Kanditutkielma

Tietojenkäsittelytieteen kandiohjelma

Oppijan kehittymisen tukeminen oppimisanalytiikalla Moodlessa

Tuomas Alanen

26.3.2022

Matemaattis-luonnontieteellinen tiedekunta

Helsingin yliopisto

Yhteystiedot

PL 68 (Pietari Kalmin katu 5)
00014 Helsingin yliopisto

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

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

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

Tekijä — Författare — Author

Tuomas Alanen

Työn nimi — Arbetets titel — Title

Oppijan kehittymisen tukeminen oppimisanalytiikalla Moodlessa

Ohjaajat — Handledare — Supervisors

Prof. D.U. Mind, Dr. O. Why

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Kanditutkielma	26.3.2022	15 sivua

Tiivistelmä — Referat — Abstract

Kirjoita tiivistelmä tähän.

Varmista, että kaikki pakolliset kohdat lomakkeella on täytetty.

Listaa ACM CCS 2012 -luokituksesta 1-3 polkua kuvaamaan työtäsi. Kts englanninkielinen ohje.

ACM Computing Classification System (CCS)
General and reference → Document types → Surveys and overviews
Applied computing → Document management and text processing → Document management → Text editing

Avainsanat — Nyckelord — Keywords

learning analytics, Moodle

Säilytyspaikka — Förvaringsställe — Where deposited

Helsingin yliopiston kirjasto

Muita tietoja — övriga uppgifter — Additional information

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

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

Tekijä — Författare — Author

Tuomas Alanen

Työn nimi — Arbetets titel — Title

Oppijan kehittymisen tukeminen oppimisanalytiikalla Moodlessa

Ohjaajat — Handledare — Supervisors

Prof. D.U. Mind, Dr. O. Why

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Bachelor's thesis	March 26, 2022	15 pages

Tiivistelmä — Referat — Abstract

Write your abstract here.

In addition, make sure that all the entries in this form are completed.

Finally, specify 1–3 ACM Computing Classification System (CCS) topics, as per https://dl.acm.org/ccs. Each topic is specified with one path, as shown in the example below, and elements of the path separated with an arrow. Emphasis of each element individually can be indicated by the use of bold face for high importance or italics for intermediate level.

ACM Computing Classification System (CCS)
General and reference → Document types → Surveys and overviews
Applied computing → Document management and text processing → Document management → Text editing

Avainsanat — Nyckelord — Keywords

learning analytics, Moodle

Säilytyspaikka — Förvaringsställe — Where deposited

Helsinki University Library

Muita tietoja — övriga uppgifter — Additional information

Sisällys

1 Johdanto

Tutkimuskysymykset:

1. Millaista dataa Moodlesta saadaan oppimisanalytiikan prosessin käyttöön
2. Miten saatua dataa voidaan hyödyntää oppijan tukemiseksi oppimisanalytiikan avulla

HOX! Punaisella merkityt tekstiosuudet ovat tutkimuspäiväkirjan sisältöä, johon olen kirjannut ylös erilaisia havaintoja ja hyviä lähdeaineistoja talteen hyödynnettäväksi myöhemmissä vaiheissa. Normaalit tekstiosuudet ovat varsinaista kandidaatin tutkielman sisältöä.

2 Oppimisen analysoinnin tarpeet

Halutaan tutkia opiskelijan kurssisuorituksien vaikutusta opiskelijan menestykseen kurssilla.

Learning analytics cycle - data -> analyisis -> action (Hasan et al., 2020)

2.1 Oppimisanalytiikka pedagogisena työkaluna

Data voidaan jaotella kahteen laatuun, aktiiviseen ja passiiviseen. Madden et al., 2007 Mikkola, 2019

Määritelmä Siemens, 2013

Table 1 Long ja Siemens, ei julkaisupäivää

AnalytiikkaÄly-hanke, tsekkaa tää -> datan käsittelystä tarkasta tukevien lähteiden varalta Kokkonen, ei julkaisupäivää E. Kaila (UTU/HY)

Romero ja Siemens toistuvia nimiä

Learning analytics LA <–> Educational Data Mining EDM

Virtual Learning Environment VLE <–> Learning Management System LMS

2.2 Moodle datalähteenä

Mitä kaikkea saadaan ulos

SQL

1. raa'alla voimalla SQL hakuina kannasta => aktiviteettitunnisteet apuna?
2. tauluja listattuna (Romero et al., 2014)

Moodlen LA API => onko tästä mihinkään? ainakin pikavilkaisu paikallaan

Ohtuprojektin projektikeskusteluun heittämäni oppimisanalytiikka-aiheiset viestit liittyen opiskelijan toiminnann seuraamiseen aikaleimojen perusteella, teksti mallilla kirjoiteltu mitä tiedetään

Teoriassa opiskelijan käyttämää aikaa oppimisympäristössä voidaan mitata tallentamalla aikaleimoja sivulatauksista, mutta tämä ei ole kovin luotettava menetelmä - perustelen sen sillä, että tällä ei voida mitata ollenkaan mitä käyttäjä tekee - keskittyykö hän tehtävään, vai onko hän hakemassa kahvia.

Teoriassa sen järkevä toteutus vaatisi 3. osapuolen seurantapalikan, kuten kaikkien rakastama (jopa lähes GDPR-compliant...) Google Analytics, jolla saataisiin oikeasti kerättyä relevanttia dataa siitä, miten käyttäjä toimii sivulla. Yksi vaihtoehto olisi pyörittää omaa Matomo-instanssia, mutta tämä vaatii palvelimelle asennettavaksi Moodlesta riippumattoman sovelluksen, ja tämä olisi aina palvelinkohtainen - jokainen joutuisi siis itse ensin asentamaan Matomon. Sama palvelinriippuvuus tulee Google Analyticsin kanssa, kun siinäkin puhutaan sivustokohtaisesta seurannasta.

Kolmas vaihtoehto on toteuttaa itse analytiikkascripti... tämä olisi tietysti teknisesti mahdollista, mutta kuvaan story-point asteikolla kohdetta eeppiseksi. Lisäksi tämä vaatii tutkimista kuinka me saataisiin edelleenkään trackattua sitä, mikä meitä oikeasti kiinnostaa - mitä käyttäjä tekee, halutaanko mitata hiiren liikettä sivulla, aikaa sivulatauksesta mikrofonin nauhoitusnapin painoon, jotain muuta?

tl;dr; jälkimmäinen voidaan toteuttaa kyllä, mutta tämä tieto ei mielestäni ole sellaista luotettavuudeltaan, minkä avulla voitaisiin tehdä yhtään mitään päätelmiä opiskelijan työskentelystä johtuen teknisistä rajoitteista. Asioita voidaan kyllä mitata ja esittää, mutta itsellä ainoana kysymysmerkkinä on tiedon luotettavuus kuvaamaan mitattua asiaa. Jos toteutetaan sivulatauksien mittaamisen lisäksi JS-script, joka seuraa opiskelijan käyttäytymistä aktiviteetissä, on täten todennäköisyys saada luotettavampaa tietoa mitattua korkeampi.

Kaivelin vähän papereita tähän liittyen ja törmäsin muutamaan:

Google analytics for time behavior measurement in Moodle (Filvà et al., 2014) // Tässä tehtiin oppimisanalytiikkaa yleisemmin käyttäen Google Analyticsia Moodlen osana. Tässä on yhdistetty myös puhuttu Student Dashboard, ja esimerkkejä myös GA:n avulla mitatusta time spent on page arviosta, joka edustaa luotettavampaa mittausta kuin pelkkä sivulatauksien välin mittaaminen.

Exploring Student Interactions: Learning Analytics Tools for Student Tracking (Conde et al., 2015) // Tässä hyvä esimerkki, että Moodlen oppimisanalytiikkaan on paljon hyviä työkaluja, mutta mikään ei suoranaisesti vastaa meidän tarpeeseen (aktiviteettikohtainen analysointi arvioinnin tueksi), vaan käsittelevät Moodlea LMS/VLE -kokonaisuutena. Nä-

mäkin ovat järkiään kokonaan erillinen palvelu tai Moodleen asennettava lisäpalikka. Samasta hauska huomio, että keskimääräisesti aika moni Moodleen ja oppimisanalytiikkaan liittyvä paperi tulee Espanjasta.

Tiivistetysti näyttäisi siltä, että vaihtoehdot on joko 3. osapuolen GA-tyylinen ratkaisu tai purkkaa oma versio kasaan, joka lienee se järkevin tapa mun osint-tiedustelutiedon perusteella

Erilaiset saatavat aineistot

1. Gradebook
    - quizzes
    - assignments
    - workshop etc
    - forum
2. Events
    - core course module viewed
    - modassign submission created
    - modassign submission updated
    - koko event api ylipäätään
3. events in the report logs -> Agudo-Peregrina et al., 2014
4. suorituskertojen määrä
5. pikapalaute (block_point of view)
6. muu palaute
7. eri muuttujat (Mwalumbwe ja Mtebe, 2017)

2.3 Päättelymahdollisuudet

Oppimisanalytiikan kolmijako Daniel @ eoppimiskeskus, 2017

Mitä voidaan päätellä

1. opiskelijan menestys kurssilla
2. dropout alert
3. opetusmenetelmien tehokkuus
4. haasteet oppimateriaalissa - katseltu useasti ja tehtävä menee edelleen päin honkia

3 Datamalli oppijan kehittymisesta

Oppimisanalytiikassa yhdistelemällä tilastollisia menetelmiä ja predicative modelling käyttämällä voidaan kohdentaa ohjausta opiskelijoiden haasteisiin oppimisessa ja tarjoamalla kohdistettua tukea saatavan datan avulla (Ranjeeth et al., 2020). Käytettävät prediktiiviset mallit voivat olla mitä vain datanlouhinta-, koneoppimis- ja keinotekoisilla menetelmillä.

Yksittäisiä opiskelijoita verrataan prediktiivisen mallin avulla muodostettuihin malleihin, jotka kuvaavat keskimääräistä opiskelijaa (Wolff et al., 2013). Esimerkiksi prediktiivisen mallinnuksen avulla voidaan ennustaa esimerkiksi kuinka opiskelija tulee menestymään kurssilla; onko pääsemässä kurssia läpi? Tämä tapahtuu edellä kuvatulla tavalla, ja ennusteen perusteella katsotaan onko opiskelija vaarassa olla läpäisemättä kurssia.

3.1 Yleistetty malli

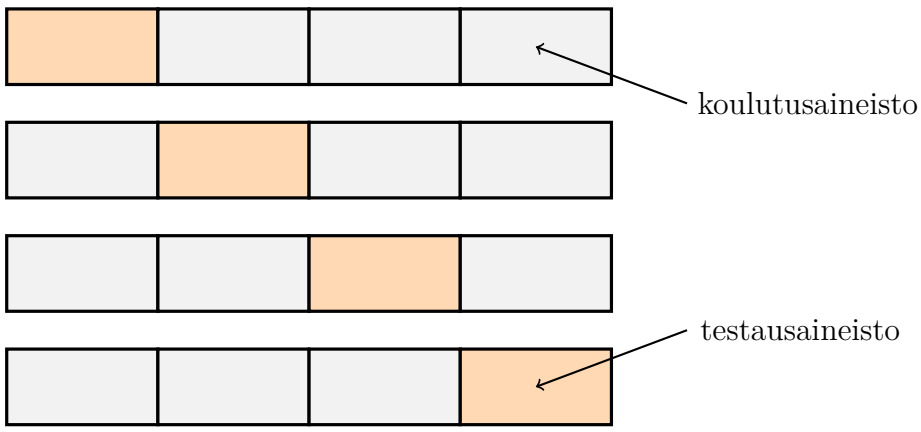
Ennen datan syöttämistä millekään analysiontia tai luokittelua tekevälle mallille, tulee aineistolle suorittaa esikäsittely (Romero et al., 2014). Esikäsittelyssä ensin tarvittava data kerätään kasaan ja ryhmitellään sopivasti järkeviin kokonaisuuksiin. Datan ollessa kasassa ryhmiteltynä, poistetaan siitä kaikki epäolennainen ja virheellinen sisältö. Kohdistaaksemme analyysin opiskelijoihin, täytyy aineistosta tunnistetaa käyttäjät sekä heidän eri asiointisessiot. Tämän jälkeen valitaan sopivat selittävät muuttujat jättääksemme kaikki korreloivat ja toisteiset muuttujat pois. Tämän jälkeen isoista data-aineistoista poistetaan aiempien vaiheiden jälkeen turhiksi jääneet kentät, jotka olisivat epäolennaisia prosessille tai toisteisia. Lopuksi tarkastellaan mahdollisuutta muodostaa uusia muuttujia olemassa olevien muuttujien perusteella. Esimerkiksi voidaan normalisoida muuttujan arvot jollekin tietylle välille tai muutetaan esitystapaa sopivammaksi.

Datamallin rakentamisessa on useita vaiheita, ja yleensä puhutaan iteratiivisesta prosessista (Hämäläinen ja Vinni, 2010). Iteratiivisen prosessin aikana kokeillaan useita erilaisia malleja, datan esitysmuotoja ja algoritmien asetuksia löytääksemme parhaan mahdollisen. Valitun mallin toimivuus voidaan todentaa luokittelun onnistumisella - jos mallille tulee lian monta virhettä, voidaan mallin soveltuvuus kyseenalaistaa.

Oppimisanalytiikassa usein käytetään luokittelua, kuten myös opetuksessa yleisesti opettajien arvioidessa oppijoiden tietotasoa, motivaatiota ja käytöstä (Hämäläinen ja Vinni, 2010). Oppimisanalytiikassa luokittelua tehdään selitettävän muuttujan arvoa ennustavalla mallilla, joka hyödyntää selittävien muuttujien arvoja. Luokittimia voidaan tehdä joko ammattilaisten käsityönä tai nykyisin yleisemmällä tavalla opettaa luokitin luokittelemaan olemassa olevalla datalla.

Useissa oppimisanalytiikkaa käsitelleissä tutkimuksissa on kokeiltu erilaisia luokittelualgoritmejä data-aineistolle löytääkseen parhaiten toimivan mallin (Akçapınar et al., 2019). Kokeiltuja algoritmejä ovat naiivi Bayes, Classification Tree, Random Forest, tukivektorikone (SVM), neuroverkko, CN2 rules ja k-lähinaapurimenetelmä. Parhaiten toimivaa mallia voidaan etsiä esimerkiksi suorituskykymittauksin, jossa tarkastellaan tarkkuutta, herkkyyttä, yksityiskohtaisuutta ja F-Measurea.

Yksi tapa toteuttaa malli on käyttää ristivalidointia (Deisenroth et al., 2020). Yksi tälläinen ristivalidoinnin malli on k-kertainen ristiinvalidointi. Tässä mallissa aineisto jaetaan k osaan, joista yhtä osaa käytetään testiaineistona $\mathcal{V}$ ja $k-1$ osaa koulutusaineistona $\mathcal{R}$. Tällöin käytettävästä koulutusaineistosta käytetään suurin osa mallin kouluttamiseen, mutta samasta aineistosta saadaan myös tarkastusaineisto. Ristiinvalidoinnissa käydään läpi kaikki mahdolliset k vaihtoehtoa valita testausaineisto eli jaetaan data-aineisto kahteen osaan $D = \mathcal{R} \cup \mathcal{V}$, missä $\mathcal{R} \cap \mathcal{V} = \emptyset$. Näiden k-suorituskerran muodostamien mallien suorituskyky tarkastellaan keskiarvona.



Kuva 3.1: Ristiinvalidoinnissa data-aineisto jaetaan k-osaan, missä $k-1$ osaa ovat koulutusaineistoa (harmaalla merkityt osuudet) ja yksi osa testausaineistoa (oranssilla merkitty osuus) (Deisenroth et al., 2020).

Mallin kouluttamisen jälkeen koulutusaineistolla $\mathcal{R}$ tarkastellaan koulutetun mallin f suorituskykyä testausaineiston $\mathcal{V}$ avulla (Deisenroth et al., 2020). Tälle halutaan laskea kes-

kineliövirheen neliöjuuri. Näin toteutettaessa saadaan jokaiselle k-osalle toteutettua malli $f^{(k)}$ koulutusaineiston $\mathcal{R}^{(k)}$ avulla. Näiden pohjalta voidaan laskea tarkastusaineiston $\mathcal{V}^{(k)}$ avulla empiirinen riski $R(f^{(k)}, \mathcal{V}^{(k)})$. Tämän avulla ristiinvalidointi pystyy arvioimaan odotetun yleistysvirheen

$$\mathbb{E}_{\mathcal{V}}[R(f, \mathcal{V})] \approx \frac{1}{K} \sum_{k=1}^{K} R(f^{(k)}, \mathcal{V}^{(k)}).$$

Prosessissa käytettävässä arvioinnissa on kaksi lähdettä, joista toinen on ettei rajatulla koulutusaineistolla välttämättä saada parasta mahdollista $f^{(k)}$ ja toinen on, ettei testausaineistolla saada tarkkaa arviota riskistä $R(f^{(k)}, \mathcal{V}^{(k)})$.

Mallia toteutettaessa on huomioitava myös ylisovittamisen vaara, jotta mallin tarkkuus ei kärsisi (Hämäläinen ja Vinni, 2010). Ylisovittamisessa malli on sovitettu koulutusaineistoon niin tarkasti, että se huomioi jopa kaikki erikoistapaukset sekä koulutusdatan virheet. Tämä ilmenee yleensä mallin ollessa liian monimutkainen suhteessa käytettävään data-aineiston kokoon. Mallin sovittamisessa on löydettävä sopiva taso, sillä liian yksinkertaisella mallilla se ei pysty välttämättä tulkitsemaan data-aineistoa ja täten malli on alisovitettu; se ei kuvaa todellisuutta tai kuvaa sitä todella vähän.

Yksi tapa tehdä luokittelua on käyttää naiivia Bayesin luokitinta (Natingga, 2018). Naiivi Bayesin luokitin pohjautuu Bayesin teoreemaan

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)},$$

missä A ja B ovat tapahtumia, $P(A)$ on todennäköisyys tapahtumalle A olla tosi ja $P(A|B)$ on ehdollinen todennäköisyys tapahtumalle A olla tosi, mikäli tapahtuma B on tosi. Naiivissa Bayesin luokittimessa datapisteiden joukolle annetaan luokka, joka on Bayesin teoreeman perusteella todennäköisin. Tämä tapahtuu laskemalla todennäköisyys sille, kuinka todennäköisesti asia A tapahtuu, jos ehto B saa tietyn arvon.

Bayesin teoreemaa voidaan hyödyntää myös useamman probabilistic eventsin tapauksessa, tällöin käytetään laajennettua Bayesin teoreemaa (Natingga, 2018). Jos määritellään tapahtumat $B_1, \ldots, B_n$ ovat ehdollisesti riippumattomia tapahtumasta A. Tällöin Bayesin teoreema voidaan esittää muodossa

$$P(A|B_1, \ldots, B_n) = \frac{P(B_1, \ldots, B_n|A) \cdot P(A)}{P(B_1, \ldots, B_n)}.$$

Nämä satunnaismuuttujat voivat olla diskreettejä tai jatkuvia seuraten todennäköisyysjakaumaa, kuten normaalijakaumaa.

Käytettäessä Bayesilaista todennäköisyyttä, pitää huomioida ovatko vertailtavat tapahtumat riippumattomia toisistaan (Natingga, 2018). Jos vertaillaan esimerkiksi lämpötilaa ja vuodenaikaa keskenään, niin havaitaan näiden välillä olevan riippuvuus: talvella on kylmää ja kesällä lämmintä. Tämä estää Bayesin teoreeman käyttämisen luokittelemiseen. Tällöin data-aineistoa voidaan kuitenkin hyödyntää tekemällä osittaista analyysiä niille tapahtumille, jotka eivät ole riippuvia toisistaan.

Barber ja Sharkey, 2012 omassa tutkimuksessaan yrittäessään tunnistaa opiskelijoita, jotka ovat vaarassa saada hylätyn kurssilta, jolle osallistuu. Toisena mallina tutkimuksessa käytettiin naiivia Bayesia ja kymmenkertaista ristivalidointia. Selittäville muuttujille oli annettu eri painoarvoja riippuen kurssin viikosta. Selittävinä muuttujina oli henkilöön liittyviä taustatietoja, suoritettujen opintopisteiden suhde yritettyihin opintopisteisiin sekä toimintaa verkko-oppimisympäristön keskustelualueella. Verrattuna logistiseen regressioon, lisättyjen selittävien muuttujien kanssa nähtiin kurssin viikolla 0 35 %-yksikön parannus ennustustarkkuudessa datamäärän ollessa pienempi ja eron kaventuessa huomattavasti lähemmäs toisiaan viikolla 3 datamäärän kasvettua, missä logistisella regressiolla keskimäärin 94% ennustuksista onnistui ja naiivilla Bayesillä 95% onnistui.

Useiden eri mallien välisessä vertailussa naiivi Bayes oli ennustustamisen osalta paras algoritmi (Kotsiantis et al., 2004).

Toinen mahdollisuus tehdä tilastollista analyysia kerätylle oppimisdatalle on regressioanalyysi (Song, 2018; Romero ja Ventura, 2010; Papamitsiou ja Economides, 2014). Regressioanalyysi voidaan tehdä usealla eri tavalla, joita ovat esimerkiksi yksinkertainen lineaarinen regressio, usean selittäjän lineaarinen regressio ja logistinen regressio. Regression avulla voidaan ennustaa kuinka opiskelija tulee menestymään eri selittävien muuttujien vaikutus huomioiden.

Lineaarinen regressio kuvaa selittävän muuttujan ja selitettävän muuttujan yhteyttä toisiinsa (Ross, 2017). Yksinkertaisessa lineaarisessa regressiosa selittäviä ja selitettäviä muuttujia on molempia yksi. Yksinkertainen lineaarinen regressio voidaan esittää kaavana

$$Y = \alpha + \beta x + e,$$

jossa x kuvaa selittävää muuttujaa ja y kuvaa selitettävää muuttujaa. Parametrit α ja β ovat tuntemattomia suureita, estimaattoreita, jotka estimoidaan datan perusteella. Muuttuja e kuvaa satunnaista virhettä, jolle yleensä tehdään olettamus sen noudattavan normaalijakaumaa odotus arvolla 0 ja varianssilla σ^2. Varianssin oletetaan olevan sama riippumatta selittävistä muuttujista x.

Parametrien α ja β estimointi voidaan tehdä pienimmän neliösumman estimoinnilla (Ross, 2017). Käytettäessä pienimmän neliösumman estimointia, halutaan löytää sellaiset arvot estimaateille α ja β, joilla virheen neliösumma $\sum_{i=1}^{n} \epsilon_i^2$ on mahdollisimman pieni. Pienimmän neliösumman estimaatit $\hat{\alpha}$ ja $\hat{\beta}$ parametreille α ja β saadaan laskettua kaavoista

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

ja

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x},$$

missä $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ ja $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$.

Estimoidussa regressioviivassa $y = \hat{\alpha} + \hat{\beta}x$ estimaatti $\hat{\alpha}$ kuvaa suoran kulmakerrointa ja estimaatti $\hat{\beta}$ kuvaa suoran vakiota, eli y-akselin kohtaa missä suora leikkaa y-akselin (Ross, 2017). Tämän estimoidun regressioviivan avulla voidaan ennustaa selitettävän muuttujan y arvoja käyttäen selittävän muuttujan x arvoja.

Yksinkertainen lineaarinen regressio voidaan laajentaa usean selittäjän lineaariseksi regressioksi, joka kuvaa kuinka useampi selittävä muuttuja x vaikuttaa selitettävään muuttujaan Y (Ross, 2017). Matemaattisena kaavana esitettynä tämä olisi

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e,$$

jossa Y on selitettävä muuttuja, ja x_i kuvaa selittäviä muuttujia, missä $i = 1, \cdots, k$. Regressioparametrejä yhtälössä kuvaa $\beta_0, \beta_1, \cdots, \beta_k$ ja satunnaisvirhettä e.

Kuten naiivin Bayesin osalta, täytyy myös regressiossa välttää sellaisia selittäviä muuttujia, jotka korreloivat keskenään; toisin sanoen näiden arvot ovat riippuvaisia toisistaan eivätkä täten ole riippumattomia (Daoud, 2017). Tätä ilmiötä kutsutaan multikollineaarisuudeksi. Ilmiö voidaan havaita tapauksissa, joissa tapahtuu suurta vaihtelua estimoitujen kertoimien osalta lisättäessä tai poistettaessa selittäviä muuttujia tai suurta vaihtelua kertoimissa muutettaessa tai poistettaessa yksittäisiä datapisteitä.

Yhdistelläksemme intervallilla liikkuvia muuttujia kategoristen muuttujien kanssa, kuten onko henkilöllä jokin tietty ominaisuus, voidaan käyttää dummy-muuttujia kuvaamaan näitä arvoja (Ross, 2017). Tämän avulla pystytään hyödyntämään sellaisia selittäviä muuttujia, jotka eivät ole lähtökohtaisesti numeerisessa muodossa. Jos esimerkiksi usean selittäjän lineaarisessa regressiossa muuttuja x_3 kuvaa onko opiskelija tutkinto-opiskelija, voitaisiin tämä esittää numeraalisessa muodossa seuraavasti:

$$x_3 = \begin{cases} 1 = \text{opiskelija on tutkinto-opiskelija} \\ 0 = \text{opiskelija ei ole tutkinto-opiskelija} \end{cases}.$$

Agudo-Peregrina et al., 2014 käytti tutkimuksessaan usean selittäjän lineaarista regressiota etsiessään eri relaatioita opiskelijan toiminnan verkko-oppimisympäristössä ja akateemisen menestyksen väliltä. Riippumattomia selittäviä muuttujia olivat jokainen eri tyyppinen interaktio verkko-oppimisympäristössä ja riippuvana selittävänä muuttujana akateeminen menestys oli esitetty jokaisen opiskelijan saamana kurssin päättöarvosanana. Tutkimuksessa löydettiin merkittäviä relaatioita eri tyyppisten opiskelijan interaktioiden ja akateemisen menestyksen väliltä.

3.2 Yksitäiseen oppijaan kohdennetut ehdotukset

1. arviointi- ja muiden seuraamisperiaatteiden muotoileminen malliksi
2. kurssiarvosanan ennustaminen kurssin edistyessä
3. suositeltavat jatkokurssit
4. Educational Data Mining EMD, ainakin (Romero ja Ventura, 2010)
5. etiikka?! (Kaila et al., 2019)

3.3 Ehdotuksien tulkinnan rajoitteet

1. virhearviot
2. yhden asian tajuamatta jääminen !== huono kurssimenestys
3. model bias
4. mallien todennäköisyydet - kuinka todennäköisesti tämä pitää paikkansa. voidaanko 72 prosentin todennäköisyyttä pitää sellaisena, että se toimii luotettavana ohjauksen työkaluna?

tilastollinen malli kuvaa optimia, ja verrataan kuinka data sopii tähän malliin

4 Yhteenveto

Lähteet

Agudo-Peregrina, Á. F., Iglesias-Pradas, S., Conde-González, M. Á. ja Hernández-García, Á. (helmikuu 2014). "Can We Predict Success from Log Data in VLEs? Classification of Interactions for Learning Analytics and Their Relation with Performance in VLE-supported F2F and Online Learning". *Computers in Human Behavior* 31, s. 542–550. ISSN: 0747-5632. DOI: 10.1016/j.chb.2013.05.031.

Akçapınar, G., Altun, A. ja Aşkar, P. (lokakuu 2019). "Using Learning Analytics to Develop Early-Warning System for at-Risk Students". *International Journal of Educational Technology in Higher Education* 16.1, s. 40. ISSN: 2365-9440. DOI: 10.1186/s41239-019-0172-z.

Barber, R. ja Sharkey, M. (huhtikuu 2012). "Course Correction: Using Analytics to Predict Course Success". *ACM International Conference Proceeding Series*. DOI: 10.1145/2330601.2330664.

Conde, M. Á., Hérnandez-García, Á., J. García-Peñalvo, F. ja Séin-Echaluce, M. L. (2015). "Exploring Student Interactions: Learning Analytics Tools for Student Tracking". Teoksessa: *Learning and Collaboration Technologies*. Toim. P. Zaphiris ja A. Ioannou. Cham: Springer International Publishing, s. 50–61. ISBN: 978-3-319-20609-7. DOI: 10.1007/978-3-319-20609-7_6.

Daoud, J. I. (joulukuu 2017). "Multicollinearity and Regression Analysis". *J. Phys.: Conf. Ser.* 949, s. 012009. ISSN: 1742-6588, 1742-6596. DOI: 10.1088/1742-6596/949/1/012009.

Deisenroth, M. P., Faisal, A. A. ja Ong, C. S. (2020). *Mathematics for Machine Learning*.

eoppimiskeskus (elokuu 2017). *Oppimisanalytiikka tulee – oletko valmis?*

Filvà, D. A., Guerrero, M. J. C. ja Forment, M. A. (kesäkuu 2014). "Google Analytics for Time Behavior Measurement in Moodle". Teoksessa: *2014 9th Iberian Conference on Information Systems and Technologies (CISTI)*, s. 1–6. DOI: 10.1109/CISTI.2014.6877095.

Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U. ja Sattar, M. U. (tammikuu 2020). "Predicting Student Performance in Higher Educational Institutions Using Video Learning Analytics and Data Mining Techniques". *Applied Sciences* 10.11, s. 3894. ISSN: 2076-3417. DOI: 10.3390/app10113894.

Hämäläinen, W. ja Vinni, M. (lokakuu 2010). ”Classifiers for Educational Data Mining”. Teoksessa: *Handbook of Educational Data Mining*. Toim. C. Romero, S. Ventura, M. Pechenizkiy ja R. Baker. Vol. 20103384. CRC Press, s. 57–74. ISBN: 978-1-4398-0457-5 978-1-4398-0458-2. DOI: 10.1201/b10274-7.

Kaila, E. T., Kurvinen, E. ja Apiola, M.-V. (2019). ”Ethical Considerations in Learning Analytics: Tethics”. *CEUR Workshop Proceedings* 2505, s. 61–63. ISSN: 1613-0073.

Kokkonen, H. (ei julkaisupäivää). ”Effects of Data Cleaning on Machine Learning Model Performance” (), s. 32.

Kotsiantis, S., Pierrakeas, C. ja Pintelas, P. (toukokuu 2004). ”PREDICTING STUDENTS’ PERFORMANCE IN DISTANCE LEARNING USING MACHINE LEARNING TECHNIQUES”. *Applied Artificial Intelligence* 18.5, s. 411–426. ISSN: 0883-9514, 1087-6545. DOI: 10.1080/08839510490442058.

Long, P. ja Siemens, G. (ei julkaisupäivää). ”Penetrating the Fog: Analytics in Learning and Education” (), s. 6.

Madden, M., Fox, S., Smith, A. ja Vitak, J. (joulukuu 2007). *Digital Footprints*.

Mikkola, J. (huhtikuu 2019). *Mitä oppimisanalytiikka on?* https://analytiikkaaly.fi/2019/04/04/mita-oppimisanalytiikka-on/.

Mwalumbwe, I. ja Mtebe, J. S. (2017). ”Using Learning Analytics to Predict Students’ Performance in Moodle Learning Management System: A Case of Mbeya University of Science and Technology”. *THE ELECTRONIC JOURNAL OF INFORMATION SYSTEMS IN DEVELOPING COUNTRIES* 79.1, s. 1–13. ISSN: 1681-4835. DOI: 10.1002/j.1681-4835.2017.tb00577.x.

Natingga, D. (2018). *Data Science Algorithms in a Week - Second Edition*. 2nd edition. Packt Publishing. ISBN: 1-78980-607-0.

Papamitsiou, Z. ja Economides, A. A. (2014). ”Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence”. *Journal of Educational Technology & Society* 17.4, s. 49–64. ISSN: 1176-3647.

Ranjeeth, S., Latchoumi, T. P. ja Paul, P. V. (tammikuu 2020). ”A Survey on Predictive Models of Learning Analytics”. *Procedia Computer Science*. International Conference on Computational Intelligence and Data Science 167, s. 37–46. ISSN: 1877-0509. DOI: 10.1016/j.procs.2020.03.180.

Romero, C., Romero, J. R. ja Ventura, S. (2014). ”A Survey on Pre-Processing Educational Data”. Teoksessa: *Educational Data Mining*. Toim. A. Peña-Ayala. Vol. 524. Cham: Springer International Publishing, s. 29–64. ISBN: 978-3-319-02737-1 978-3-319-02738-8. DOI: 10.1007/978-3-319-02738-8_2.

Romero, C. ja Ventura, S. (marraskuu 2010). ”Educational Data Mining: A Review of the State of the Art”. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 40.6, s. 601–618. ISSN: 1558-2442. DOI: 10.1109/TSMCC.2010.2053532.

Ross, S. M. (tammikuu 2017). ”Introductory Statistics”. Teoksessa: *Introductory Statistics (Fourth Edition)*. Toim. S. M. Ross. Oxford: Academic Press, s. 797–800. ISBN: 978-0-12-804317-2. DOI: 10.1016/B978-0-12-804317-2.00031-X.

Siemens, G. (lokakuu 2013). ”Learning Analytics: The Emergence of a Discipline”. *American Behavioral Scientist* 57.10, s. 1380–1400. ISSN: 0002-7642. DOI: 10.1177/0002764213498851.

Song, D. (heinäkuu 2018). ”Learning Analytics as an Educational Research Approach”. *INTERNATIONAL JOURNAL OF MULTIPLE RESEARCH APPROACHES* 10, s. 102–111. DOI: 10.29034/ijmra.v10n1a6.

Wolff, A., Zdrahal, Z., Nikolov, A. ja Pantucek, M. (2013). ”Improving Retention: Predicting at-Risk Students by Analysing Clicking Behaviour in a Virtual Learning Environment”. Teoksessa: *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK ’13*. Leuven, Belgium: ACM Press, s. 145. ISBN: 978-1-4503-1785-6. DOI: 10.1145/2460296.2460324.