DATAN MALLINTAMINEN: OPETUKSELLINEN TIEDONLOUHINTA

Jyväskylän yliopisto

TIIVISTELMÄ

Tekijä Miika Mikkonen, Jeremias Collianer

Työn nimi
Datan mallintaminen: opetuksellinen tiedonlouhinta harjoitustyö

Aika (pvm.) 3.27.2020 | Sivumäärä 14

Tiivistelmä

Tutkielmassa tarkastellaan millä tavoin voidaan ennustaa opiskelijoiden opintomenestystä tutkimalla heidän toimintojaan tietyn aktiviteetin aikana. Tutkielman tarkoitus on selvittää, onko mahdollista olemassa olevien datapisteiden avulla ennustaa opintomenestystä. Opintomenestystä tarkastellessa on tärkeää kiinnittää huomiota isompiin otoksiin perusjoukon sisällä ja kiinnittää huomiota näiden joukkojen välisiin eroavaisuuksiin, mieluummin kuin tekee vertailuja yksittäisten alkioiden välillä.
Tutkielmassa käydään läpi otosten välisiä eroavaisuuksia huonommin menestyneillä opiskelijoilla ja paremmin menestyneiden opiskelijoiden välillä. Opintomenestyksen ennustaminen on mahdollista, kun normalisoidaan tarpeellinen toimintojen määrä ja hajautetaan tämä toimintojen määrä sopivalle aikavälille, jotta liian suuria aikaeroja toimintojen väliin tai liian suuria toimintomääriä ei tule tunnin ajalle. Tämä auttaa pitämään opiskelijan toiminnat laadukkaina ja tarkoituksenomaisina, jotta liiallisia ”toimintarypäyksiä” ei pääse käymään.

Asiasanat: Opetuksellinen tiedonlouhinta, opintomenestys,

SISÄLLYS

KUVALUETTELO

TAULUKOT

1 JOHDANTO

Opetuksellista tiedonlouhintaa on pyritty hyödyntämään laajalti opintojen kehittämisessä. Opetuksellinen tiedonlouhinta tarkoittaa tiedonlouhinta tapojen keksimistä, luomista ja kehittämistä joiden avulla pyritään ymmärtämään oppimista ja sen menestystekijöitä. Tässä paperissa tiedonlouhintaa toteutetaan mallintamalla oppilas dataa, jonka pohjalta voimme kehittää ratkaisuja opintomenestyksen ennustamiseen.

Tutkimme paperissa Oppimisen hallintajärjestelmän (LMS) tuottamaa dataa, jossa näkyy opiskelijan pistemäärä kurssilla sekä aktiviteetin aikaleima. Aikaleima syntyy aina kun opiskelija on tehnyt, ”klikannut” (tutustunut, lukenut, yrittänyt palauttaa tai muuten vaan klikannut) jotain liittyen kurssilla suoritettavaan aktiviteettiin ”välikoe/vaikeahko harjoitustehtävä” jota voi tehdä/suorittaa useamman kerran LMS:ssä. Opiskelija on tunnistettavissa datassa näkyvästä pistemäärästä. Pistemäärää vastaa opiskelijan menestystä kurssilla (mitä korkeampi sitä paremmin opiskelija on menestynyt).

Oletuksena emme tutki korrelaatiota opiskelijan ”klikkailun” ja päivän ajankohdan (aamu, päivä, ilta, yö) välillä. Opiskelijoilla on mahdollisesti erilaiset päivärytmit, jotka vaikuttavat ”klikkailun” ajankohtaan. Kuitenkin otamme huomioon, millä aikaväleillä opiskelija on vuorovaikuttanut aktiviteetin kanssa ja kuinka pitkiä aikoja tai lähellä kurssin loppua etc. Kurssin kesto oletetaan datassa esiintyvien aikaleimojen perusteella ensimmäisestä aikaleimasta viimeiseen. Oletamme myös, että kaikki ”klikkaukset” ovat samanarvoisia koska emme voi tietää mitä kunkin aktiviteetin aikaleiman kohdalla on tehty. Tiedämme vain, että aktiviteetin kanssa on vuorovaikutettu ja kuka näin on tehnyt pistemäärän perusteella.

Pyrimme paperissa mallintamaan selkeästi aikaleimojen esiintyvyyden korrelaatiota opintomenestykseen. Tutkimme aikaleimojen aikavälejä, niiden keskihajontaa ja varianssia, esiintyvyyttä ja esiintyvyystiheyttä jne. Aikaleimojen avulla tutkimme myös, klikkaileeko opiskelija useampina päivinä vähän vai pyrkiikö opiskelija yhden päivän aikana tekemään kaiken työn ja kuinka nämä kaksi tapaa ennustavat opintomenestystä. Pyrimme pystyä mallien avulla selit-

tämään mikä näistä vaikuttajista on suurin vaikuttava tekijä opintomenestyksen kannalta.

Paperin ensimmäisessä osassa esitämme ratkaisuehdokkaat, joiden pohjalta luodaan mallit, joita käytetään datan korrelaatioiden esittämiseen sekä analysoimiseen. Mallien avulla pyrimme luomaan kuvaa siitä, kuinka vahvasti mikäkin datasta louhittu attribuutti vaikuttaa yhden opiskelijan opintomenestykseen. Toisessa esitämme millä muilla ratkaisutavoilla dataa voitaisiin tutkia ja kuinka niitä voi verrata käytettyihin tapoihin

Kolmannessa ja neljännessä kappaleessa vertailemme ratkaisutapoja ja esitämme kuinka laajalti löydetyt ratkaisut pätevät opetuksellisessa tiedonlouhinnassa, ja mitkä vaikuttavat näiden ratkaisujen esiintymiseen.

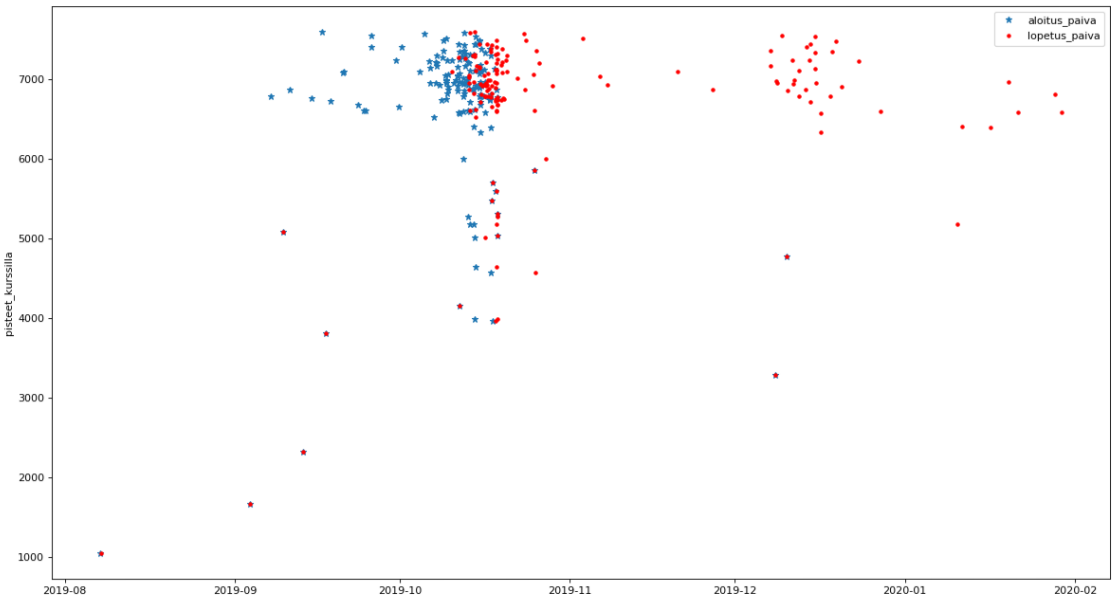
2 MALLIT

2.1 Datan alustaminen

Aineistossa esiintyi 6789 datapistettä ja 150 eri opiskelijaa perustuen opiskelijan pisteisiin kurssin aikana. Data on jaoteltu pisteiden mukaan 10-osa ja 4-osa fraktaaleihin. Päivämäärät on jaettu aikaan, tunteihin ja päivämääriin. Data on myös jaettu kahteen taulukkoon yhteen, jossa on edellä mainitut muuttujat sekä toiseen, jossa on opiskelijoiden pisteet ja heidän toimintojensa määrä. Näin saamme helpommin analysoitua aineiston dataa.

2.2 Aikajakso ja toimintojen määrä

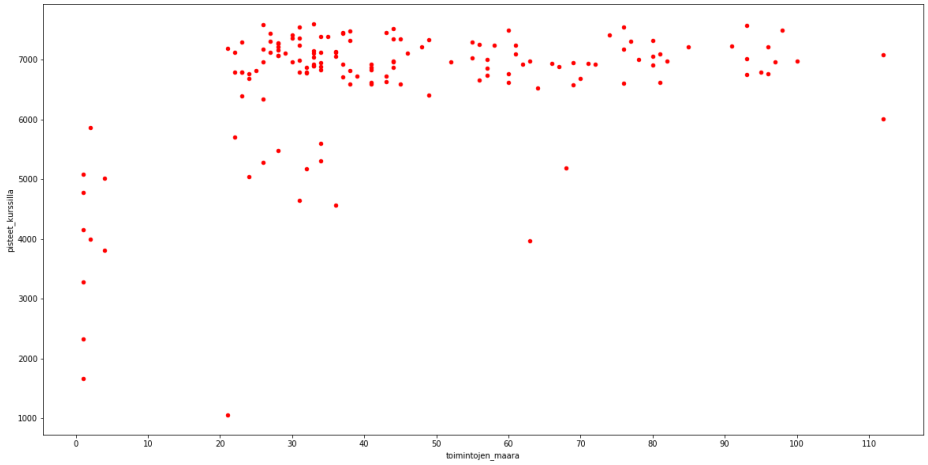
Lähdimme tutkimaan analysoitavaa dataa ottamalla yhden opiskelijan aloitus- ja lopetus päivämäärän ja katsomaan onko menestyneimpien ja vähemmän menestyneimpien opiskelijoiden pitkäjänteisyydellä jotain tekemistä menestyksen kanssa.



Kuva 1 näyttää opiskelijan toimintojen aloitus- ja lopetus päivämäärän.

Tutkimalla tätä pitkäjänteisyyttä huomasimme, että menestyksekkäimmät opiskelijat (top 25%) eivät kukaan klikkailut vain yhtenä päivänä. Siltikään vahvoja päätelmiä ei ole tehtävissä liittyen siihen onko parempi tehdä kauemmin vai vain kauemmin kuin yhden päivän. Kaavioissa syntyneet isot ryhmittymät johtuvat todennäköisesti välikoe/harjoitustehtävien julkaisupäivästä.

Toinen vaikuttava tekijä, jota halusimme tutkia, oli toimintojen määrä ja vaikuttaako toimintojen määrä itsessään menestykseen. Tutkimme tätä kahdelta kannalta. Ensin katsoimme jokaisen yksittäisen opiskelijan toimintojen määrää tutkimalla niiden aikaleimojen esiintyvyyttä aineistossa.



Kuva 2 opiskelijan toimintojen määrä

Yksi piste edustaa opiskelijaa ja hänen toimintojensa määrää. Malli, esittää että korrelaatiota toimintojen määrän ja menestyksen välillä olisi olemassa. Alle 20 toimintoa tehneitä opiskelijoita ei yksikään menestynyt kurssilla eikä päässyt yli aineiston pisteiden keskiarvon ~6663(*Taulukko 1.)*. Vaikka toimintojen määrä voi selittää menestyksen määrää. On myös hyvä muistaa, että menestykseen voi olla muita vaikuttavia tekijöitä. Toimintojen määrä voi myös olla harhaanjohtava sillä toimintoihin laskettiin kaikki opiskelijan ”klikkaukset” mukaan. Myös ”huti klikkaukset” joilla ei ole virkaa analysoidessa aineistoa.

Taulukko 1 opiskelijoiden pisteiden tunnusluvut.

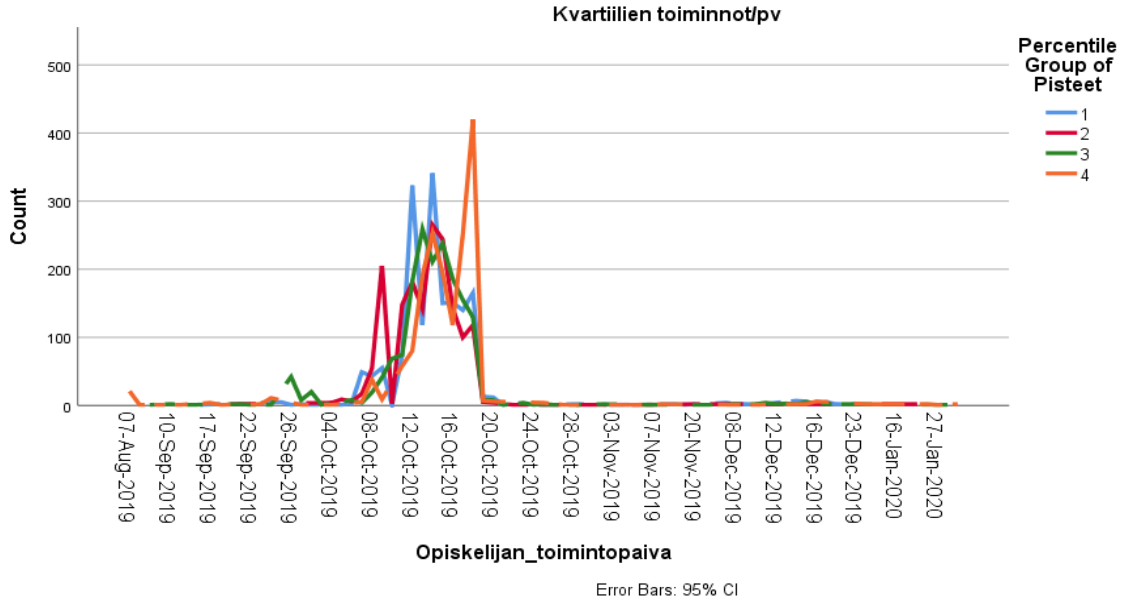
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Opiskelija	150	1055	7594	6636.26	1083.947
Valid N (listwise)	150				

2.3 Kvartiilien vertailu

Koska toimintojen määrä voi olla yksi mahdollinen selittävä tekijä opintomenestykselle niin tutkitaan sitä lisää. Jaoimme datan pisteiden mukaan kvartiileihin. Tämä mahdollistaa menestyksekkäimpien ja vähiten menestyksekkäimpien opiskelijoiden vertailun keskenään ja näin voidaan yleistää tuloksia paremmin kuin vain yksittäisten opiskelijoiden pisteiden kanssa. Kvartiili 1 (Q1) on par-

haiten menestynyt 25% ja alakvartiili (Q4) taas huonoimmat 25% Kvartiilien vertailu tehtiin niin että luotiin aikajana datapisteiden avulla, jonka voi olettaa edustavan kurssin pituutta, ja syötettiin toimintojen määrä y-akselille.



Kuva 3 kvartiilien välisten toimintojen määrän erot.

Näin saadaan näkymään kvartiilien päiväkohtaisten toimintojen määrien väliset erot. Tässä nähdään, että jokainen kvartiili on tehnyt eniten toimintoja lokakuun puolivälin aikoihin. Saadaan myös kuva siitä, että jokainen kvartiili on tehnyt toimintoja paljon lyhyen ajan sisään. Kuitenkin kvartiili 1 näyttää olevan hajautuneempi monelle päivälle kuitenkin pääosa toiminnoista keskittyy yhteen tai muutamaan päivään.

Taulukko 2 kvartiilien opiskelijoiden päivien jakauma, jolloin toimintoja on tehty.

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Opiskelijan_toimintopaiva * Percentile Group of Pisteet	6789	100.0%	0	0.0%	6789	100.0%

Report

Opiskelijan_toimintopaiva

Percentile Group of Pisteet	Mean	N	Std. Deviation	Last	First
1	15-OCT-2019	1696	9 12:19	19-OCT-2019	16-SEP-2019
2	13-OCT-2019	1675	7 14:21	18-OCT-2019	07-OCT-2019
3	14-OCT-2019	1710	9 00:01	19-OCT-2019	11-OCT-2019
4	15-OCT-2019	1708	12 20:22	07-AUG-2019	09-OCT-2019
Total	14-OCT-2019	6789	9 22:53	07-AUG-2019	16-SEP-2019

Datasta (Kuva 5) nähdään että poikkeava neljännes on alakvartiili, josta yksikään opiskelija ei päässyt yli keskiarvon (Kuva 6). Alakvartiilin päivien keskihajonta on huomattavasti suurempaa kuin muilla kvartiileilla joka osoittaa että alakvartiilissä olevat opiskelijat ovat ”klikkaileet” hajaantuneemmin. Tämä indikoisi, että alakvartiilin opiskelijoiden pisteet menevät lähempänä normaali käyrän mukaan kvartiileilla ei nähdä juurikaan eroja toimintojen määrän jakautumisen kanssa.

Taulukko 3 kvartiilien toimintojen erot

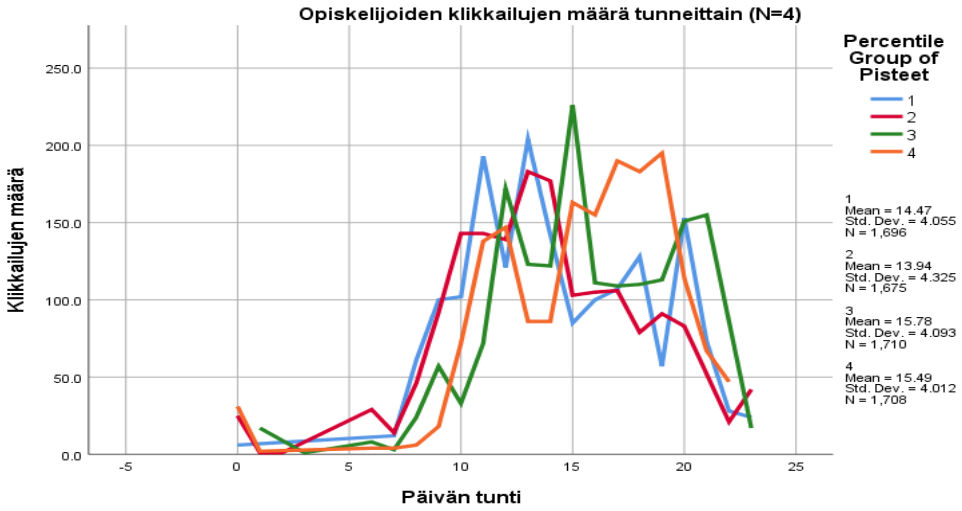
Report

Percentile Group of Opiskelija		Toiminnot	Opiskelija
1	Mean	48.43	7376.97
	N	37	37
	Std. Deviation	22.563	113.330
	Sum	1792	272948
	Minimum	23	7213
	Maximum	98	7594
2	Mean	50.66	7062.11
	N	38	38
	Std. Deviation	25.650	80.584
	Sum	1925	268360
	Minimum	21	6948
	Maximum	112	7211
3	Mean	47.66	6826.53
	N	38	38
	Std. Deviation	21.508	77.873
	Sum	1811	259408
	Minimum	22	6678
	Maximum	96	6942
4	Mean	34.08	5262.78
	N	37	37
	Std. Deviation	27.374	1450.804
	Sum	1261	194723
	Minimum	1	1055
	Maximum	112	6662
Total	Mean	45.26	6636.26
	N	150	150
	Std. Deviation	24.995	1083.947
	Sum	6789	995439
	Minimum	1	1055
	Maximum	112	7594

Tällä tiedolla vahvistetaan väittämää, että vaikkakin toimintojen määrä on vaikuttavana tekijänä opintomenestykseen niin se ei siltikään yksiselitteisesti ole menestystekijä.

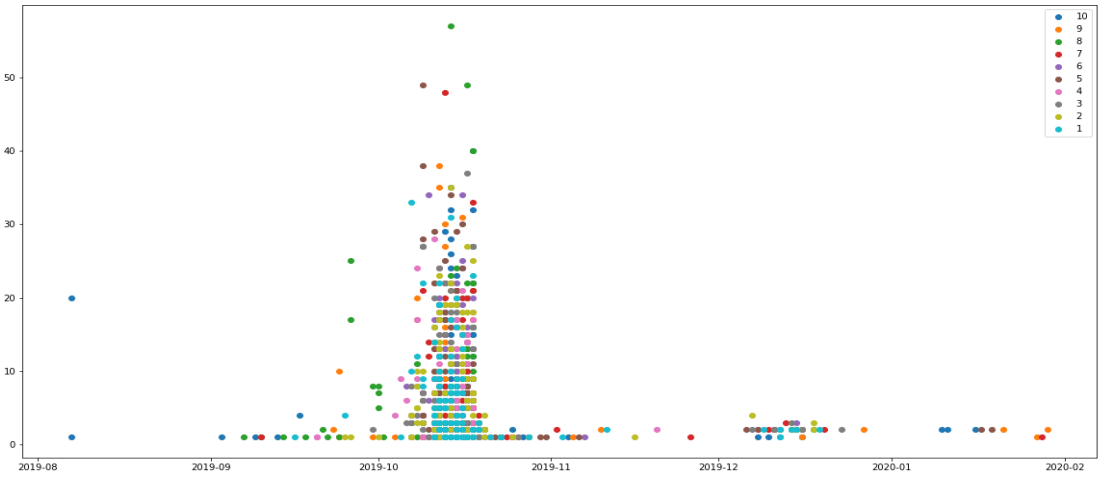
Koska toiminnot eivät näytä olevan laajalti jakaantuneet useammille päiville eri lailla kvartiilien väleillä (Kuva 5), niin lähdimme tutkimaan päivänsisäisiä ajankäytöllisiä eroja opiskelijoiden kesken. Tutkitaan näin ollen opiskelijan tekemien ”klikkauksien” määrää tuntikohtaisesti. Tutkitaan, onko vaikutusta sillä, kuinka paljon ”klikkauksia” opiskelija tekee tunnissa tai mihin aikaan päivästä.

Pidetään opiskelijat jakautuneena kvartiileihin ja johdetaan datasta aika muuttujasta päivien tunnit omaan muuttujaansa. Näiden tuntien perusteella meillä on jokaiselle datapisteelle oma tunti. Lasketaan kvartiilissa olevien tuntien esiintymät ja sijoitetaan ne päivän mittaiselle aikajanalle. (Kuva 7)



Kuva 4 kvartiilien toimintojen määrä tunneittain.

Kuvaajasta nähdään, että jokaisen kvartiilin tunnusnumerot ovat lähes samanlaiset ajankohdan puolesta. Eroavaisuutena jälleen nähdään alakvartiili (Q4). Nähdään että suurin osa työskentelystä tapahtuu klo 15 jälkeen sekä muutamaan tuntiin tehdään paljon ”klikkauksia”. Muiden kvartiilien puolesta nähdään, että suurempia huippuja toimintojen määrässä esiintyy mutta näiden välillä tapahtuu aina isompi tiputus toimintojen määrässä. Toimintojen määrän tippuminen voi kertoa, että opiskelijat pitävät taukoja tai tekevät jotain muuta opintoihin aktiviteettiin liittyvää/liittymätöntä toimintaa. Tiputus voi johtua myös siitä, että opiskelijat ovat eri päivinä tehneet eri aikavälein toimintoja.



Kuva 5 10-osa fraktaalien opiskelijoiden toimintojen määrä tunneittain. 1 = menestyneimmät 10%, 10 = huonoiten menesty-neet 10%.

Tätä kuvaaja tukee myös Kuva 8 jossa datapisteet ovat opiskelijan toimintojen määrä tunnin sisällä. Kuvaaja osoittaa, että menestyneimmät opiskelijat eivät tee paljoa toimintoja yhden tunnin sisällä vaan toimintojen määrä jakaantuu useammalle tunnille päivän/päivien aikana.

3 TULOKSET

Datasta pystyy johtamaan informaatiota siitä vaikuttaako toimintojen määrä ja/tai ajankohta opintomenestykseen. Pelkkä toimintojen määrä näyttää datan mukaan vaikuttavan opintomenestykseen jollain tasolla. Alle 20 toimintoa tehneistä opiskelijoista ei yksikään päässyt top 75% (Q3) pisteiden puolesta (Kuva 6), jossa minimi toimintojen määrä oli 22. Kuva 1 kuvaus pitkäjänteisyydestä näyttäisi myös, että mitä isompi väli aloituksen ja lopettamisen välillä on, niin sitä varmemmin pääset top 50% (Q2). Tämä korrelaatio voi johtua monesta tekijästä. Aloitus- ja lopetuspäivämäärä ei kerro meille aktiivisten päivien määrää, joka selittäisi enemmän vaikuttaako pitkäjänteisyys menestykseen. Aloitus- ja lopetuspäivämäärien välin korrelaatio voi myös johtua siitä, että opiskelija on voinut harjoitella aihetta aktiviteetin ulkopuolella, joka osoittaisi, että enemmän opiskelua toisi parempia tuloksia. Tästä johtuen ei tiedetä, johtuuko tämä korrelaatio oikeasti pitkäjänteisyydestä vai onko opiskelija vain viivyttänyt aktiviteetin palautusta pitemmälle aikavälille.

Vaikkakin kuva 1 ei tuota meille tarpeeksi informaatiota, joita voidaan päätellä voimme yhdistää siitä saatavaa informaatiota kuvan 4 ja 8 kanssa. Kuva 4 ja 8 osoittavat meille, että suurin osa toiminnoista tehtiin lokakuun puolenvälin aikana ja tarkemmin 8. – 20.10. Tällä aikavälillä tehtiin ~94% kaikista toiminnoista sekä jokaisen kvartaalin toiminnot kohdistuivat suurimmilta osin tälle aikavälille.

Taulukko 4 kvartiilien väliset toimintojen määrät aikavälillä 8.lokakuuta - 20.lokakuuta.

Opiskelijan_toimintopaiva * Percentile Group of Pisteet Crosstabulation

Count

		Percentile Group of Pisteet				
		1	2	3	4	Total
Opiskelijan_toimintopaiva	08-OCT-2019	43	54	19	38	154
	09-OCT-2019	55	205	41	9	310
	10-OCT-2019	1	3	69	0	73
	11-OCT-2019	74	148	74	58	354
	12-OCT-2019	323	181	182	80	766
	13-OCT-2019	118	142	259	191	710
	14-OCT-2019	341	265	211	253	1070
	15-OCT-2019	150	244	238	195	827
	16-OCT-2019	151	143	186	118	598
	17-OCT-2019	140	100	155	250	645
	18-OCT-2019	165	118	130	420	833
	19-OCT-2019	13	5	8	7	33
	20-OCT-2019	12	4	8	6	30
Total		1586	1612	1580	1625	6403

$$Q1 = \frac{1586}{1696} \approx 0.94$$

$$Q2 = \frac{1612}{1675} \approx 0.96$$

$$Q3 = \frac{1580}{1710} \approx 0.92$$

$$Q4 = \frac{1625}{1708} \approx 0.95$$

Tässä nähdään, että vaikkakin kuva 1 indikoisi että pitempi aikaväli on vaikuttavana tekijänä niin se ei vaikuta olevan merkittävin ennustava tekijä. Kuvassa olevat huomattavan ajan päässä tehdyt toiminnot voivat olla palautteen katsomista joka taas ei vaikuta opistomenestykseen kyseisellä kurssilla. Kuva 8 kuvaaja taas näyttää, että menestyneimmät opiskelijat eivät ole tehneet pääasiassa yli 20 toimintoa/tunti, joka kertoisi, että pelkästään ajankohta ja toimintojen määrä ei kerro meille kaikkea tarvittavaa tietoa. Kuva 8 kuitenkin osoittaisi, että tärkeimpänä tekijänä menestymiseen olisi menestyneimmät opiskelijat suorittavat toimintansa tasaisemmin ympäri päivä, joka taas voi tuottaa laadukkaampaa tulosta, kun taas heikommin menestyvät opiskelijat pyrkivät saada paljon tehtyä mahdollisimman lyhyessä ajassa. Datasta on mahdollista ennustaa menestymistä toimintojen määrään, aktiivisten päivien (ainakin yksi toiminta/pv) ja päivien sisäisen tuntien keskihajonnan kanssa.

3.1 Mitä eri lailla?

Datasta voitaisiin tutkia vielä tarkemmin aikaleimojen välisiä eroja. Ratkaisuistamme käy kyllä ilmi, että aikaleimojen hajonnaisuus pidemmälle ajalle on parempi kun se että kaikki aikaleimat ovat lyhyen ajan sisällä. Aikaleimat voisivat kertoa, onko olemassa jonkinlaista optimaalista aikaeroa päivien sisässä. Raapaisimme vain pintaa aktiivisten päivien tutkimisessa. Datasta pystyisi yksityiskohtaisemmin tarkastelemaan onko menestyneimpien oppilaiden aktiivisten päivien välillä suurempia tai pienempiä eroja kuin huonommin menestyneillä.

LÄHTEET

- Aldowah, Hanan; Al-Samarraie, Hosam; Fauzy, Wan Mohamad Julkaisussa: Telematics and Informatics April 2019, Vol.37, pp.13-49
- Bakhshinategh, Behdad; Zaiane, Osmar; ElAtia, Samira; Ipperciel, Donald, Julkaisussa: Education and Information Technologies 2018, Vol.23(1), pp.537-553
- Piety, Philip; Hickey, Daniel; Bishop, M, 2014. Proceedings of the Fourth International Conference on Learning Analytics And Knowledge. Association for Computing Machinery, New York, NY, USA.

LIITTEET

- Python_koodit.html