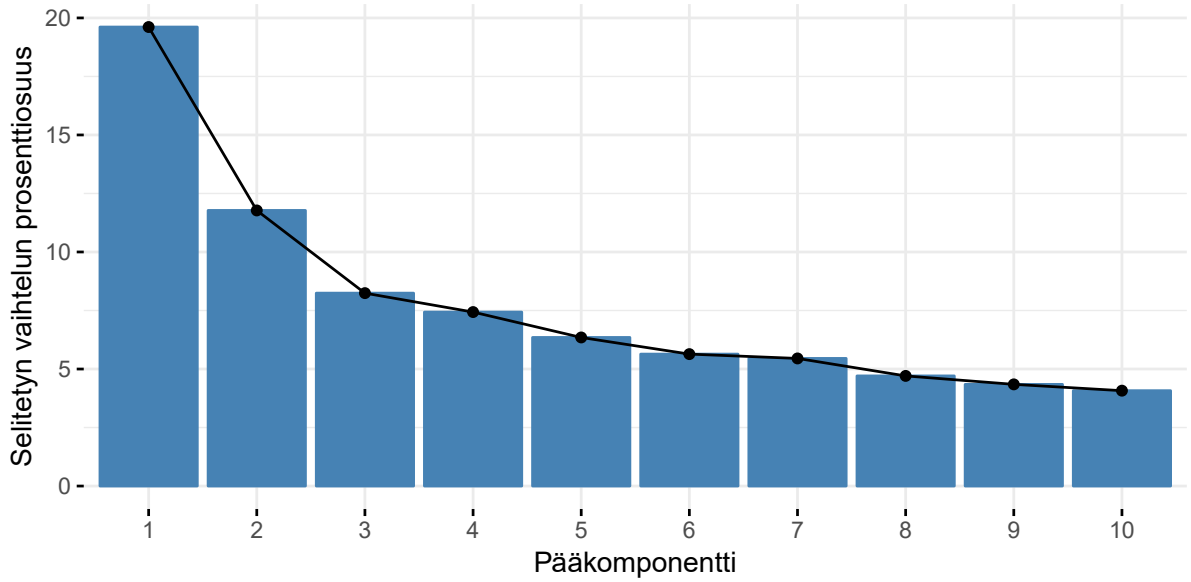
Ryhmä	N	Keskiarvo	Hajonta
Muut	48	42.6	25.0
TTK-vastan.	142	54.0	22.1

kuvaa vastausten korrelaatioita. Tähän tarkasteluun otettiin vain kysymykset, joissa opiskelijat saivat asteikolla 1–5 antaa mielipiteensä aiheen tärkeydestä, ja lopulta kuvaajaan sisällytettiin vain korrelaatiot, jotka olivat itseisarvoltaan suurempia kuin 0.3. Tarkoituksena on hahmottaa, mitkä kysymyksistä antavat samansuuntaista tietoa ja tarpeettomia säilyttää analyysivaiheessa itsenäisinä muuttujina.

Kaikki yllättävät yhteydet ovat tässä vaiheessa mielenkiintoisia. Kuvaajasta havaitaan esimerkiksi vasemmassa alalaidassa oleva ryhmä kysymyksiä, jotka koskevat opintojen aikataulua. Näissä erityisen vahvasti korreloivat tavoitteet valmistua luonnontieteiden kandidaatiksi kolmessa vuodessa ja filosofian maisteriksi viidessä vuodessa. Huomataan myös, että tavoitteiden saavuttamisen tärkeys korreloi monen muun kysymyksen kanssa. Useimmat kuvaajassa esitetyistä korrelaatioista ovat positiivisia. Vastaavanlaisia havaintoja voitaisiin tehdä muodostamalla kysymyksistä suoraan korrelaatiomatriisi, mutta kuvaaja esittää saman asian havainnollisemmassa muodossa.



Kuva 3: Taustatietokyselyn vastausten korrelaatiot esitettynä graafisessa muodossa. Vahvempi väri kuvastaa itseisarvoltaan suurempaa korrelaatiota. Kuvaajassa on otettu huomioon vain korrelaatiot, jotka ovat itseisarvoltaan yli 0.3:n vahvuisia. Erityisesti huomataan kuvaajan alalaitaan muodostuvat vahvat korrelaatiot opintojen edistymistä koskevien vastausten välillä.



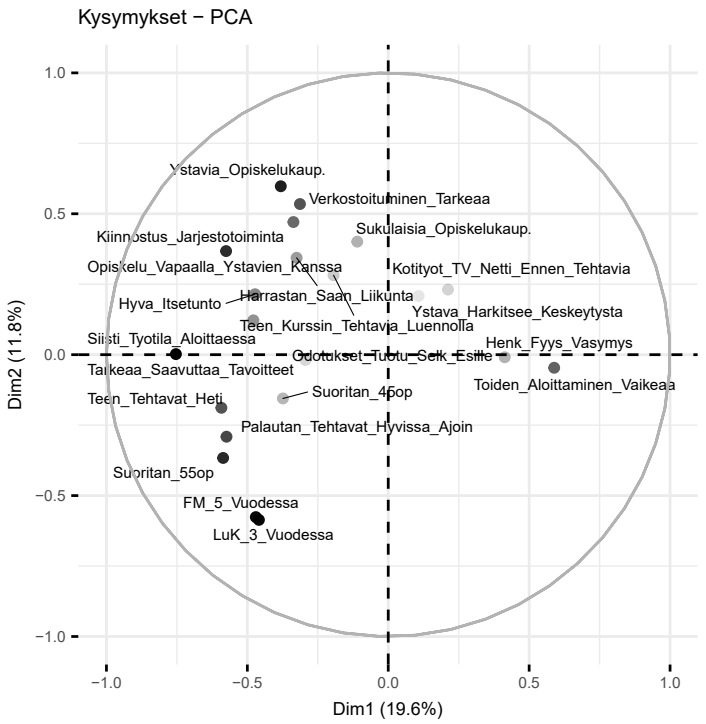
Kuva 4: Vaihtelun selittäminen pääkomponenttianalyysissa. Ensimmäiset pääkomponentit erottelevat aineistoa eniten. Kuvaajasta huomataan, että tarvitaan useita pääkomponentteja ennen kuin suurin osa vaihtelusta selitetään.

4.2.1 Kysymysten jaottelu pääkomponenttianalyysillä

Kun muuttujia on paljon ja niiden välillä voi olla suuriakin korrelaatioita, yksi analysointiin valmistava ratkaisutapa on käyttää pääkomponenttianalyysia. Yksinkertaistettuna pääkomponenttianalyysissä saamme esitettyä usean dimension (usean kysymyksen) aineiston visuaalisessa muodossa laskemalla uudet ulottuvuudet. Tarkoituksena on saada hahmotettavampi, esimerkiksi kahden pääkomponentin muodostama kuvaaja erottelemaan opiskelijoita toisistaan. Nämä pääkomponentit muodostuvat useista kysymyksistä. Tästä eteenpäin taustatietokyselyn kysymyksistä puhutaan vain muuttujina.

Kuvasta 4 nähdään, kuinka hyvin aineistosta muodostetut pääkomponentit selittävät siinä esiintyvää varianssia. Selittävyysaste ei ole kovin korkea yhdelläkään pääkomponentilla. Ensisijainen tavoite tällä pääkomponenttianalyysilla on selvittää, mitkä taustatietokyselyn muuttujista ovat vaikutukseltaan samansuuntaisia. Tässä työssä ei käydä tarkemmin, kuinka hyvin pääkomponenttianalyysilla voitaisiin erotella eri opintopistemääriä suorittaneita opiskelijoita toisistaan.

Kuvassa 5 on jaoteltu taustatietokyselyn muuttujia (kysymyksiä) pääkomponenttianalyysin avulla. Kuvaajan akseleiden prosenttiluvuista nähdään kahden ensimmäisen pääkomponentin selittämä vaihtelu opiskelijoiden vastauksissa. Muuttujien sijoittuminen kuvaajassa esittää, mihin kysymyksistä on vastattu samansuuntaisesti. Ensimmäinen havainto voidaan tehdä y-akselin jakamien muuttujien välillä. Kuvaajassa vasemmalla puolella olevilla muuttujilla on positiivisempi konnotaatio kuin oikealle puolelle jäävissä muuttujissa. Toinen huomio ei ole aivan yhtä selvä, mutta kuitenkin nähtävissä: Kuvaajan vasempaan yläsektoriin sijoittuu muuttujia, jotka viittaavat sosiaaliseen toimintaan. Vastaavasti vasemmassa alasektorissa on tavoitteellisia muuttujia. Tavoitteelliset muuttujat liittyvät opintojen etenemiseen, joka vaatii suunnitelmallisuutta



Kuva 5: Taustatietokyselyn muuttujien sijoittuminen pääkomponenttianalyysissa. Kuvaajan akselit muodostuvat kahdesta tärkeimmästä pääkomponentista, joilla saavutetaan eniten erottelevuutta aineistossa. Yksittäisen muuttujan (ympyräkiekon) sijainti kuvaa muuttujan projektiopistettä kahdelle pääkomponentille. Muuttujan tummuus kuvaa muuttujan osuutta siitä vaihtelusta, mitä kuvatut pääkomponentit selittävät. Tummaksi värjätyn muuttujan osuus on esitetyissä pääkomponenteissa suurempi kuin haaleaksi värjätyn muuttujan.

ja itsekuria.

Kaikilla muuttujilla ei ole samanlaista painoarvoa pääkomponentteja muodostettaessa, ja tätä on kuvattu muuttujien pallojen tummuudella. Tumma täyte viittaa muuttujan osuutta siitä vaihtelusta, mikä pääkomponentilla pystytään selittämään. Kaikkia muuttujia kuvaajaan ei ole otettu mukaan, koska on ajateltu kuvaajan mahdollisimman selkeää luettavuutta. Huolimatta siitä, että kaikkia muuttujia ei ole tarkasteltu, kuvaajasta saadaan arvokasta tietoa muuttujien jakamiseksi aihepiireihin (ks. luku 6).

5 Luokittelu- ja regressiomallien soveltaminen aineistoon

Tässä luvussa muodostetaan ensin pääkomponenttianalyysin tuloksia apuna käyttäen uudet aihepiirimuuttujat. Tämän jälkeen käytetään aiemmissa luvuissa esiteltyjä menetelmiä aineiston analysointiin. Klusteroinnissa käytetään vain aihepiirimuuttujia. Regressioanalyysissa aihepiirimuuttujien lisäksi otetaan muun muassa opiskelijoiden ajankäytöstä kertovia muuttujia ja satunnaismetsän mallissa käytetään aihepiirimuuttujien kanssa kaikkia taustatietokyselyn muuttujia. Jaotellaan tautatietokyselyn muuttujia sopiviin aihepiireihin seuraavalla tavalla:

1) Verkostot: Aihepiiri muodostuu muuttujista, jotka mittaavat opiskelijoiden kiinnostusta viettää aikaa kanssaopiskelijoiden seurassa.

2) Taito: Aihepiiri muodostuu muuttujista, jotka kertovat opiskelijan taidoista ja koulumenestyksestä sekä kiinnostuksesta opiskeltavaa alaa kohtaan.

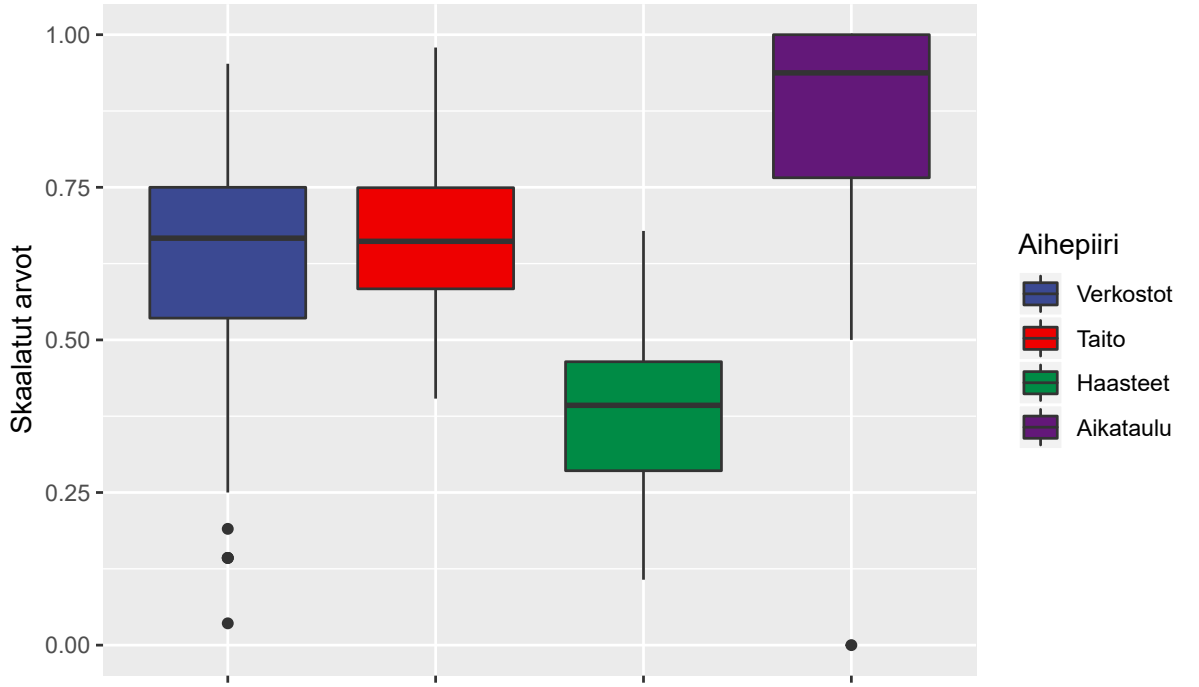
3) Haasteet: Aihepiiri muodostuu ongelmia ja (opintoja haittaavia) suunnitelmia mittaavista muuttujista, jotka voivat vaikuttaa negatiivisesti opiskelujen etenemiseen.

4) Aikataulu: Aihepiiri muodostuu muuttujista, jotka kuvaavat opiskelijan omia arvioita valmistumisajoista ja opintopistemääristä.

Kysymysten luokittelu perustui subjektiiviseen arviointiin ja pääkomponenttianalyysilla saatuihin havaintoihin pääkomponenttien sisältämien muuttujien osuuksista. Muutama aihealue erottui kysymyksissä selkeimmin, minkä perusteella aihepiirejä muodostettiin lopulta neljä kappaletta. Aihepiirit on luotu 4–13 muuttujasta, jotka vastaavat aihepiirin kuvausta. Aihepiirit on skaalattu välille [0,1]. Pääkomponenttianalyysin tuloksia ei sovellettu suoraan, koska pelkästään muuttujien jakaminen aihepiireihin taustatietokyselyn vastauksiin perustuen ei ole tarkoituksenmukaista.

Kuvassa 6 on laatikko-janakuviot opiskelijoiden vastauksista muodostettuihin aihepiirimuuttujiin. Aihepiiri *Aikataulu* on muodostettu vain neljästä kysymyksestä, ja laatikko-janakuvion perusteella suurin osan vastanneista opiskelijoista suhtautuu luottavaisesti opiskelujen etenemiseen. Havaitaan kuitenkin yksi opiskelija, jolla ei ole aikataulullisia tavoitteita kyselyn perusteella. Muut aihepiirimuuttujat ovat jakaantuneet tasaisemmin lähemmäs skaalan puolta väliä, eikä samanlaisia poikkeavia havaintoja ole nähtävissä.

Vertaillaan aluksi, miten valitut aihepiirit korreloivat opintopisteiden kanssa. Korrelaatio on helppo ja hyvä merkki siitä, kannattaako aihepiiriä edes ottaa lähempään tarkasteluun, mutta toisaalta korrelaatio ei vielä takaa perusteltua relaatiota muuttujien välille. Taulukko 4 esittää kaikkien neljän valitun aihepiirin ja opintopisteiden korrelaatiot. Korrelaatioiden tilastolliseen merkitsevyyteen ei ole kiinnitetty huomiota. Korrelaatiot toimivat suuntaa antavina arvoina muuttujien välisistä suhteista eivätkä itsessään ole tässä työssä kiinnostavia.

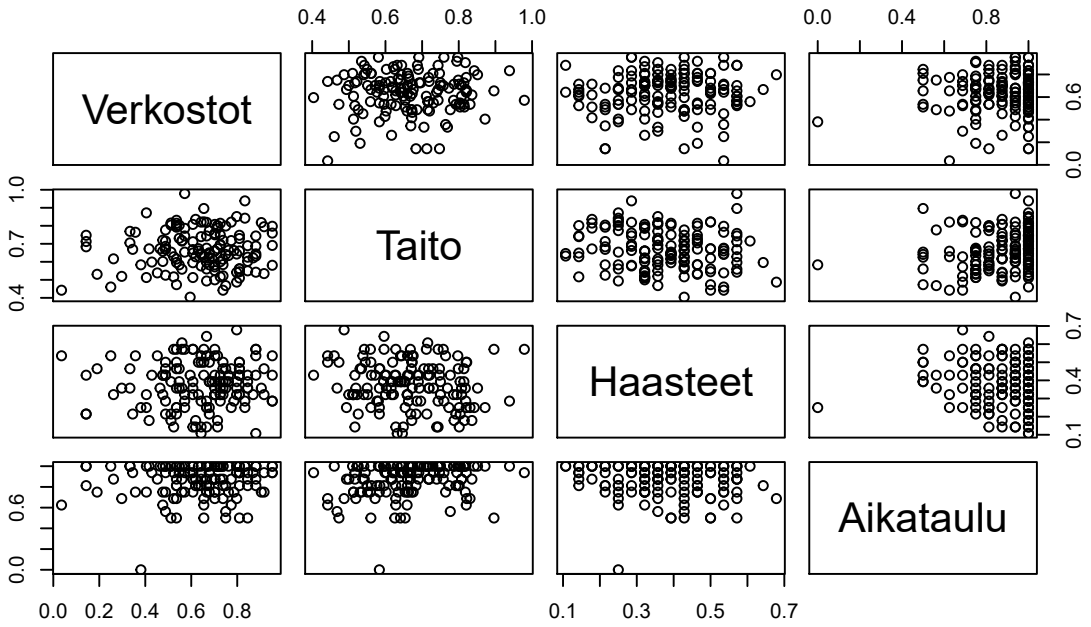


Kuva 6: Neljän aihepiirimuuttujan jakaumat. Kunkin laatikon poikkiviiva on mediaani, ja ala- ja yläreunat ovat 1. ja 3. kvartiilin kohdalla. Aikataulu-aihepiirissä vastaukset ovat keskittyneet skaalan yläpäähän yhtä havaintoa lukuunottamatta. Muissa aihepiireissä muuttujan arvot ovat jakaantuneet tasaisemmin keskemmälle skaalaa.

Taulukko 4: Neljän aihepiirimuuttujan korrelaatiot opintopisteiden kanssa. Kolme aihepiirimuuttujaa korreloi opintopisteiden kanssa ja ainoastaan muuttujan Verkostot kohdalla korrelaatiota ei havaita juuri lainkaan.

	Korrelaatio
Verkostot	0.05
Taito	0.28
Haasteet	-0.23
Aikataulu	0.39

Aihepiirien korrelaatiot opintopisteisiin ovat tässä työssä mielenkiintoisin korrelaatioiden vertailu, mutta tarkastellaan myös aihepiirien keskinäisiä korrelaatioita. Ensinnäkin, jos aihepiirit korreloivat voimakkaasti keskenään, on syytä miettiä, mistä tämä johtuu, ja kysyä olisiko perusteltua yhdistää korreloivat aihepiirit yhdeksi kattavammaksi aihepiiriksi. Toisekseen yllättävät korrelaatiot saattavat herättää pohdintaa, onko muuttujien jako aihepiireihin ollut mielekäs. Kolmanneksi korrelaatioden kartoittaminen jo tässä vaiheessa on ennaltaehkäisevä keino multikollineaarisuudelle, joka saattaisi tuottaa regressioanalyysissä ongelmia. Kuva 7 esittää aihepiirimuuttujien keskinäisiä korrelaatioita. Korrelaatioissa ei ole tämän perusteella mitään erityistä huomioitavaa.



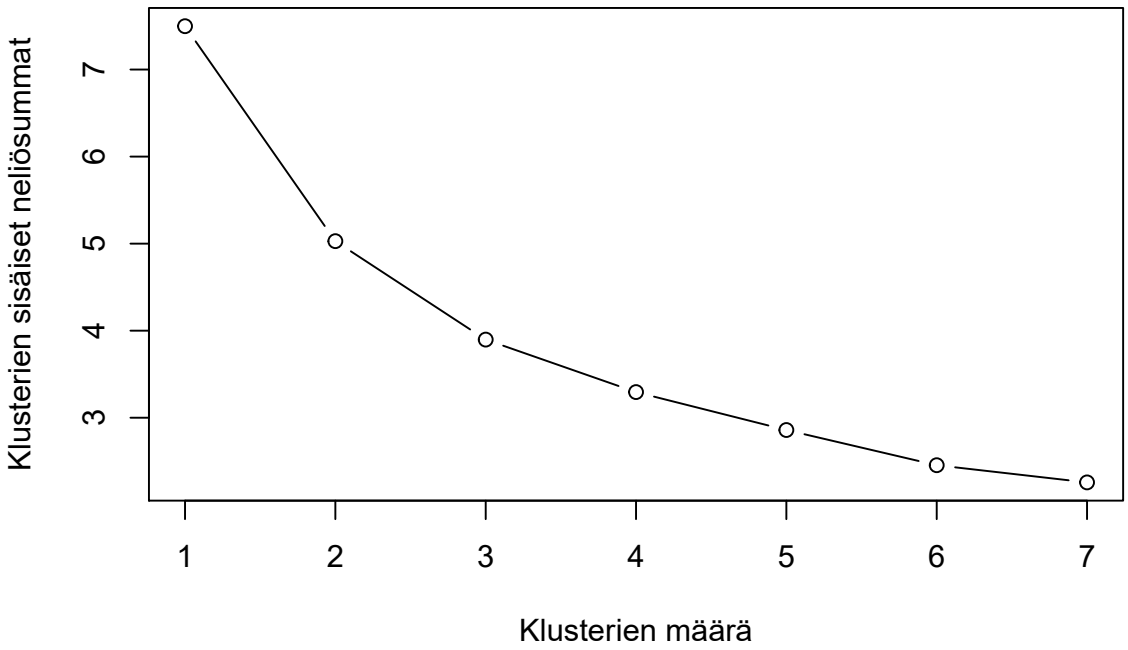
Kuva 7: Aihepiirimuuttujien keskinäiset korrelaatiot. Lineaarista yhteyttä ei ole kuvaajan perusteella nähtävissä minkään aihepiirimuuttujaparin kesken. Tämän perusteella aihepiirien voidaan katsoa olevan itsenäisiä muuttujia eli ne eivät mittaa samoja asioita.

5.1 Klusterointi

Tarkastellaan opintopistekertymiä muodostamalla opiskelijoista klustereita valittujen aihepiirien perusteella. Aihepiiri *Verkostot* korreloi vain heikosti opintopisteiden kanssa, joten tätä aihepiiriä ei oteta klusterointiin mukaan. Klusterointiin käytetään kolmea muuta aihepiiriä. Klusteroinnissa käytetään aiemmin esiteltyä *Fuzzy C Means* -klusterointia.

Ensimmäinen tehtävä on klusterien lukumäärän päättäminen, joka tehdään kuvaajan perusteella. Kuvassa 8 on vaaka-akselilla klusterien kokeiltu lukumäärä ja pystyakselilla klusterien sisäinen varianssi. Klusterien lukumääräksi valitaan yleensä se kohta, jossa murtoviiva muodostaa selkeimmän taitoskohdan (Kodinariya & Makwana 2013). Tässä työssä valittujen klustereiden lukumäärä haluttiin pitää pienenä, jotta opintopistekeskiarvojen tilastolliseen tarkasteluun jää jokaisessa klusterissa riittävästi opiskelijoita. Halutaan kuitenkin klusterien määrää valittaessa, että klustereihin jää klusterin sisällä keskenään samanlaisia opiskelijoita. Näiden perusteluiden pohjalta klustereiden määräksi valitaan kolme. Myös kaksi klusteria olisi ollut perusteltu valinta, mutta kolmella klusterilla saadaan opiskelijoita eroteltua toisistaan vielä ilman, että klusterien opiskelijamäärät pienenisivät ongelmallisiksi.

Tarkastellaan seuraavaksi, miten muodostettujen klustereiden opiskelijat vertautuvat valittujen aihepiirimuuttujien suhteen. Kuvassa 9 yksi ympyrä vastaa yhtä opiskelijaa sen mukaan, mikä on opiskelijan arvo kyseisessä aihepiirimuuttujassa. Ympyrän väri kuvaa, mihin klusteriin opiskelija kuuluu. Opiskelijat on sijoitettu siihen klusteriin, mistä ne ovat klusterointimenetelmällä saaneet suurimman todennäköisyyden kuulua. Aihepiirit



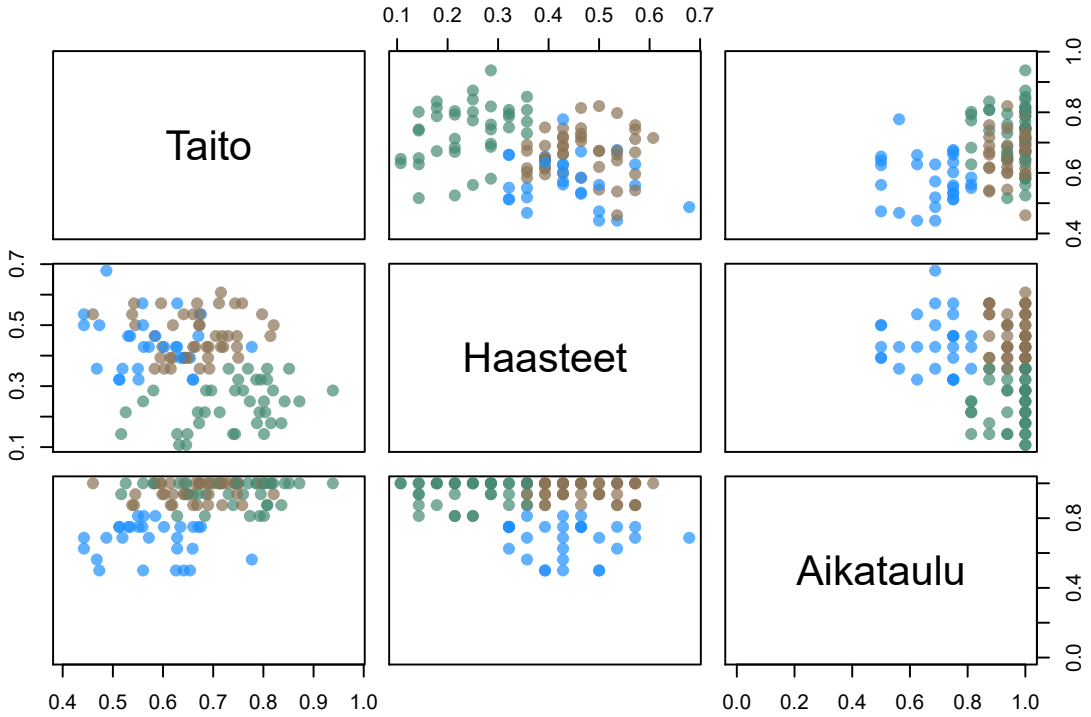
Kuva 8: Klusterien sisäisen varianssin riippuvuus klusterien lukumäärästä. Vaaka-akselilla on esitettynä klusterien kokeiltu lukumäärä ja pystyakselilla summat neliöidyistä havaintojen poikkeavuuksista Fuzzy C Means -klusteroinnilla määrätyistä klusterikeskuksista. Kuvaajan murtoviivan perusteella päätetään valittujen klusterien lukumäärä. Hyväksi klusterien määräksi voidaan ajatella kohtaa, jossa murtoviivassa näkyy selvä taitos. Tämä perustuu siihen, ettei klusterikeskusten lisäämisestä taitoskohdan jälkeen saada merkittävää parannusta klusterien varianssien pienentämiseksi

Aikataulu ja *Haasteet* eroavat selkeästi toisistaan. Näitä aihepiirejä verrattaessa *Taito*-aihepiiriin nähdään yhden klusterin opiskelijoiden erottuvan selkeämmin kahden muun klusterin opiskelijoista.

Klusterit on luotu kolmen aihepiirimuuttujan perusteella ja klustereiden opintopistekeskiarvot on laskettu klusterien muodostamisen jälkeen. Klusterikohtaisia tunnuslukuja on esitetty taulukossa 5. Muuttujien *Taito* ja *Aikataulu* keskiarvot ovat pienimpiä ensimmäisessä klusterissa. Ensimmäinen klusteri on myös opintopistekeskiarvoltaan pienin. Toisessa ja kolmannessa klusterissa opintopistekeskiarvot ovat melkein samat ja käytännössä eroa nähdään ainostaan muuttujan *Haasteet* keskiarvoissa.

5.1.1 Klusterikohtaisten opintopistekeskiarvojen luottamusvälit

Edellä laskettujen opintopistekeskiarvojen epävarmuuden estimoimiseksi käytetään BC_a-bootstrapia. BC_a-bootstrapilla lasketaan jokaisen klusterin opintopistekeskiarvoille luottamusvälit.



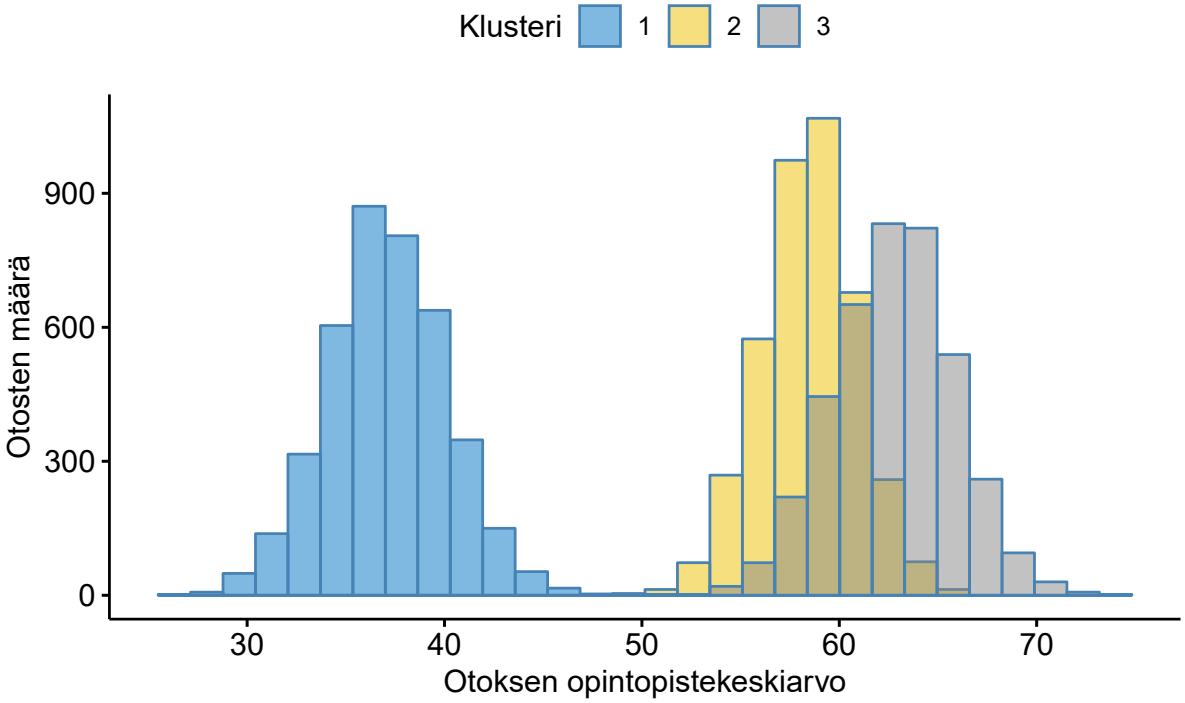
Kuva 9: Opiskelijat on jaettu kolmeen klusteriin sen perusteella, mihin klusteriin opiskelijoilla on suurin todennäköisyys kuulua. Klusterit erottaa toisistaan värien perusteella. Yksi opiskelija on sijoitettu kuvaajaan aina opiskelijan kahden eri aihepiirimuuttujan arvon perusteella. Klusterit on muodostettu kuvassa esitettyjen kolmen aihepiirin mukaan, joten on luonnollista, että kuvaajassakin klusterit erottuvat toisistaan.

Taulukko 5: Aihepiirimuuttujien ja opintopisteiden klusterikohtaiset keskiarvot. Klusterit tehtiin Fuzzy C Means -algoritmilla aihepiirimuuttujien perusteella. Opintopisteitä ei otettu huomioon klustereiden muodostamisvaiheessa, mutta myös opintopistekeskiarvoissa nähdään selkeitä eroja klustereiden välillä. Sarake N kertoo opiskelijoiden määrän klusterissa.

Klusteri	N	Taito	Haasteet	Aikataulu	Opintopisteet
1	39	0.60	0.42	0.69	37.10
2	57	0.67	0.44	0.94	58.54
3	46	0.73	0.27	0.93	62.61

Kuva 10 esittää opintopisteiden klusterikohtaisten keskiarvojen bootstrap-otoksia. Havaitaan, että klusterien välillä on joitakin eroja. Taulukko 6 esittää bootstrap-otosten perusteella lasketut opintopisteiden odotusarvojen klusterikohtaiset luottamusvälit. Ensimmäisen klusterin luottamusvälin yläraja ja kahden muun klusterin alaraja poikkeavat arviolta 10 opintopisteen verran 95%:n luottamustasolla.

Klusteroinnissa käytettiin vain aihepiirimuuttujia. Aihepiirimuuttujiin ei sisällytetty kysymyksiä, joihin vastattiin avoimella vastauskentällä. Esimerkiksi laskuharjoituksiin käytetyllä ajalla ei ole käytännöllistä ylärajaa, ja skaalaaminen muiden kysymysten



Kuva 10: Opintopisteiden klusterikohtaisten keskiarvojen bootstrap-otokset. Kuvaajaa varten on otettu jokaisesta klusterista 4000 bootstrap-otosta opiskelijoiden keräämistä opintopisteistä. Näille jokaisen klusterin 4000 otokselle on erikseen laskettu opintopistekeskiarvot, joiden perusteella histogrammit on piirretty kuvaajaan.

Taulukko 6: Kolmesta aihepiiristä muodostettujen klustereiden opintopisteiden odotusarvojen keskivirheet ja luottamusvälit. Toisen ja kolmannen klusterin luottamusvälit eivät risteä ensimmäisen klusterin luottamusvälin kanssa, joten ensimmäinen klusterin opintopisteiden odotusarvo eroaa tilastollisesti kahden muun klusterin odotusarvoista.

Klusteri	Keskivirhe	LV_Alaraja	LV_Yläraja
1	3.02	31.28	43.18
2	2.43	53.81	63.40
3	3.05	56.44	68.37

kanssa ei ole yksiselitteistä. Regressiomalleissa tämä kokoluokkaongelma ei ilmene samoin, koska kokoluokkaerot tasoittuvat kertoimien mukana. Malliin voidaan tuoda näin ollen lisää muuttujia.

Taulukko 7: Lineaarisen regressiomallin kertoimien piste-estimaatit arvioitaessa muuttujien vaikutusta opintopistekertymään. Taulukon vasemmalla puolella ovat mallissa olevien muuttujien nimet. Estimaatit kuvaavat, kuinka paljon yhden yksikön lisääminen vaikuttaisi opiskelijan opintopistekertymän määrän. Keskivirhe ja luottamusväli kuvaavat estimaatin epävarmuutta. Suurimmasta osasta mallin muuttujia ei voida piste-estimaatin luottamusvälin perusteella sanoa vaikuttaako muuttujan arvon kasvattaminen opintopistemääriin negatiivisesti vai positiivisesti. Laskuharjoituksiin (demoihin), kotitehtäviin ja tietokoneella pelaamiseen käytetyt ajat ovat tuntimääriä viikossa, kun taas nukkumiseen käytetty aika on tuntimäärä vuorokaudessa. Työpaikkaa kysyvä muuttuja on binäärinen ja ikä on kysytty vuosissa.

	Estimaatti	Keskivirhe	95% luottamusväli
Vakio	-29.35	28.71	[-83.1, 32.7]
Ikä	0.32	0.39	[-0.7, 1]
Aika_Kotitehtaviin	0.08	0.37	[-0.7, 0.7]
TyopaikkaKylla	-0.38	5.38	[-12.9, 13.1]
Aika_Luennoilla	0.18	0.24	[-0.4, 0.6]
Aika_Pelaaminen_TiKo	-0.04	0.14	[-0.3, 0.2]
Aika_Uni	3.22	2.31	[-1.2, 7.9]
Aika_Demoihin	-1.01	0.61	[-2.1, 0.1]
Verkostot	0.69	9.50	[-17, 21.2]
Taito	31.09	15.95	[2.1, 64.7]
Haasteet	-21.42	14.63	[-49.7, 5.2]
Aikataulu	48.77	11.37	[19.5, 76.3]

5.2 Lineaarinen regressioanalyysi

Klusteroinnissa tarkasteltiin aihepiirimuuttujia, joista muodostettujen klusterien opintopistekeskiarvoissa huomattiin eroja. Aihepiirit erottelivat opiskelijoita ja korreloivat opintopisteiden kanssa, joten ei ole perusteltua jättää niitä pois regressiomalleistakaan. Regressiomalliin voidaan ottaa myös muuttujia, jotka skaalausongelman vuoksi jätettiin pois klusteroinnista. Malliin tuodaan uusia muuttujia, joista suurin osa koskee opiskelijoiden ajankäyttöä. Mallin uudet muuttujat ja aihepiirimuuttujat toimivat selittävinä muuttujina, kun rakennetaan lineaarinen regressiomalli, jonka vastemuuttujana on opintopistemäärä. Mallin tuloksia kuvataan taulukossa 7.

Ajankäytöstä kertovien muuttujien kertoimet ovat yleisesti hyvin lähellä nollaa. Varsinkin kotitehtäviin ja tietokoneella pelaamiseen käytetyt ajat ovat kertoimiltaan hyvin pieniä. Mallin selkeyden vuoksi on suotavaa poistaa ylimääräisiä muuttujia mallista. Tähän on olemassa useampia menetelmävaihtoehtoja, mutta tässä työssä käytetään aiemmin esiteltyä Lasso-regressiota. Tällä menetelmällä saadaan poistettua muuttujia kokonaan mallista pienentämällä osa muuttujien kertoimista nollaksi.

Lasso-regressiossa suurin osa taulukon 7 muuttujien kertoimista estimoidaan nollaksi eli ne jätetään kokonaan pois mallista. Aihepiirimuuttujien lisäksi malliin jäävät muuttujat laskuharjoituksiin sekä nukkumiseen käytetyistä ajoista. Poistetut muuttujat eivät vaikuta opintopisteiden osalta merkittäviltä, joten jatketaan analyysia Lasso-regressioon jääneillä muuttujilla. Taulukossa 8 ovat jäljelle jääneiden muuttujien piste-estimaatit, keskivirheet ja luottamuvälit.

Taulukon 8 regressiomallin selityskerroin R^2 on 0.25, joka kertoo, kuinka paljon valitut muuttujat selittävät kerättyjen opintopisteiden vaihtelusta. Taulukossa 8 esitetyt

Taulukko 8: Lasso-regression mallissa muuttujia on karsittu tavalliseen lineaariseen malliin verrattuna. Tämä auttaa määrittämään muuttujien todellista vaikutusta, kun vähemmän tärkeitä muuttujia on poistettu sekoittamasta mallia.

	Estimaatti	Keskivirhe	95% luottamusväli
Vakio	-14.9	23.4	[-58.1, 36.4]
Aika_Demoihin	-0.9	0.5	[-1.8, 0]
Aika_Uni	2.8	2.2	[-1.5, 7.4]
Taito	31.0	15.4	[3.4, 62.8]
Haasteet	-25.7	13.8	[-51.8, -0.5]
Aikataulu	48.3	10.8	[23.9, 75.2]

luottamusvälit viittaavat, että kolmen muuttujan kohdalla voidaan luotettavasti arvioida niiden vaikutusta. Muuttujat *Taito* sekä *Aikataulu* ovat vaikutukseltaan positiivisia ja muuttuja *Haasteet* puolestaan negatiivinen.

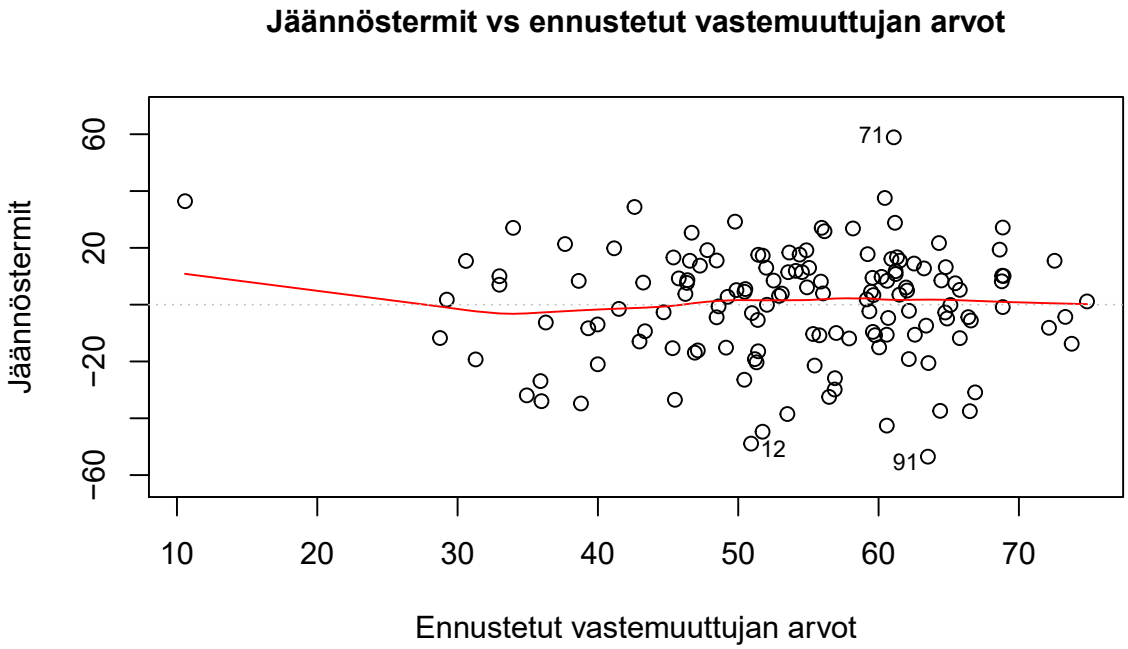
5.2.1 Lineaarisen regressiomallin oletukset

Seuraavaksi tarkastellaan lineaarisen mallin oletuksia. VIF-arvot ovat keino multikollineaarisuuden tarkasteluun. Käytännössä multikollineaarisuus tarkoittaa, että muuttuja on jollain tavalla edustettuna useampaan kertaan regressiomallissa. Taulukosta 9 nähdään VIF-arvojen olevan lähellä arvoa 1, joten multikollineaarisuutta ei näiden muuttujien kesken juuri ole. Mallin muuttujien VIF-arvojen ollessa matalat voidaan jatkaa mallin oletusten tarkastelua tekemättä muutoksia mallin muuttujien valintaan.

Taulukko 9: Regressiomallin muuttujien VIF-arvot. Kaikkien muuttujien VIF-arvot ovat lähellä arvoa yksi, joten multikollineaarisuudesta ei ole viitteitä eikä muuttujien karsimista tarvitse tältä osin harkita.

	VIF-arvo
Aika_Demoihin	1.03
Aika_Uni	1.07
Taito	1.09
Haasteet	1.11
Aikataulu	1.07

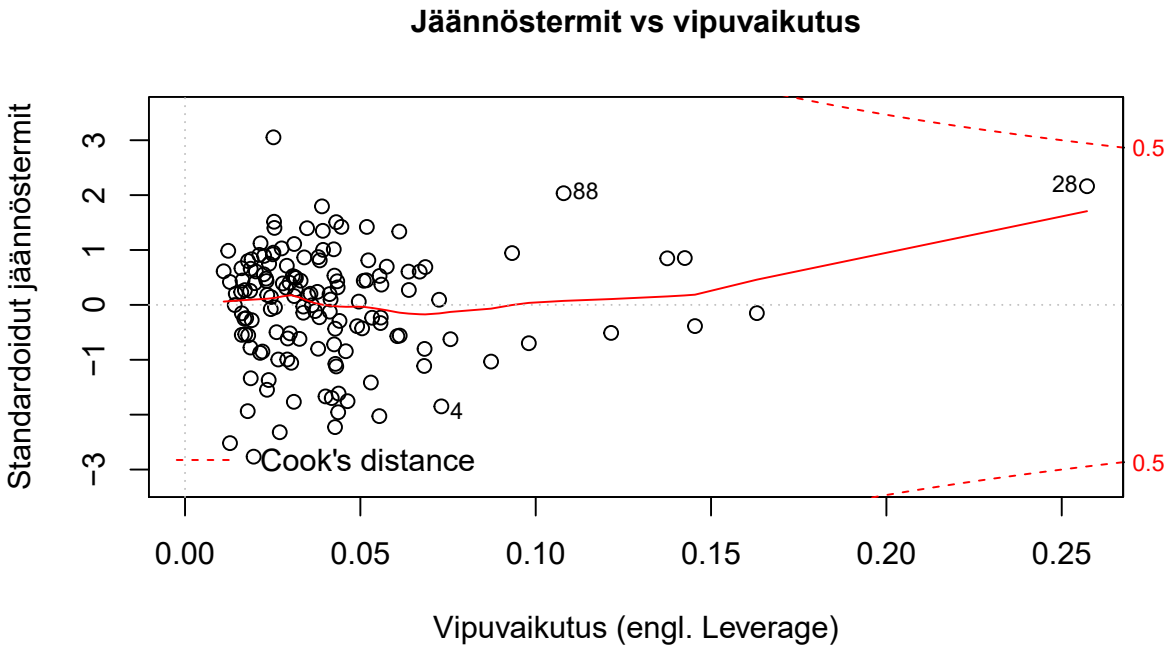
Kuva 11 tarkastelee lineaarisen mallin vaatimia oletuksia. Huomataan, että jäännöstermien varianssi pysyy suurinpiirtein samansuuruisena riippumatta vastemuuttujan odotusarvosta eli homoskedastisuus oletus täyttyy. Kuvaajan perusteella muuttujien välinen kovarianssi ei ole ongelma, koska jäännöstermit ovat jakaantuneet tasaisesti nollan molemmin puolin. Jäännöstermit eivät myöskään muodosta selkeää kuviota, joka viittaisi mallin epäsopivuuteen, vaan muutamaa poikkeavaa havaintoa lukuunottamatta jäännöstermit ovat jakaantuneet oletusten mukaisesti.



Kuva 11: Taulukon 8 muuttujista rakennetun lineaarisen regressiomallin jäännöstermit. Jokainen mallin opiskelija on sijoitettu kuvaajaan ennustetun vastemuuttujan ja jäännostermin arvojen perusteella. Punainen käyrä kuvastaa jäännöstermeihin sovitettua keskiarvoa ennustettujen arvojen suhteen. Tästä havaitaan jäännöstermien jakaantuneen tasaisesti nollan molemmin puolin. Tasaisesti jakaantuneet jäännöstermit ilman selkeitä kuviomaisia piirteitä viittaavat, että mallin oletukset ovat riittävän hyvin voimassa.

Kuvasta 12 voidaan erottaa poikkeavia havaintoja. Vaaka-akselin suure kertoo, kuinka paljon yksittäisen opiskelijan vastemuuttujan ennustettu arvo poikkeaa ennusteiden keskiarvosta. Tästä käytetään nimitystä vipuvaikutus (engl. *Leverage*). Pystyakselilla on standardoitujen jäännöstermien arvot. Kuvaajalla arvioidaan tarvetta poistaa havaintoja (opiskelijoita) mallista. Ongelmallisiksi tulkitaan havainnot, jotka on ennustettu poikkeamaan selvästi muista ennustetuista arvoista samalla, kun näiden havaintojen jäännöstermit ovat suhteellisen suuria muihin havaintoihin verrattuna. Kuvaajaan 12 on merkitty Cookin-etäisyys, joka auttaa havaitsemaan tällaiset havainnot. Havainto 28 on vaikutukseltaan suurin. Havainto 28 on opiskelija, jolla ei ole ollut *Aikataulu*-aihepiirimuuttujan perusteella tavoitteita mutta joka on silti kerännyt kohtuullisesti opintopisteitä. Yhtäkään havaintoa ei ole kuvaajan tai tarkastuksen perusteella tarpeellista poistaa mallista.

Tarkastelujen perusteella voidaan todeta oletusten täyttyvän lineaarisen regressiomallin käyttämiseksi.



Kuva 12: Kuvaajassa on standardoitujen jäännöstermien riippuvuus vipuvaikutuksesta. Vipuvaikutus kertoo, kuinka paljon yksittäisen opiskelijan vastemuuttujan ennustettu arvo poikkeaa ennusteiden keskiarvosta. Kuvaajan perusteella havaitaan poikkeavia havaintoja, jotka vaikuttavat suhteettoman paljon mallin estimaattien arvoihin. Ongelmallisiksi muodostuvat havaintojen arvot, jotka ovat jäännöstermeiltään ja vipuvaikutuksiltaan suuria. Cookin-etäisyyttä käytetään apuna päätöksessä havaintojen poistamiselle.

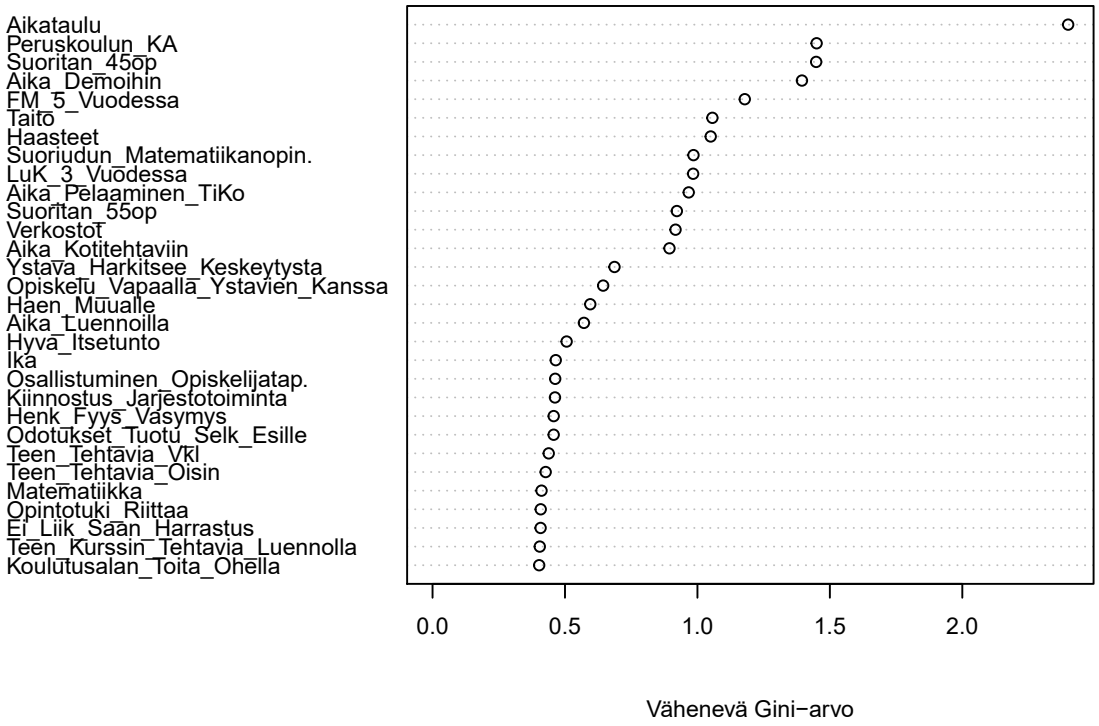
5.3 Satunnaismetsä

Satunnaismetsää käytetään tässä työssä luokittelutarkoitukseen. Kiinnostuksen kohteena on, saavuttaako opiskelija lukuvuoden aikana vähintään 45 opintopistettä. Tämä raja on kiinnostava, koska 45 opintopistettä vaaditaan 9 kuukauden opintotukien saamiseksi. Opiskelijat pyritään luokittelemaan siis kahteen ryhmään; *Kyllä saavuttavat* ja *Eivät saavuta* (asetettua opintopisterajaa).

Aineisto jaetaan ensiksi kahtia *opetus*- ja *testaus*-osuuksiin. Ensimmäisessä tarkastelussa käytetään kaikkia taustatietokyselyn muuttujia sekä niistä tehtyjä aihepiirimuuttujia. Kuva 13 esittää, mitkä solmukohdat ovat merkityksellisiä. Tähän on käytetty *vähenevää Gini* -arvoa, joka kertoo, kuinka suuri rooli muuttujalla on havaintojen jakamisessa.

Opetusaineistoon sovitettua luokittelumallia kokeillaan seuraavaksi testausaineistoon. Taulukosta 10 havaitaan, että suurin osa opiskelijoista saavuttaa kyseisen rajan, ja näiden opiskelijoiden ennustamisessa on onnistuttu kohtalaisesti. Malli ennustaa vain 4 opiskelijan jäävän rajan alle, vaikka todellinen luku on 18 opiskelijaa.

Kuvasta 13 erottuu selkeästi tärkeimpänä aihepiirimuuttuja *Aikataulu*, mutta myös muut aihepiirimuuttujat sijoittuivat arvioinnissa korkealle. Satunnaismetsän pitäisi olla robusti sekoittaville muuttujille, mutta testataan mallia jättämällä siihen pelkästään aihepiirimuuttujat. Tämän mallin tulokset esitetään taulukossa 11.



Kuva 13: Kuvaajasta nähdään, mitkä muuttujista ovat merkityksillisimpiä opiskelijoiden luokittelemisessa. Kaikki aiemmin muodostetut aihepiirit sijoittuvat vertailussa kuvaajan ylälaitaan ja ovat näin ollen merkityksellisiä opiskelijoiden jakamisessa. Aikataulu-muuttuja erottuu muista muuttujista kaikista selkeimmin.

Taulukko 10: Tässä ennustetaulussa on käytetty kaikkia taustatietokyselyn muuttujia sekä näistä muodostettuja aihepiirimuuttujia. Malli ennustaa vain harvan opiskelijan jäävän asetetun 45 opintopisteen rajan alle lukuvuoden aikana.

Ennuste	Todellinen	
	Ei	Kyllä
Ei	4	2
Kyllä	14	39

Taulukossa 12 on tärkeimpiä tunnuslukuja taulussa 11 esitetyistä ennustusten tuloksista. Ennusteen tarkkuus (engl. *accuracy*) oli noin 0.75 (44/59). Ennusteen spesifisyys (engl. *specificity*) on 0.22 (4/18), joka kertoo osuuden oikein ennustetuista opintopisterajan alle jäävistä opiskelijoista verrattuna kaikkiin opintopisterajan alle jääviin opiskelijoihin. Mallin spesifisyys on heikko, mutta malli ennustaa kuitenkin tarkasti kaikki opiskelijat, jotka tulevat asetetun opintopisterajan saavuttamaan. Tätä kuvataan sensitiivisyydellä (engl. *sensitivity*), joka ennusteessa on 0.98 (40/41). Toisin sanoen malli ennustaa pienen osan opiskelijoista jäävän asetetun opintopisterajan alle, mutta näistä opiskelijoista neljä havaintoa viidestä todella jäi asetetun opintopisterajan alapuolelle.

Samanlaiset huomiot voidaan tehdä molempien satunnaismetsämallien ROC-käyristä

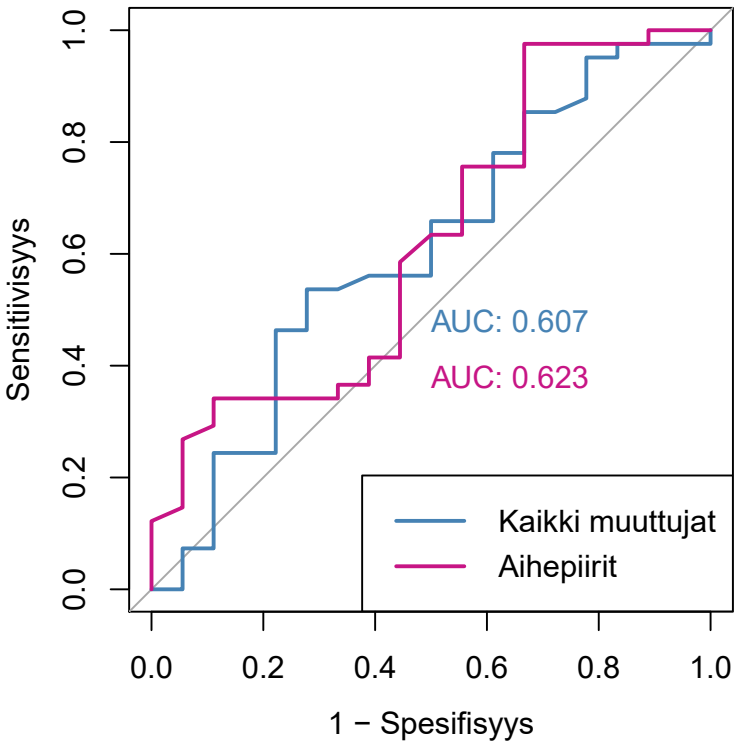
Taulukko 11: Satunnaismetsän ennustetaulu. Tämä on saatu mallista, johon on otettu vain aihepiirimuuttujia mukaan. Ennustetaulun perusteella havaitaan parannusta malliin, johon on otettu mukaan myös kaikki taustatietokyselyn muuttujat. Tämäkin malli ennustaa vain harvan opiskelijan jäävän asetetun opintopisterajan alle, mutta suurin osa näistä opiskelijoista on luokiteltu oikein.

Ennuste	Todellinen	
	Ei	Kyllä
Ei	4	1
Kyllä	14	40

Taulukko 12: Satunnaismetsän ennustemallin yhteenveto. Luottamusväli on laskettu mallin tarkkuudelle. Mallin sensitiivisyyden arvo on korkea, mutta spesifisyyden arvo puolestaan matala.

	Arvo
Tarkkuus	0.75
Luottamusväli	[0.62, 0.85]
Sensitiivisyys	0.98
Spesifisyys	0.22

(kuva 14), johon on kuvattu sensitiivisyyden ja spesifisyyden arvoja mallin eri kynnysarvoilla. Tällä tarkoitetaan sitä luokittelussa käytettävää todennäköisyyden arvoa, jolla päätetään kumpaan luokkaan opiskelija luokitellaan. Kuvaajaan on myös merkitty käyrän alle jäävän alueen pinta-ala (AUC). AUC-arvot ovat kummassakin mallissa lähellä 0.6:a. Huomataan mallien täydentävän toisiaan, joten ei ole aivan selvää kumpi malleista luokittelee opiskelijat paremmin. Bootstrapilla lasketut luottamusvälit kaikkien muuttujien mallin AUC-arvolle on [0.44, 0.61] ja vastaavasti vain aihepiirimuuttujia sisältävälle mallille [0.46, 0.63].



Kuva 14: ROC-käyrien vertailu kaikkien muuttujien satunnaismetsässä ja vain aihepiirimuuttujia sisältävässä satunnaismetsässä. Malleissa kynnysarvona on todennäköisyys, jolla opiskelijoita luokitellaan joko saavuttamaan asetettu opintopisteraja tai jäämään asetetun opintopisterajan alle. Kun kynnysarvo opiskelijoiden luokittelemiselle opintopisterajan saavuttamiseksi on matala, niin aihepiirimuuttujista koostuva malli suoriutuu paremmin. Vastaava havainto tehdään, kun tämä kynnysarvo on korkea. Vain aihepiirimuuttujista koostuva malli suoriutuu kuitenkin heikosti, kun kynnysarvot ovat lähempänä todennäköisyyden puoltaväliä.

6 Pohdinta

Tämän työn tarkoituksena oli arvioida ja ennustaa, mitkä tekijät vaikuttavat opiskelijoiden keräämiin opintopisteisiin ensimmäisen opiskeluvuoden aikana. Analyysien tuloksien perusteella kysymysten jaotteleminen aihepiireihin ja näistä muodostetut uudet muuttujat olivat kaikissa käytetyissä menetelmissä merkityksellisimpiä. Klusteroinnissa näillä aihepiirimuuttujilla oli selkeitä eroja klusterien opintopistekeskiarvojen välille ja regressioanalyysissä muuttujien vaikutus opintopistemäärän ennustamisessa oli selkeä.

Opintopisteiden kertyminen on korkeakoulusta valmistumisen välttämätön edellytys. Tässä työssä keskityttiin tulevaisuuden teknologioiden laitoksen opiskelijoiden ensimmäiseen opiskeluvuoteen sen sijaan, että olisi tutkittu koko tutkinnon kattavia opintoja. Ensimmäinen opiskeluvuosi on luonnollisesti tärkeä opintojen alkuunpääsemisen ja korkeakouluun sopeutumisen kannalta. Oikein kohdistetuilla kysymyksillä saatiin kerättyä mielekästä tietoa suunnitellusta opintojen etenemisestä. Samanlaiset aiheet nousivat merkittäviksi sekä ensimmäisen vuoden opintopistemäärien (tämä tutkimus) että valmistumisen ennustamisen kannalta (aiempi tutkimus). Opintojen edistymiseen

vaikuttavien muuttujien selvittäminen auttaa opinto-ohjauksen suunnittelussa, jolloin opintojen edistymistä haittaaviin tekijöihin voidaan tarvittaessa, ja opiskelijan niin toivoessa, puuttua korkeakoulun puolelta.

Aiemmat opintomenestystä koskevat tutkimukset ovat pääasiassa keskittyneet ennustamaan opintojen valmistumista tai keskeytymistä ja selittämään näihin vaikuttavia tekijöitä (Rautopuro ja Korhonen 2011). Motivaation puutteen ja elämänongelmien on aiemmissa tutkimuksissa havaittu lisäävän opintojen keskeyttämisen riskiä. Myös tässä työssä opiskelijan motivaatiota ja taitoa sekä haasteita mittaavat kysymykset olivat merkityksellisiä. Näistä kysymyksistä muodostetut aihepiirimuuttujat *Taito* ja *Haasteet* olivat sekä luokittelu- että regressiomalleissa vaikuttavimpien muuttujien joukossa ennustettaessa opintopistemääriä. Lineaarisessa regressioanalyysissa aihepiirimuuttujan *Taito* vaikutus oli positiivinen (piste-estimaatin arvo 31, 95% luottamusväli [3.4, 62.8]). Tämä tarkoittaa, että jokaista *Taito*-muuttujan 0.1 yksikön lisäystä kohden opintopisteiden odotusarvo kasvaa 3.1:llä. Aihepiirimuuttujan *Haasteet* vaikutus oli negatiivinen (piste-estimaatin arvo -26, 95% luottamusväli [-51.8, -0.5]). *Haasteet*-muuttujassa jokainen 0.1 yksikön lisäys laskee opintopisteiden odotusarvoa 2.6:lla. Aihepiirimuuttujien mahdolliset arvot olivat välillä [0,1], mutta kuvasta 6 havaittiin, että arvojen vaihteluvälit olivat todellisuudessa tätä mahdollista vaihteluväliä pienempiä.

Aihepiirimuuttujien *Taito* ja *Haasteet* lisäksi muodostettiin kaksi muuta aihepiirimuuttujaa: *Verkostot* ja *Aikataulu*. Aihepiirimuuttuja *Verkostot* muodostettiin kysymyksistä, jotka mittasivat kiinnostusta opintojen sosiaaliseen puoleen. Aiemmissa tutkimuksissa (Ashraf et al. 2018) yhteenkuuluvuuden tunteella on havaittu olevan merkitystä valmistumisen kannalta, mutta tässä tutkimuksessa muuttujan *Verkostot* perusteella tällä ei kuitenkaan ollut tilastollista merkitsevyyttä kerättyjen opintopisteiden suhteen. Muuttujaa *Verkostot* voi kuuluvuuden tunteen lisäksi tarkastella myös toiselta näkökannalta. Opintoihin liittyvien oheistoimien, kuten opiskelijatapahtumiin osallistumisen, voidaan ajatella haittaavan opintojen etenemistä.

Aiemmissa tutkimuksissa motivaation hallintaan on havaittu vaikuttavan tavoitteiden asettamisen (Cheung 2004), johon liittyvistä kysymyksistä tässä työssä tehtiin oma aihepiirimuuttujansa *Aikataulu*. Tämä muuttuja osoittautui kaikkein vaikuttavimmaksi; Lineaarisessa regressiomallissa muuttujan piste-estimaatti oli 48 (95% luottamusvälillä [23.9, 75.2]), eli jokaista muuttujan 0.1 yksikön lisäystä kohden opintopisteiden odotusarvo kasvaa 4.8:lla. Satunnaismetsässä *Aikataulu*-muuttuja oli kaikista tärkein muuttuja jakamaan aineistoa puiden solmukohdissa (vähenevän Gini-arvon perusteella).

Klusteroinnilla saatiin tässä työssä onnistuneesti eroteltua opiskelijoita (taulukko 5). Klusterit muodostettiin kolmen aihepiirimuuttujan perusteella (*Taito*, *Haasteet* ja *Aikataulu*). Klustereita muodostettiin kolme kappaletta, joista ensimmäisen klusterin opintopistekeskiarvo oli selkeästi kahden muun klusterin opintopistekeskiarvoja pienempi. Ensimmäisen klusterin *Aikataulu*-muuttujan keskiarvo oli myös selkeästi pienempi verrattuna kahteen muuhun klusteriin. Myös muuttujissa *Taito* ja *Haasteet* klustereiden välillä oli eroja, mutta ne eivät olleet yhtä selkeitä kuin *Aikataulu*-muuttujassa. Toinen ja kolmas klusteri erosivat keskiarvoissa toisistaan käytännössä vain *Haasteet*-muuttujan osalta, eikä näiden klustereiden välillä ollut opintopistekeskiarvoissa tilastollisesti merkitsevää eroa. Tämän perusteella klusterien erot opintopistekeskiarvoissa seurasivat lähinnä klusterikohtaisista eroista *Aikataulu*-muuttujan keskiarvoissa. Opiskelijan oma arvio opintojen etenemisestä on näin ollen merkittävä tekijä, kun ennustetaan opiskelijoiden saavuttamia opintopistemääriä.

Taustatietokyselyyn eivät vastanneet kaikki tulevaisuuden teknologioiden laitoksella aloittaneet opiskelijat tarkastelluilta vuosilta (2015–16 ja 2016–17). Vastanneiden ja vastaamatta jättäneiden opiskelijoiden eroja tarkasteltiin vain kerättyjen opintopistemäärien ja suoritettujen kurssien suhteen eikä muita merkityksellisiä mittareita ollut saatavilla. Opintopistemäärissä oli näiden kahden ryhmän välillä eroja. Tutkimukseen osallistuneet opiskelijat saavuttivat keskimäärin noin. 10 opintopistettä enemmän lukukauden aikana, kuin opiskelijat, jotka eivät tutkimukseen osallistuneet. Tämä voidaan tulkita niin, että opiskeluidensa etenemisestä kiinnostuneet opiskelijat vastaavat helpommin vapaaehtoisiin kyselyihin kuin opiskeluidensa etenemisestä vähemmän kiinnostuneet opiskelijat.

Kysely suoritettiin vain lukuvuoden alussa, joten tietoa opintoja koskevien suunnitelmien mahdollisista muutoksista ei ollut käytettävissä. Kyselyn vastauksien totuudenmukaisuutta ei pystytty arvioimaan luotettavilla mittareilla, joten opiskelijoiden vastaukset eivät välttämättä vastaa täysin heidän ajatuksiaan. Pääkomponenttianalyysissä samankaltaisten muuttujien samansuuntainen vaikutus viittaa kuitenkin yhtenäisyyteen vastauksien välillä. Opiskelijoilla ei myöskään ollut syytä valehdella vastauksissaan, koska kyselyyn vastaamisesta ei ollut opiskelijoille suoranaista hyötyä.

Tämän työn aineiston perusteella menetelmien ennustevoima ei ole riittävä luotettavien ennustemallien rakentamiseen. Lineaarisen regressioanalyysin estimoidut vastemuuttujien arvot erosivat selkeästi oikeista vastemuuttujien arvoista. Tämä nähtiin jäännöstermien kuvasta 11, jossa jäännöstermit olivat suuria suhteessa ennustettuihin vastemuuttujien odotusarvoihin. Lineaarinen regressiomalli pystyi selittämään 25% kerättyjen opintopisteiden vaihtelusta. Mallin luotettavuutta arvioitaessa poikkeavien havaintojen käsitteleminen tuotti vaikeuksia, koska epäjohdonmukaisuuksia taustatietokyselyn vastausten ja opintopistemäärien välillä ei voida selittää epäilyksellä yksilön vastausten totuudenmukaisuuteen. Näin ollen yksilöitä ei myöskään poistettu mallista.

Opiskelijoita luokiteltiin satunnaismetsä-menetelmällä sen mukaan, saavuttavatko he asetetun opintopisterajan. Kuvaajasta 14 nähtiin opiskelijoiden luokittelemisen onnistuminen kahden eri mallin ROC-käyristä. Vain aihepiirimuuttujia sisältävässä mallissa saatiin korkea sensitiivisyys (0.98) luokittelemalla vain muutama opiskelija jäämään opintopisterajan alapuolelle. Vaikka nämä opiskelijat olivat suurimmaksi osaksi luokiteltu oikein, mallin spesifisyys (0.22) jäi matalaksi. Mallin AUC-arvon luottamusväli ([0.46, 0.62]) sisältää arvon 0.5, joka saataisiin myös satunnaisella luokittelulla. Tämän perusteella valitut muuttujat eivät ole vielä riittäviä luotettavien luokittelumallien käyttämiseksi.

Tämän työn tulokset antavat aiemman tutkimuksen kanssa samansuuntaisia havaintoja niistä merkitsevistä muuttujista, jotka vaikuttavat opintopisteiden kertymiseen. Luotettavien ennusteiden tai luokittelujen tekemiseksi nämä muuttujat eivät ole vielä riittäviä. Tuloksista voidaan kuitenkin päätellä, että näiden muuttujien avulla ennusteita ja luokitteluja pystytään jatkossa parantamaan satunnaisiin malleihin verrattuna.

Tutkimuksen jatkon kannalta kysymyksiä täytyisi kohdentaa vielä tarkemmin merkityksellisiksi havaittuihin aihepiireihin. Suunnitelmia tai motivaatiota voitaisiin kysellä myös myöhemmässä vaiheessa lukuvuotta, jolloin havaittaisiin mahdolliset muutokset opiskelijoiden tilanteessa. Nämä mahdolliset muutokset toisivat mukanaan myös uusia tutkimuskysymyksiä esimerkiksi opiskelijoiden motivaation muutoksiin liittyen.

7 Yhteenveto

Opiskelijoiden opintojen edistäminen on sekä oppilaille että oppilaitoksille oleellinen kysymys. Tässä tutkimuksessa tarkasteltiin tilastollisin keinoin, mitkä aiheet ovat opiskelijoille merkittäviä tekijöitä opintopisteiden keräämisessä ensimmäisen opiskeluvuoden aikana. Tuloksista havaittiin, että tarkkojen ennustemallien luominen on vaikeaa, mutta eri lähestymistavoista huolimatta samat aihepiirit nousivat useissa analyyseissa merkittäviksi vaikuttajiksi. Varsinkin opiskelijoiden oma arvio opintojen etenemisestä oli yhteydessä opintopistemäärään. Epävarmuus yksittäisten muuttujien vaikutuksista opintopistemääriin oli kuitenkin suurta. Tarkemmat johtopäätökset vaativat lisää aineistoa ja jatkotutkimusta.

Oppimisanalytiikan yleistyessä kehitetään yhä useampia tapoja mitata opiskelijoiden suoriutumista. Osaltaan tämän työn tarkoituksena on ollut antaa suuntaviivoja sille, mihin oppimisen etenemisen mittaamisen kannalta voimavaroja kannattaa suunnata. Ongelma on monimutkainen, eikä ole mahdollista arvioida jokaisen opiskelijan suoriutumista tarkasti. Suuremmassa mittakaavassa keskeisiin ongelmiin puuttuminen ja opiskelijoiden tukeminen on kuitenkin mahdollista, ja aiheena sellainen, johon jatkossa resursseja myös keskitetään.

Viitteet

Ashraf R, Godbey J, Shrikhande M ja Widman T. 2018. "Student Motivation and Perseverance: Do They Explain College Graduation?" *Journal of the Scholarship of Teaching and Learning* 18 (3): 87–115.

Bezdek J. 1973. "Cluster Validity with Fuzzy Sets." *Journal of Cybernetics* 3: 58–73.

Breiman L. 2001. "Random Forests." *Machine Learning* 45: 5–32.

Chatterjee S ja Hadi A. 1986. "Influential Observations, High Leverage Points, and Outliers in Linear Regression." *Statistical Science* 1 (3): 379–416.

Cheung E. 2004. "Goal Setting as Motivational Tool in Student's Self-Regulated Learning." *Educational Research Quarterly* 27 (3): 3-9.

Cortez P ja Silva A. 2008. "Using Data Mining to Predict Secondary School Student Performance." *Proceedings of 5th Annual Future Business Technology Conference*, Porto, Portugal, 5-12.

Efron B. 1987. "Better Bootstrap Confidence Intervals." *Journal of the American Statistical Association* 82 (397): 171–185.

Efron B ja Tibshirani R. 1986. "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy." *Statistical Science* 1: 54–77.

Fawcett T. 2006. "An Introduction to Roc Analysis." *Pattern Recognition Letters* 27 (8): 861–874.

Fonti V ja Belitser E. 2017. "Feature Selection Using Lasso." https://beta.vu.nl/nl/Images/werkstuk-fonti_tcm235-836234.pdf 9.12.2019

Franke G., 2010. Multicollinearity. *Wiley International Encyclopedia of Marketing.* https://onlinelibrary.wiley.com/doi/full/10.1002/9781444316568.wiem02066 9.12.2019

Gaertner M ja McClarty K. 2015. "Performance, Perseverance, and the Full Picture of College Readiness." *Educational Measurement: Issues and Practise* 34 (2): 20–33.

Han H, Guo X ja Yu H. 2016. "Variable Selection Using Mean Decrease Accuracy and Mean Decrease Gini Based on Random Forest." *2016 7th IEEE international conference on software engineering and service science (icsess)*: 219-224.

Hastie T, Tibshirani R ja Friedman J. 2009. *The Elements of Statistical Learning Data Mining, Inference, and Prediction.* 2nd ed. Springer. New York.

Häkkinen P ja Arvaja M. 1999. "Kollaboratiivinen oppiminen teknologia ympäristöissä." *Teoksessa A. Eteläpelto ja P. Tynjälä (Toim.) oppiminen ja asiantuntijuus. Työelämän ja koulutuksen näkökulmia. Juva: WSOY.*

Järvinen H, Pääkkönen K, Rantala H ja Väänänen M. 2018. "Oppimisanalytiikka Suomessa – Nykytilanne, tulevaisuus ja haasteet." Tampereen Ammattikorkeakoulu.

Jolliffe I. 2002. *Principal Component Analysis.* 2nd ed. Springer. New York.

Kodinariya T ja Makwana P. 2013. "Review on Determining Number of Cluster in K-Means Clustering." *International Journal* 1 (6): 90–95.

Liaw A ja Wiener M. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22.

Rautopuro J ja Korhonen V. 2011. “Yliopisto-opintojen keskeyttämisriski ja opintoihin kiinnittymisen ongelmat.” *Teoksessa: M. Mäkinen, V. Korhonen, J. Annala, P. Kalli, P. Svärd ja V-M Värri (toim.) Korkeajännityksiä - kohti osallisuutta luovaa korkeakoulutusta. Tampere University Press*: 36-58.

Tibshirani R. 2011. “Regression Shrinkage and Selection via the Lasso: A Retrospective. *”Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (3): 273–282.

Tuononen T, Parpala A, Haarala-Muhonen A ja Lindlom-Ylänne S. 2016. “Yliopisto-opintojen aikainen työssäkäynti: ajanhallinta- ja itsesäätelytaitojen merkitys opintojen etenemiselle.” *Tiedepolitiikka*, no. 4: 53–60.

Weisberg S. 2005. *Applied Linear Regression.* 3rd ed. John Wiley & Sons, Inc. Hoboken, New Jersey.

Xie L, Wang Y, Chen L ja Yue G. 2010. “An Anomaly Detection Method Based on Fuzzy C-Means Clustering Algorithm.” *Second International Symposium on Networking and Network Security (ISNNS, 10) Jinggangshan, PR China*: 2-4.

Yang F ja Li F. 2018. “Study on Student Performance Estimation, Student Progress Analysis, and Student Potential Prediction Based on Data Mining.” *Computers and Education* 123: 97–108.