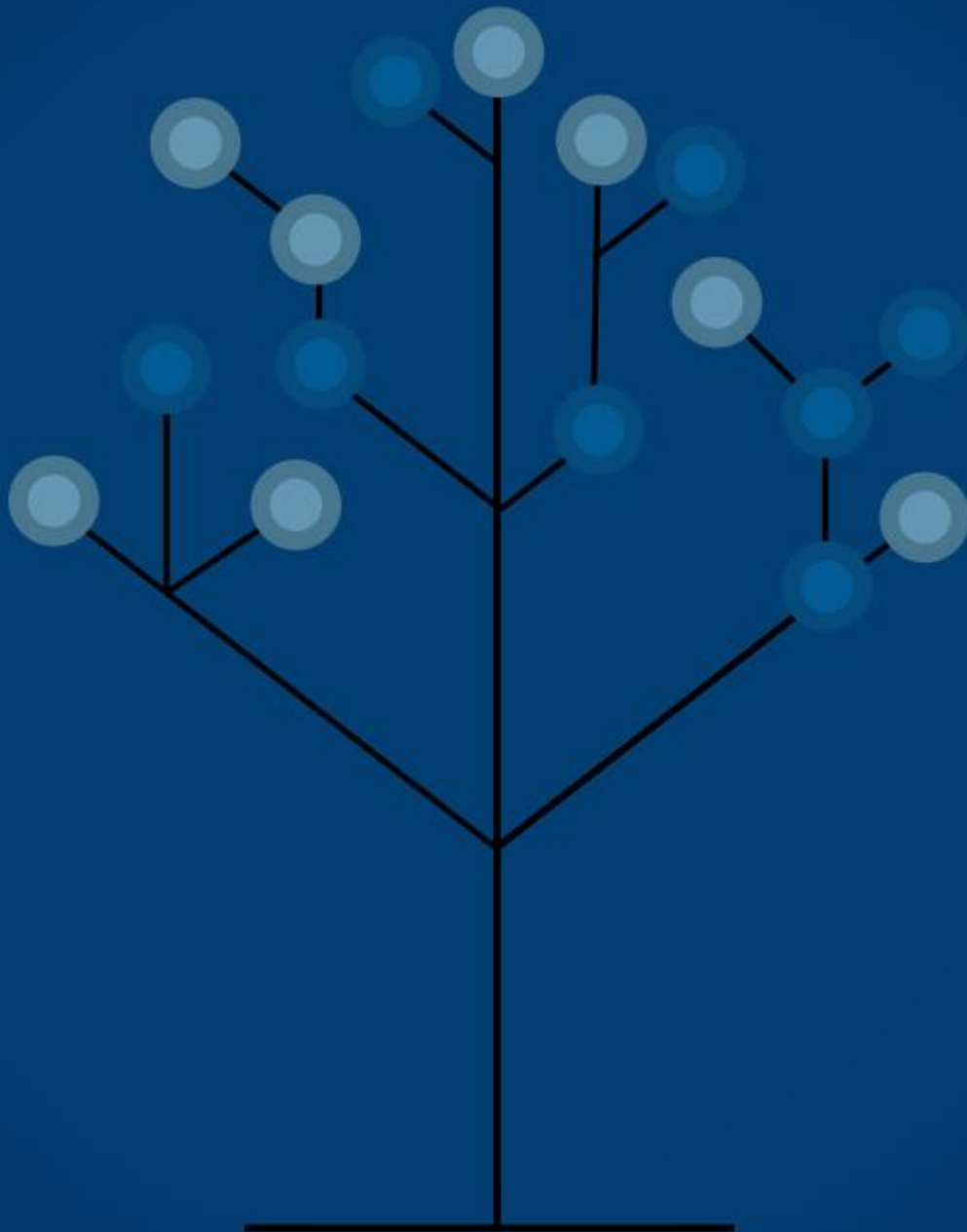


GRAVEN NAAR SUCCES

Educational Data Mining voor het Voorspellen van
Studiesucces: Een Exploratief Onderzoek



Damion E. Verboom
6 januari 2020

GRAVEN NAAR SUCCES

Educational Data Mining voor het Voorspellen van Studiesucces: Een Exploratief Onderzoek

HOGESCHOOL SAXION

Academy Mens & Arbeid, Lectoraat Brain & Technology

Scriptietraject

“Toegepaste Psychologie”

AFSTUDEER SCRIPTIE

Auteur:

Verboom, Damion E.

Achternaam, Voornaam & -Letter

426731

Student nummer

Eerste begeleider:

Farfan, M. A.

Achternaam, Voorletter(s)

Docent Onderzoeker, Drs.

Positie, Titel

Tweede begeleider:

de Graaf, J. W.

Achternaam, Voorletter(s)

Lector Brain & Technology, PhD

Positie, Titel

Plaats & Datum:

Deventer, 6 januari 2020

Voorwoord

Voor u ligt de scriptie “Educational Data Mining voor het Voorspellen van Studiesucces: Een Exploratief Onderzoek.” De scriptie is geschreven als onderdeel van het afstudeertraject van de opleiding Toegepaste Psychologie aan Hogeschool Saxion te Deventer. De scriptie is de laatste proef van deze opleiding en dient aan te tonen dat ik beschik over hbo werk- en denkniveau.

Naast dat het een toepassing van vaardigheden was, was het een geweldig leerproces. Om met de data te werken heb ik gebruik gemaakt van Python, een programmeertaal. En hoewel ik daar al bekend mee was, kon ik daar nog veel in leren. Dankzij deze ervaring beschik ik nu over de vaardigheden die in de toekomst zeer goed van pas zullen komen. Ook was de scriptie voor mij een eerste stap in het veld van Educational Data Mining, wat het zeer uitdagend maakte aangezien ik er nog veel in kon (en kan) leren.

Dat ik deze scriptie heb kunnen schrijven heb ik grotendeels aan Marco Farfan te danken. Hij was een prettige begeleider die veel mij veel ruimte en vrijheid gaf om mij in het veld van EDM te kunnen verdiepen. Ook bood Marco mij een werkplek aan in zijn kantoor om aan het onderzoek te kunnen werken. Hierdoor heb ik veel geleerd over de gang van zaken binnen de hogeschool op operationeel, tactisch en strategisch niveau. Ook bedank ik Jan Willem de Graaf voor zijn feedback om de scriptie te kunnen verbeteren. Tot slot gaat mijn dank uit naar Ana Daniela Rebelo voor haar hulp in het ontwerpen van de omslag.

Damion Verboom

Amersfoort, 29 december 2019

Inhoudsopgave

Samenvatting	5
Hoofdstuk 1. Inleiding van het onderzoek	6
1.1 Aanleiding van het onderzoek	6
1.2 Onderzoeksvraag	8
1.3 Doelstelling van het onderzoek	9
1.3.1 Omschrijving van de opdrachtgever	9
1.3.2 Doelstelling van het onderzoek	9
Hoofdstuk 2. Theoretisch kader	10
2.1 Onderwijs	10
2.2 Educational Data Mining	13
2.2.1 Classification	16
2.2.1.1 Decision trees	19
2.2.1.2 Bayesian networks	20
2.2.2 Voorspellen van studiesucces in EDM	20
2.3 Studiesucces	22
2.3.1 Voorspellers van studiesucces	23
2.4 Variabelen en het conceptueel model	25
Hoofdstuk 3. Onderzoeksdesign	27
3.1 Onderzoeksmethode	27
3.2 Onderzoeksdoelgroep	29
3.3 Onderzoeksinstrumenten	30
3.4 Procedure	30
3.4.1 Data understanding	30
3.4.2 Data preparation	31
3.4.3 Modelling en deployment	31
3.5 Analyses	31
Hoofdstuk 4. Onderzoeksresultaten	33
4.1 Uitvoering	33
4.2 Resultaten per deelvraag	36
4.2.1 Deelvraag 1	36
4.2.2 Deelvraag 2	38
4.2.3 Deelvraag 3	39
4.2.4 Deelvraag 4	40

Hoofdstuk 5. Conclusie, discussie en aanbevelingen	46
5.1 Conclusie en discussie	46
5.1.1 Antwoord op de hoofdvraag en deelvragen	46
5.1.2 Discussie	48
5.1.3 Betrouwbaarheid, validiteit en bruikbaarheid	49
5.2 Aanbevelingen	50
Literatuurlijst	53
Bijlage A: Feature selection	58
Bijlage B: Data preparation	60
Bijlage C: Resultaten modelling	74
Bijlage D: Dataverwerking met Python	76
Bijlage E: Decision trees	81

Samenvatting

Het doel van dit onderzoek is het ontwikkelen van een model op basis van technieken uit het veld Educational Data Mining (EDM) die door individuen zowel bekend als onbekend met het veld (EDM) in gebruik genomen kan worden om studiesucces (al dan niet doorstromen naar het tweede leerjaar) mee te kunnen voorspellen. Er werd antwoord gezocht op de vraag in hoeverre de beschikbare data van Hogeschool Saxion daartoe in staat was.

Om antwoord op deze vraag te geven werden gegevens afkomstig uit het intakeassessment (algemeen, persoonlijkheid en capaciteiten), Blackboard Learn (klik-gedrag in de cursussen) en Bison (toets-cijfers) verzameld van eerstejaars studenten van de opleiding Toegepaste Psychologie ($n = 812$) van cohorten 2016-2017, 2017-2018 en 2018-2019. Vervolgens werden op basis van feature selection variabelen geselecteerd voor diverse decision tree algoritmes (genaamd C5.0, CART, CHAID, Exhaustive CHAID en QUEST) om modellen te ontwikkelen. De decision trees werden beoordeeld en geëvalueerd op basis van het aantal correcte voorspellingen.

Uit de resultaten blijkt het mogelijk om voor het merendeel van de studenten studiesucces correct te kunnen voorspellen (73.74%). Er blijkt verder ook dat de belangrijkste voorspellers toets-cijfers aan het begin van het leerjaar en online klik-gedrag direct voor de toets-periodes de belangrijkste voorspellers zijn, respectievelijk afkomstig uit Blackboard en Bison. Met het intakeassessment kan geen studiesucces voorspeld worden.

Tot slot werd bediscussieerd dat de data niet optimaal benut kon worden, waardoor een model dat met een hogere mate van zekerheid in staat is studiesucces te voorspellen achterwegen blijft. Dit wijst erop dat acties ondernomen moeten worden om de aanwezige gegevens te optimaliseren voordat men nader gebruik maakt van EDM.

Hoofdstuk 1. Inleiding van het Onderzoek

Hoofdstuk 1 betreft de aanleiding, probleemstelling en mogelijke oplossingen (paragraaf 1.1), onderzoeksvraag inclusief deelvragen (paragraaf 1.2) en de doelstelling van dit onderzoek (paragraaf 1.3).

1.1 Aanleiding van het onderzoek

In het hoger onderwijs wordt veel aandacht besteed aan studiesucces. In 2016 verscheen de evaluatie van de in 2012 geïntroduceerde landelijke prestatieafspraken omtrent de kwaliteit van onderwijs, waaruit bleek dat met name hogescholen hun doelen op het aspect studiesucces (uitval, switch en het bachelor rendement) niet realiseerden (Bussemaker, 2016). Dit is problematisch voor zowel studenten als hogescholen. Voor studenten worden de kansen op de arbeidsmarkt vergroot indien zij de opleiding binnen de nominale studieduur plus één jaar afronden bij dezelfde instelling (bachelor rendement, Onderwijsinspectie, 2019). Voor hogescholen is studiesucces van belang vanwege de jaarlijkse financiële bijdrage vanuit het Rijk, genaamd de lumpsum, welke bestaat uit een vast en variabel bedrag, waarvan de laatste afhankelijk is van studiesucces. Daarnaast is het de maatschappelijke taak van het hoger onderwijs kwalitatief hoogwaardig onderwijs aan te bieden en is het de grootste leverancier van hoog opgeleide professionals in Nederland (Jonkman, z.d.).

De opdrachtgever, de Academie van Mens & Arbeid (AMA) binnen hogeschool Saxion te Deventer, heeft de afgelopen jaren diverse onderzoeken uitgevoerd aangaande voorspellers en correlaten van studiesucces van studenten van de opleiding toegepaste psychologie (TP-studenten) in Deventer (Ambagts, 2018; Jansen, 2019; Wilmer, 2019). De resultaten, in termen van uitval en bachelor rendement, die de onderzochte studenten behalen zijn namelijk lager dan Saxion nastreeft. Zo stroomt slechts 55% van de eerstejaars TP-studenten door naar het tweede leerjaar (Saxion, 2019), wat 20% minder is dan het landelijke gemiddelde (Vereniging Hogescholen, 2018; Onderwijsinspectie, 2019). Verder blijkt ook dat het bachelor rendement van TP-studenten gemiddeld 39% is (Ambagts, 2018), terwijl het Saxion-brede gemiddelde bachelor rendement rond 60% ligt (Saxion, 2018).

Zoals hiervoor genoemd heeft de AMA diverse onderzoeken uitgevoerd in het kader van studiesucces van TP-studenten. Zo werd door Ambagts (2018) onderzocht in hoeverre studiesucces voorspelt kon worden op basis van demografische gegevens en het intakeassessment. Het intakeassessment is in de basis een vragenlijst waarmee persoonlijkheidskenmerken, capaciteiten en algemene gegevens (zoals wat de studenten ertoe geleidt heeft zich voor de opleiding aan te melden) in kaart worden gebracht. Uit het onderzoek

bleken deze gegevens onvoldoende over voorspellende waarde te beschikken. De aanbeveling luidde om in vervolg onderzoek meer data en data bronnen met elkaar te combineren en opnieuw de voorspellende waarde te analyseren. Ook Wilmer (2019) maakte in haar onderzoek gebruik van de gegevens vanuit het intakeassessment en onderzocht de samenhang ervan met studiesucces. Wederom kon studiesucces onvoldoende verklaard worden met de gegevens vanuit het intakeassessment. Uit haar onderzoek (Wilmer, 2019) bleek echter wel dat op basis van geslacht en studiemotivatie (gemeten met het intakeassessment) risico studenten gesignaleerd kunnen worden. Door de onderzoeker werd aanbevolen meer data met aanvullende voorspellers te gebruiken. Tot slot maakte Jansen (2019) gebruik van andere gegevens, namelijk data afkomstig uit de data base van de digitale leeromgeving Blackboard Learn, om studiesucces te voorspellen. Tevens bleek dat deze gegevens afzonderlijk onvoldoende in staat waren studiesucces te verklaren en werd aangeraden alle beschikbare data met elkaar te combineren.

Niet alleen binnen Saxion, maar ook wereldwijd wordt geregeld onderzoek gedaan naar voorspellers van studiesucces, zoals bijvoorbeeld persoonlijke, academische en curriculaire factoren. Op basis van de resultaten is men echter vaak onvoldoende in staat exact aan te wijzen om welke studenten het gaat (Baars, Stijnen, & Splinter, 2017). Hetzelfde probleem is dus ook bij Saxion aanwezig, zoals blijkt uit het feit dat de Saxion-brede ingezette interventies tot op heden onvoldoende in staat zijn het succes van de TP-studenten naar het gewenste niveau te krijgen. Daarom wilt de AMA uiteindelijk met de uitkomsten van dit onderzoek interventies in kunnen zetten om de prestaties van de TP-studenten te kunnen verbeteren, ten einde het bachelor rendement te vergroten van de opleiding TP. Echter, zoals blijkt uit Helal et al. (2019), wilt men een interventie vroegtijdig en effectief inzetten, dienen deze studenten en de relevante kenmerken en factoren wel eerst geïdentificeerd te worden.

In de eerdere onderzoeken (Ambagts, 2018; Wilmer, 2019; Jansen, 2019) werd daarom aangeraden de diverse databronnen met elkaar te combineren, aangezien de data bases van de AMA afzonderlijk van elkaar onvoldoende in staat zijn studiesucces te voorspellen. Het combineren leidt tot een completer beeld van de student. De omvang die het combineren van de data vervolgens met zich mee brengt maakt het echter zeer complex en tijdrovend om mee te werken. Om met deze grote hoeveelheid data toch een model te kunnen ontwikkelen, dient Educational Data Mining (EDM) ingezet te worden, aangezien dit precies is waar de kracht van EDM ligt. EDM is in staat nieuwe patronen en voorspellers te ontdekken, waartoe men zelf binnen een redelijk tijdsbestek niet in staat is. Daarnaast kan met EDM een vroegtijdig waarschuwingssysteem voor al dan niet succesvolle studenten ontwikkeld worden in de vorm

van een door diverse partijen gemakkelijk te interpreteren en inzetbaar model (Heppen & Therriault, 2008). EDM biedt zo de mogelijkheid om tot nieuwe inzichten te komen. Het wordt in de literatuur omschreven als het toepassen van een verzamelingen van technieken voor het ontdekken van patronen in een reeds aanwezig collectie van data binnen een onderwijssetting (Brookshear, 2007; Fernandes et al., 2019).

Anderen (Baker & Yacef, 2009; Romero, Ventura, Pechenizkiy, & Bakar 2010; Abu Tair & El-Halees, 2012; Saa, 2019) waren al succesvol in het toepassen van EDM, waarbij de meest voorkomende, effectieve en ingezette techniek classificatie is, waarmee bijvoorbeeld decision trees (besluitbomen) gegenereerd kunnen worden. Deze bieden het voordeel dat zij voor zowel deskundigen als ondeskundigen gemakkelijk te interpreteren zijn. Het wordt daarom regelmatig ingezet om voorspellingen te doen over toekomstige gevallen. Het lijkt een zeer geschikte methode voor de AMA om student prestaties mee in kaart te brengen en voorspellen (Baradwaj & Pal, 2011; Kabakchieva, 2013; Alharbi, Cornford, Dolder, & De La Iglesia, 2016; Márquez-Vera, Cano, Romero, Noaman, Mousa Fardoun, & Ventrúa, 2016).

1.2 Onderzoeksvraag

Het bovenstaande leidt ertoe dat onderzocht wordt of het combineren van alle beschikbare gegevens het mogelijk maakt een model te ontwikkelen waarmee studiesucces nauwkeurig voorspelt kan worden. Hiervoor worden de gegevens van eerstejaarsstudenten aan de opleiding Toegepaste Psychologie (TP-studenten) gebruikt. De hoofdvraag luidt vervolgens:

- In welke mate kan de beschikbare data studiesucces van TP-studenten voorspellen?

Aangezien de data waar de AMA over beschikt afkomstig is uit Blackboard Learn, het intakeassessment en Bison, worden om de hoofdvraag te beantwoorden de volgende deelvragen geformuleerd:

1. In hoeverre kan data van Blackboard Learn ingezet worden om studiesucces van TP-studenten te voorspellen?
2. In hoeverre kan data van het intakeassessment ingezet worden om studiesucces van TP-studenten te voorspellen?
3. In hoeverre kan de data van Bison ingezet worden om studiesucces van TP-studenten te voorspellen?
4. In welke mate kan de data met elkaar gecombineerd studiesucces van TP-studenten voorspellen?

Na het operationaliseren in hoofdstuk 2 zal een keuze gemaakt worden studiesucces te definiëren als het gemiddelde gewogen cijfer (GPA), aantal behaalde studiepunten (EC) of uitval dan wel persisteren (Van Rooij et al., 2017)

1.3 Doelstelling en opdrachtgever

Hieronder wordt eerst de opdrachtgever aan de hand van haar functie, doelen en structuur omschreven (sub-paragraaf 1.3.1). Daarna volgt het doel en de praktische implicatie ervan (sub-paragraaf 1.3.2).

1.3.1 Omschrijving van de opdrachtgever

De Academie Mens en Arbeid (AMA) is een academie binnen Hogeschool Saxion, te vinden in zowel Deventer als Enschede, wie als kennisorganisatie actief is op het terrein van de mens, arbeid en organisatie. Zij is actief in het delen van kennis en ervaringen tussen zowel studenten als organisaties, onder andere in de vorm van stages en afstuderen.

Naast deze vorm van participatie met het bedrijfslevens, voert de AMA praktijkgericht onderzoek uit met inzet van lectoren, docent-onderzoekers en studenten. Het Lectoraat Brain & Technology besteedt aandacht aan het ontwerpen en toepassen van innovaties vanuit de psychologie en techniek. Het Lectoraat Strategisch HRM doet onderzoek naar onder andere de gebieden personeelsbeleid en leiderschap in de mkb-sector, werknemers inzetbaarheid, ondernemend gedrag en de aantrekkelijkheid van bedrijven voor starters op de arbeidsmarkt. Het lectoraat dat is opgericht ten einde bij te dragen aan de ontwikkelingen van de Smart Industry wordt het Lectoraat Smart Industry & Human Capital genoemd. Ook bestaat de AMA nog uit het Bureau Psychodiagnostiek (BPD). Het BPD is actief op het gebied van psychologisch onderzoek, biedt dyslexieonderzoek aan en heeft een eigen test-o-theek waar studenten en docenten terecht kunnen voor testmaterialen. Tot slot verzorgt de AMA onderwijs op hbo- en post-hbo niveau, zoals (International) Human Resource Management en Toegepaste Psychologie.

1.3.2 Doelstelling van het onderzoek

Het doel van dit onderzoek is het ontwikkelen van een model met Educational Data Mining (EDM) die door individuen zowel bekend als onbekend met het veld EDM in gebruik genomen kan worden om studiesucces van de TP-studenten mee te kunnen voorspelen. Op basis van de gevonden resultaten wordt de AMA geadviseerd over het gebruik maken ervan in de vorm van een adviesrapport.

Hoofdstuk 2. Theoretisch kader

In dit hoofdstuk wordt aandacht besteed aan de theoretische achtergrond van het onderzoek. Dit wordt gedaan door eerst de context te beschrijven waarbinnen het onderzoek plaatsvindt, namelijk: onderwijs. Vervolgens wordt omschreven welke rol Educational Data Mining daarin speelt door na te gaan *wat* het is en *hoe* het gebruikt wordt. Tot slot wordt ingegaan op de theorie achter studiesucces. Het hoofdstuk wordt afgesloten met een paragraaf waarin op basis van de theorie een raamwerk voor het onderzoek wordt geschetst (paragraaf 2.4).

2.1 Onderwijs

Zoals in de inleiding van het hoofdstuk vermeld staat, wordt eerst stilgestaan bij de context waarbinnen het onderzoek valt, namelijk: onderwijs, of meer specifiek, het hoger onderwijs. In Nederland wordt in het hoger onderwijs over het algemeen onderscheid gemaakt tussen het hoger beroepsonderwijs (hbo) en wetenschappelijk onderwijs (wo). In vergelijking met het hbo is het materiaal dat wordt aangeboden in het wo vaak abstracter, de onderwijsvorm is minder praktisch en van studenten wordt een hoger studietempo en zelfstandig leren verwacht. De curricula in het hbo kenmerken zich daarentegen aan hun meer beroepsgerichte karakter. Stage is vaak onderdeel van het curriculum, wat in het wo veel minder vaak voorkomt. Daarnaast zijn de curricula merendeels ontworpen om studenten te trainen voor een specifiek beroep (Van Rooij et al., 2017).

De opleiding Toegepaste Psychologie

De opleiding Toegepaste Psychologie (TP) leidt studenten op tot toegepast psycholoog. Als afgestuurde toegepast psycholoog wordt de student in staat geacht psychologische kennis toe te passen om gedrag van mensen te kunnen beïnvloeden en veranderen. Zo kan een toegepast psycholoog bijvoorbeeld werken bij instellingen in de lichamelijke en geestelijke gezondheidszorg, op het snijvlak van mens en technologie (zoals ondersteuning bij ontwikkeling van apps) of als psychologisch medewerker, (cognitief) trainer, preventiewerker of toegepast onderzoeker (Saxion, z.d.-a). Dit tracht de opleiding te verwerkelijken door het studieprogramma te organiseren in functie van de ontwikkeling van beroeps-specifieke en algemene competenties (Saxion, z.d.-b). De beroeps-specifieke competenties zijn:

- Diagnostisch Onderzoek (DO)
- Science, Technology, Engineering & Mathematics (STEM)
- Professionele Gespreksvoering & Training (PGT)

- Praktijkgericht Onderzoek (PRO)

Onder de algemene competenties wordt verstaan:

- Creatief handelen in complexe situaties
- Probleemoplossend werken
- Methodisch & reflectief denken en handelen
- Maatschappelijk verantwoord en ethisch handelen
- Kritisch denken, sociaal communicatief handelen
- Sensitief handelen
- ICT-geletterdheid
- Proactief en ondernemend handelen

Voor de opleiding Toegepaste Psychologie (TP) geldt een nominale studielast van 60 studiepunten (EC) per studiejaar. Wanneer 240 studiepunten aan de opleiding zijn behaald, ontvangt de student het bachelor diploma. De studie is opgedeeld in twee fases: de propedeutische fase (eerste studiejaar; 60 EC) en postpropedeutische fase (180 EC). Het propedeusejaar (propedeutische fase) is bedoeld voor oriëntatie, kennis en selectie. Het resterende studieprogramma valt onder de postpropedeutische fase en is bedoeld voor verdieping en verbreding.

De opleiding is per 1 september van 2016 van een nieuw curriculum voorzien (Saxion, 2019), waardoor de beschikbare gegevens van studenten van studiejaar 2014-2015 en 2015-2016 moeilijk zijn te vergelijken met die van daarna (zij wijken te veel af). In paragraaf 3.2 en bijlage B wordt hier nader op ingegaan.

Onderwijsvorm

Het materiaal dat tijdens de opleiding wordt aangeboden dient het doel de student de algemene en beroeps-specifieke competenties eigen te maken. De voortgang hierin wordt bewaakt door de studenten te toetsen. Toetsen worden beoordeeld in de vorm van een cijfer, waarbij de student minimaal een 5.5 dient te behalen om te slagen voor de toets. Het slagen voor een toets levert vervolgens studiepunten op. Toetsing vindt plaats in de vorm van werkstukken (zoals verslagen en portfolio's), digitale (kennis)toetsen en assessments. Verder geldt voor een aantal toetsen dat het eindcijfer bepaald wordt door de resultaten op deelttoetsen. Zo wordt het cijfer voor het vak Practice Based Learning 1 bepaald op basis van de resultaten op de drie deelttoetsen van het vak (zie tabel 2.1). De toetsing vindt (over het algemeen) plaats aan het

eind van elk kwartiel. Een kwartiel is een periode van 10 lesweken, waarvan er per studiejaar vier zijn (elk semester kent twee kwartielen). Dit levert een totaal van vier kwartielen op.

Om studenten voor te bereiden op de toetsen, wordt onderwijs aangeboden in de vorm van blended learning: het integreren van (synchrone) klassikale face-to-face leerervaringen met (asynchrone) leeractiviteiten via een online platform (Garrison & Kanuka, 2004). De synchrone leeractiviteiten betreffen de hoor- en werkcolleges van de opleiding. Dit betreft wekelijks gemiddeld 20 uur (Saxion, z.d.-a). De asynchrone leeractiviteiten vinden plaats via het digitale leerplatform Blackboard Learn (Bb). Dit is een Learning Management System (LMS) waarop cursus content wordt aangeboden (Romero et al., 2010). Op Bb kunnen studenten aanvullende (cursus relevante) informatie vinden, opdrachten maken, opdrachten inleveren en communiceren met elkaar en de docenten. Met de inzet van een LMS is het mogelijk het online gedrag van studenten te inventariseren (Conijn, Kleingeld, Matzat, Snijders, & Van Zaanen, 2016). Naast de 20 contacturen wordt verwacht dat studenten wekelijks 15 uur zelfstandig aan de studie besteden (Saxion, z.d.-a). In tabel 2.1 te zien welke vakken de studenten tijdens de propedeutische fase aangeboden krijgen.

Tabel 2.1

Overzicht van aangeboden toetsen in het eerste studiejaar van de opleiding toegepaste psychologie

Toets naam	Werkvorm		Toetsing				KW
	Hc	Wc	Dig	Wk	Ass	Sc	
Inleiding AOP	x		x				1
Inleiding psychologie	x	x	x	x			1
Practice Based Learning (PBL) 1		x		x	x	x	1
Gespreksvoering (PGT)		x			x		1, 2
Diagnostisch onderzoek	x	x	x(2)				1, 2
Sociale psychologie	x		x				2
Inleiding G&T	x		x				2
PBL 2		x		x(3)			2
Biologische psychologie	x		x				3
Inleiding GZP	x		x				3
PBL 3		x		x(3)			3
Praktijkgericht onderzoek 1	x	x	x	x			3, 4

STEM1	x	x	x	x	3, 4
Cognitieve en neuropsychologie	x	x	x	x	4
Ontwikkelingspsychologie	x		x		4

Opmerking. Werkvormen: Hc, hoorcolleges; Wc, werkcolleges. Toetsing: Dig, digitaal; Wk, werkstuk; Ass, assessment; Sc, schriftelijke toets. KW, kwartiel van het vak. $x(n)$, n = aantal toetsingen.

2.2 Educational Data Mining

Nu duidelijk is binnen welke context (onderwijs) het onderzoek wordt uitgevoerd, kan nagegaan worden welke rol Educational Data Mining (EDM) daarin kan spelen door antwoord te zoeken op de vraag wat het is hoe het gebruikt wordt. Wat EDM is kan op zijn breedst beschreven worden als het inzetten van technieken waarmee patronen in reeds aanwezige collectie van data ontdekt kunnen worden binnen een onderwijssetting (Brookshear, 2007; Fernandes et al., 2019). De technieken waarnaar gerefereerd wordt behoren tot het domein van data mining: kennis vergaren uit een grote hoeveelheid data (Brookshear, 2007). Het k -Means clustering algoritme is hier een voorbeeld van, welke een data set $D = \{x_1, x_2, \dots, x_n\}$ in K onsamenhangende clusters, $C = \{C_1, C_2, \dots, C_K\}$, probeert te verdelen, waarbij elke object, x_i , aan een cluster C_k wordt toegewezen. Dit type algoritme (Clustering algoritme) wordt, bijvoorbeeld, vaak ingezet om studenten (*objecten*) met gelijke kenmerken te groeperen (toewijzen aan clusters) op basis van LMS-data (Romero et al., 2010).

Romero et al. (2010) omschrijven in hun geredigeerd werk zes globale doeleinden en de daar bijhorende meest ingezette technieken. In tabel 2.2 is hier een overzicht van weergegeven, waarin te zien is dat Association, Clustering (zoals het hiervoor genoemde k -Means algoritme) en Classification het meest ingezet worden. Zo ook voor het voorspellen van student prestaties en leeruitkomsten, het vierde toepassingsgebied. Voordat deze technieken nader omschreven worden, is het belangrijk eerst een stap terug te nemen. Het eerste onderscheid dat binnen EDM gemaakt kan worden tussen deze technieken is namelijk *supervised* en *unsupervised learning*.

Een algoritme binnen de categorie supervised learning opereert vanuit de aanname dat een gebruiker (supervisor) de objecten (studenten) op voorhand classificeert, terwijl bij unsupervised learning deze classificaties zelfstandig door het algoritme gecreëerd worden (Sathya & Abraham, 2013). Meer concreet gesproken bepaalt de gebruiker bij supervised learning welke variabelen onder de *predictor* en *respons variabelen* vallen (ook wel

onafhankelijke en afhankelijk variabelen), terwijl dit bij unsupervised learning niet het geval is. Bij unsupervised learning ontdekt het algoritme zelfstandig de predictor én respons variabelen (zoals bijvoorbeeld een cluster bij Clustering), waarna het aan de *supervisor* (gebruiker) de taak hier betekenis aan te verlenen (interpreteren).

Nu dit bekend is, wordt duidelijk wat wordt bedoeld als gezegd wordt dat zowel Association Rules als Clustering onder de categorie unsupervised learning vallen. Het wordt ook direct duidelijk dat beide van deze technieken (in eerste instantie) niet bruikbaar zijn voor het voorspellen van studiesucces, aangezien bij het voorspellen van studiesucces de responsvariabele al op voorhand bekend is en het algoritme gericht patronen dient te ontdekken in de set predictor variabelen voor elke groep (categorie, waarde of classificatie) in de respons variabele. In tweede instantie blijkt echter dat het met inzet van Clustering mogelijk is om *features* (predictor variabelen) te genereren in functie van Classification. In andere woorden, door eerst Clustering toe te passen op de set predictor variabelen voor het Classification algoritme, kunnen nieuwe predictor variabelen gegenereerd worden. Deze kunnen vervolgens gebruikt worden bij supervised learning (López, Luna, Romero, & Ventura, 2012). Dit maakt het ontwikkelen van een model echter complexer, aangezien meer kennis en expertise benodigd is binnen het domein van EDM. Het risico dat men onbruikbare of niet te generaliseren modellen ontwikkelt neemt hierdoor toe (Romero et al., 2010). Omdat dit onderzoek een eerste poging is om door middel van EDM een model te ontwikkelen voor het voorspellen van studiesucces, is ervoor gekozen het inzetten van algoritmes te beperken tot Classification, een vorm van supervised learning (Romero & Ventura, 2007). Het eerstvolgende aandachtspunt is daarom om nader in te gaan op wat Classification inhoudt.

Tabel 2.2

Het doel en de meest gebruikte methode voor de verschillende toepassingsgebieden binnen educational data mining.

Toepassingsgebied	Doel	Methode
1. Communicatie naar belanghebbenden	Cursusleiders en leraren assisteren in het analyseren van (online) activiteiten van studenten en hoe met informatie wordt omgegaan in cursussen.	Statistische Analyses Visuele rapportages Proces-mining
2. Onderhouden en verbeteren van cursussen	Optimaliseren van cursussen (zoals de inhoud, aanbod, links, enzovoort) in functie van leeruitkomsten.	Association Rules Classification Clustering
3. Aanbevelingen doen	Materiaal aanbieden (zoals opdrachten en leesmateriaal) aansluitend op de huidige staat (zoals concentratie en huidig kennisniveau) van de student.	Association Rules Classification Clustering Sequencing
4. Prestaties en leeruitkomsten voorspellen	Het eindcijfer of andere leeruitkomsten (zoals behoud van kennis of lerend vermogen in vervolgonderwijs) voorspellen.	Association Rules Classification Clustering
5. Studentkenmerken modeleren	Onder andere de huidige staat van de student (zoals motivatie, tevredenheid en progressie) identificeren of het signaleren van factoren die de leeruitkomsten van de student negatief beïnvloeden.	Statistische Analyses Classification Clustering
6. Domein Structuur Analyse	Identificeren van een domein structuur. Het kunnen voorspellen van student prestaties wordt als maat gebruikt voor het bepalen van de kwaliteit.	Association Rules Clustering Space-Searching Algorithms

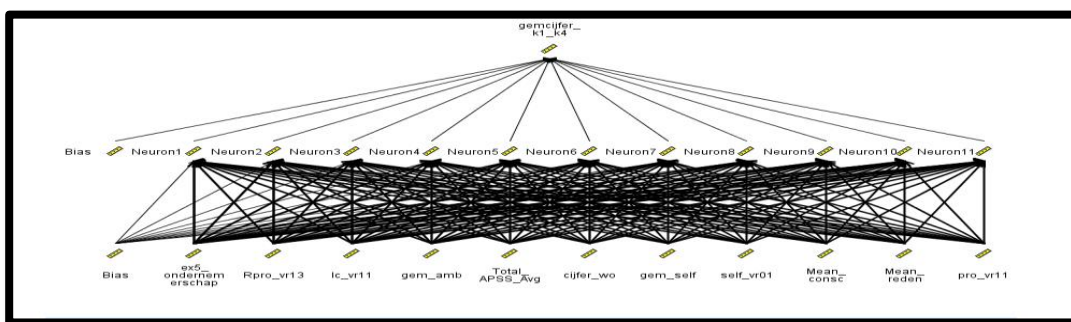
2.2.1 Classification

Het doel van Classification is voorspellen tot welke categorie (*classificatie*) een object (student) toebehoort. Een algoritme dat deze taak uitvoert wordt een *classificeerder* genoemd. Zo zijn classificeerders bijvoorbeeld in staat te achterhalen hoe gemotiveerd een student is gebaseerd op klik-gedrag binnen een LMS, waarbij de mate van motivatie te voorspellen classificatie is. Dit doet men door de predictor variabelen te selecteren, welke gekozen worden op basis van de kennis van de gebruiker (zoals een psycholoog), maar het is gebruikelijker deze te *leren* op basis van actuele gegevens (zoals data uit Bb en het intakeassessment). Eerst dient men de classificatie methode (algoritme) te bepalen. Daarna wordt de data verdeeld in twee (of drie) groepen: de train groep en test groep (en als eventuele derde groep een validatie groep). Het model *leert* patronen herkennen in de features (predictor variabelen) van de train groep die voorkomen bij de gespecificeerde classificaties (respons variabele), waarna deze geleerde patronen worden getest op de test data. Er wordt *getest* of de ontdekte patronen tevens voorkomen in de *test* groep. Als de classificeerder vervolgens veel classificaties correct voorspelt in deze test groep, wordt aangenomen dat het algoritme (de classificeerder) in de toekomst ook goede voorspellingen zal doen (Romero et al., 2010). Het leren is dan succesvol.

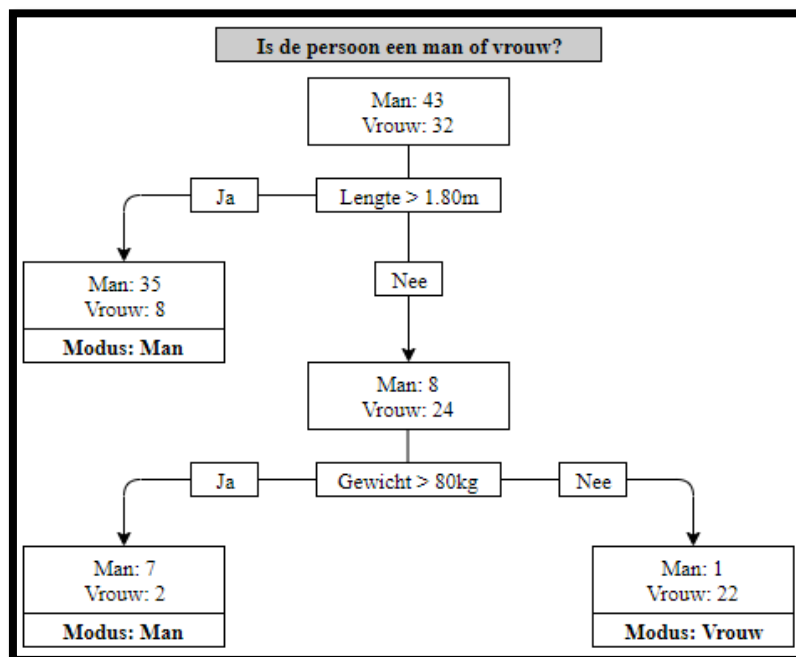
In andere woorden wordt bij classificatie dus gebruik gemaakt van een dataset waarin de records vooraf in een categorie geplaatst worden (zoals een laag, gemiddeld of hoog cijfer), waarna binnen deze data patronen *geleerd* worden die gepaard gaan met de betreffende *classificatie*. De patronen worden vervolgens getoetst op een test set in het kader van generaliseerbaarheid (Romero, Ventura, Pechenizkiy, & Bakar 2010). Een voorbeeld van een classificeerder is de Rule-based classificeerder. Deze maakt gebruik van een set ALS-DAN regels om te classificeren, waardoor het gemakkelijk geïnterpreteerd kan worden (Abu Tair & El-Halees, 2012). Een regel kan dan bijvoorbeeld zijn dat als een student in Deventer woont, een hoog aantal oefentoetsen heeft gemaakt en zich ruim van te voren heeft ingeschreven voor de toets, de student dan een hoog gemiddeld cijfer zal halen. Hoe goed het model in staat is classificaties te voorspellen wordt vervolgens uitgedrukt in een percentage genaamd *accuratesse*. Dit is de ratio van het aantal goede voorspellingen over het totaal aantal gemaakte voorspelling en wordt als volgt berekend (Romero et al., 2010):

$$Accuraatheid = \frac{\text{Aantal correcte voorspelling}}{\text{Totaal aantal voorspelling}}$$

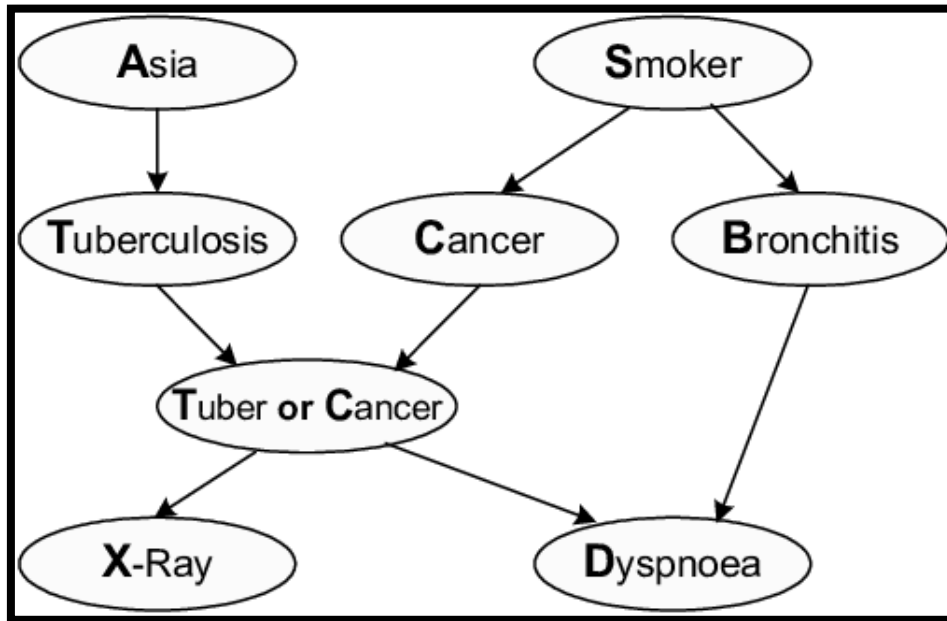
Classification is dus het classificeren van objecten. Er zijn echter veel verschillende classificeerders, met elk voordelen, nadelen en vereisten. Om een keuze te maken tussen van welk algoritme of algoritmes men gebruik gaat en kan maken, zijn een aantal criteria op zowel het niveau van de gebruiker als op niveau het algoritme van belang. Ten eerste is het voor de gebruiker van belang te weten in welke mate de resultaten te interpreteren zijn door diverse partijen. Er moet vooraf nagegaan worden waar het ingezet wordt en door wie. Een Neural Network, hoewel krachtig, is zeer complex om te interpreteren (figuur 2.1). Decision trees (figuur 2.2) zijn daarentegen gemakkelijk te interpreteren vanwege de grafische weergave, evenals bayesian classifiers (figuur 2.3).



Figuur 2.1. Schematische weergave van een Neural Network. Aangepast overgenomen uit “Welke determinanten verklaren jouw studiesucces?”, door Ambagts, J., 2018, p. 62.



Figuur 2.2. Voorbeeld van een decision tree (fictieve data).



Figuur 2.3. Het Asia Bayesian Network. Overgenomen uit “An Algorithm to Handle Structural Uncertainties in Learning Bayesian Network”, door Fernandes, C. M., Da Silva, W. T., & Ladeira, M., 2004, p. 7, in Proceedings of Ibero-American Symposium on Software Engineering and Knowledge Engineering.

Ten tweede, vanuit het perspectief van het algoritme, moet nagegaan worden welke eisen gesteld worden ten opzichte van de data om voorspellingen te kunnen doen. Zo werkt het ene algoritme (naïeve Bayes) beter met weinig data dan het andere algoritme (K-nearest neighbor). Romero et al. (2010) bieden een overzicht van verschillen tussen de diverse classificeerders (zie tabel 2.3), wat vervolgens kan helpen bij het maken van een keuze. Er wordt onderscheid gemaakt tussen (non-)lineaire regressie, accuratesse met kleine datasets, werken met incomplete data, ondersteuning voor diverse meetniveaus, interpreteerbaarheid en efficiëntie (redeneren, leren en updaten). De laatste in deze opsomming (efficiëntie) refereert naar de computationele snelheid en wordt onderverdeelt in drie categorieën: (1) beredeneren, het onderscheidend vermogen tussen (belangrijke) variabelen; (2) leren, het construeren van een model door patroonherkenning; (3) updaten, het aanpassend vermogen door toevoeging van nieuwe data. De gebruiker dient vooraf te bepalen welke criteria voor hem belangrijk is. In het geval van dit onderzoek is interpreteerbaarheid een belangrijk criterium, aangezien diverse (ondeskundige) partijen met het model moeten kunnen werken. Decision Trees, naïeve Bayes classificeerders en General Bayes classificeerders voldoen hieraan. De volgende stap is na te gaan wat deze classificeerders van elkaar onderscheidt, waar in de volgende sub-paragrafen aandacht aan wordt besteed.

Tabel 2.3

Vergelijking tussen verschillende classificeerders

Criterium	<i>DT</i>	<i>NB</i>	<i>GB</i>	<i>FFNN</i>	<i>K-nn</i>	<i>SVM</i>
Non-lineaire grenzen	+	(+)	+	+	+	+
Accuraat bij kleine datasets	–	+	(+)	–	–	+
Incomplete data	–	+	+	+	+	–
Diverse meetniveaus	+	+	+	–	+	–
Gemakkelijk te interpretern	+	+	+	–	(+)	–
Efficiënte beredenering	+	+	+	+	–	+
Efficiënt leren	(+)	+	–	–	(+)	+
Efficiënt updaten	–	+	+	+	+	–

Opmerking. Aangepast overgenomen uit “*Handbook of Educational Data Mining*”, door Romero, C. et al., 2010, p. 70, Boca Raton, FL: CRC Press. De afkortingen en symbolen in de tabel zijn: +, het model ondersteunt het criterium; –, het model ondersteunt het criterium niet; (+), ondersteuning afhankelijk van meetniveaus of datagrootte; *DT*, decision trees; *NB*, naïeve Bayes classificeerders; *GB*, general Bayes classificeerder; *FFNN*, feed-forward neural network; *K-nn*, *K*-nearest neighbor classificeerder; *SVM*, support vector machine.

2.2.1.1 Decision trees

Decision trees behoren tot de meest populaire classificerende algoritmes (Geurts, 2002; Romero et al., 2010). Een decision tree biedt in zijn meest simplistische vorm een overzicht van binaire scheidingen in de vorm van een besluitboom. Figuur 2.2 illustreert deze simplificatie. In het voorbeeld kan de gebruiker zien dat als de persoon langer is dan 1.80 meter, de persoon waarschijnlijk een man is. Is de persoon korter, dan volgt men de route naar het eerst volgende criterium (gewicht). Weegt de persoon vervolgens meer dan 80 kilogram, dan is de persoon waarschijnlijk een man, terwijl in het tegenovergestelde geval de persoon waarschijnlijk een vrouw is.

Naast dat decision trees gemakkelijk te interpretern zijn, worden zij tevens gekenmerkt door hun efficiëntie en flexibiliteit. Zij zijn namelijk snel te genereren, waaronder met datasets bestaande uit miljoenen objecten en duizenden features. Daarnaast werken zij zowel met variabelen op interval en ratio niveau, als nominale (waaronder binair) en ordinale variabelen. Een bekend nadeel aan decision trees is dat overfitting gemakkelijk plaatsvindt, waardoor feature selection en optimaliseren in complexiteit, bias en variantie belangrijke procedures zijn

om de accuratesse en generaliseerbaarheid van het model te verhogen (Geurts, 2002). In bijlage A wordt nader ingegaan op deze begrippen.

2.2.1.2 Bayesian Networks

Een Bayesian Network is een grafisch model, welke (beoogde) causaliteit tussen variabelen laat zien. Figuur 2.3 is een voorbeeld van het bekende Asia Bayesian Network, welke vaak gebruikt wordt om Bayesian Networks te illustreren. In het voorbeeld geeft Asia aan of de persoon naar Azië is geweest, wat een voorspeller is voor tuberculose. Smoker (rookt de persoon) is een voorspeller van kanker en bronchitis. Hoewel in figuur 2.3 geen waarschijnlijkheden worden weergegeven, worden deze er over het algemeen wel bij vermeldt. Wat het echter illustreert is de simpliciteit waarin het model geïnterpreteerd kan worden. Evenals een decision tree werkt dit type model met diverse meetniveaus en is het flexibel. Nadeel is echter dat bij een laag aantal classificaties (zoals binaire, waarbij slechts twee uitkomsten mogelijk zijn) de accuratesse van de Bayesian Networks lager is (Romero et al., 2010). Over het algemeen zijn Bayesian networks in vergelijking met decision trees minder sterk in staat voorspellingen te doen. Ook is het optimaliseren van de features voor een Bayesian network complexer, waardoor het meer expertise en ervaring vanuit de gebruiker vraagt. Het gebruik maken van een decision tree biedt daarom de voorkeur.

2.2.2 Voorspellen van studiesucces in EDM

Tot dusver werd ingegaan op *wat* EDM is. De vraag die vervolgens rest is *hoe* het wordt gebruikt om studiesucces te voorspellen. Ramaswami en Bhaskaran (2010) maakte hiervoor gebruik van een op CHAID (algoritme) gebaseerde decision tree voor het voorspellen van de eindprestatie (eindcijfer) van studenten (*secondary higher education*). Het eindcijfer werd verdeeld in zeven classificaties en de dataset bestond uit (socio-)demografische gegevens en prestaties tijdens de vooropleiding (in totaal 34 features) van 772 studenten. De belangrijkste voorspellers waren de voertaal van het onderwijs, (eind)cijfer van de vorige opleiding, onderwijsinstelling, woongebied en type vooropleiding. De accuratesse van het model van 44.69%, wat een toename van meer dan 5% is ten opzichte van de baseline (39.23%).

Ramesh, Parkavi en Ramar (2013) deden soortgelijk onderzoek. Hun dataset 500 studenten (*secondary higher education*) waarvan het eindcijfer voorspelt werd. De hoogste accuratesse werd behaald met het neural network Multilayer Perceptron (72.38%), gevolgd door de decision trees J48 (64.88%; J48 is een in Java verwerkte versie van de C4.5 decision tree) en REPTree (60.13%). Studiesucces werd in zeven classificaties verdeeld en als

voorspellers werden demografische en socio-economische gegevens, verkregen via vragenlijsten, gebruikt (in totaal 28 features). De belangrijkste voorspellers waren type financiering, thuis studeren, vooropleiding van de ouders en locatie van de onderwijsinstelling (vooropleiding en huidige opleiding).

Osmanbegović en Suljić (2012) voorspelde het al dan niet behalen van het eindexamen van eerstejaars studenten in het hoger onderwijs op basis van (socio-)demografische gegevens, prestaties tijdens de vooropleiding, resultaten van het intake assessment bij aanmelding en mening ten opzichte van studeren. Het eindcijfer werd van zes (A, B, C, D, E en F) naar twee (behaald en niet-behaald) classificaties getransformeerd. De dataset bestond uit 257 studenten en 12 voorspellers (features). Om te classificeren werd gebruik gemaakt van een Naive Bayes (NB) algoritme, Multilayer Percetron neural network (MLP) en C4.5 decision tree (ook wel J48). De NB classificeerder was in staat 76.65% correct te voorspellen, wat slechts 1% hoger lag dan de baseline. C4.5 en MLP waren respectievelijk in staat 73.93% en 71.2% correct te voorspellen. Eerdere prestaties tijdens de opleiding, gebruikte type studiemateriaal, score op de toelatingstest en tijd besteed aan zelfstudie waren de belangrijkste voorspellers.

Mueen, Zafar en Manzoor (2016) maakte naast algemene gegevens (zoals de onderzoeken in de vorige sub-paragraaf) gebruik van data vanuit het LMS systeem (Blackboard) en eerdere prestaties tijdens de opleiding om studiesucces te voorspellen. Studiesucces werd gedefinieerd als het behalen van de eindtoets van de cursus. De dataset bestond uit 60 studenten en 38 features. Om te classificeren werd gebruik gemaakt van de NB (Naive Bayes), MLP (Multilayer Perceptron) en C4.5 classificeerders. De NB classificeerder had de hoogste accuratesse (85.7%), gevolgd door de MLP (81.4%) en C4.5 (80.5%) classificeerders. Alle modellen waren aanzienlijk beter in staat te voorspellen dan de baseline (68.33%). De belangrijkste voorspellers waren eerdere prestaties, tussentijdse toetsen, activiteit op Blackboard en aanwezigheid tijdens de lessen.

Conijn, Kleingeld, Matzat, Snijders en Van Zaanen (2016) maakte gebruik van LMS data (activiteit op Blackboard), persoonlijkheidskenmerken, algemene kenmerken en tussentijdse toets-resultaten om studiesucces te voorspellen. De dataset bestond uit 888 studenten en 24 features. Het cijfer werd als maat voor studiesucces gebruikt, waarbij driemaal op binair niveau geclassificeerd werd. Namelijk cijfer < 3 (ja/nee), cijfer < 5.5 (ja/nee) en cijfer > 8 (ja/nee). Hoewel de modellen onderling gelijke prestaties lieten zien, werd geconcludeerd dat de decision trees het meest geschikt zijn, aangezien deze gemakkelijker te interpreteren zijn ten opzichte van andere classificeerders (Support Vector Machines). Met de J48 (C4.5) en JRip decision trees werd respectievelijk een accuratesse van 67.1% en 66.9% behaald met een

combinatie van alle dataset. Met een SVM werd een accuratesse behaald van 68.7%. De accuratesse van alle algoritmes waren aanzienlijk hoger dan de baseline van 50.7%.

Dekker, Pecheniznky en Vleeshouwers (2009) voorspelde studiesucces (persisteren; ja/nee) van eerstejaars studenten op basis van data voor en tijdens het de opleiding. In totaal werden 13 features gebruikt van 648 studenten. Er werden modellen geconstrueerd op basis van variabelen beschikbaar vóór de opleiding (pre-university only), tijdens de opleiding (university only) en een vóór en tijdens de opleiding (complete dataset). Decision trees (J48 en CART) hadden een accuratesse van 79% en 80%, wat een verbetering is ten opzichte van de baseline classificeerder OneR (75%). De belangrijkste voorspellers waren prestaties voor vakken tijdens de opleiding.

Wat de hiervoor genoemde werken laten zien is dat de prestaties van de classificeerders kunnen variëren. Zo presteert een neural network (NN) de ene keer beter dan een decision tree (Ramesh, Parkavi, & Ramar, 2013), terwijl in andere onderzoeken dit andersom is (Osmanbegović & Suljić, 2012). Osmanbegović en Suljić (2012) benadrukken verder het belang van interpreteerbaarheid, evenals Conijn et al. (2016), binnen de context het onderwijs. Met name decision trees krijgen de voorkeur, aangezien deze van alle classificeerders het gemakkelijkst te interpreteren zijn (Geurts, 2002).

2.3 Studiesucces

Tot op dit punt in het onderzoek is bekend binnen welke context het onderzoek plaatsvindt, wat EDM is en hoe studiesucces ermee voorspelt wordt. Het is echter nog onbekend wat de theorie achter studiesucces is. Wat blijkt is dat studiesucces een complex begrip is. Het komt in de literatuur veel voor en de gehanteerde definities variëren sterk van elkaar. York, Gibson en Rankin (2015) deden daarom literatuuronderzoek naar de meest voorkomende definiëringen, met het doel tot een nieuw kader te komen op basis waarvan studiesucces gedefinieerd kon worden. Het onderzoek resulteerde in de volgende begrippen voor studiesucces: 1) academische prestaties, 2) tevredenheid, 3) opleiding gerelateerde vaardigheden en competenties, 4) persisteren, 5) behalen van doelen en 6) succes in het werkveld. Academische prestaties wordt veruit het meest gebruikt (67.7%), gevolgd door opleiding gerelateerde vaardigheden en competenties (48.4%) en persisteren (22.6%). De eerste daarvan, academische prestaties, wordt gemeten aan de hand van het gemiddelde gewogen cijfer (GPA; Grade Point Average). De tweede, vaardigheden en competenties, worden gemeten met behulp van vragenlijsten. De laatste, persisteren, wordt gemeten door het al dan niet behalen van een diploma (ook wel uitval of dropout genoemd). Andere onderzoeken,

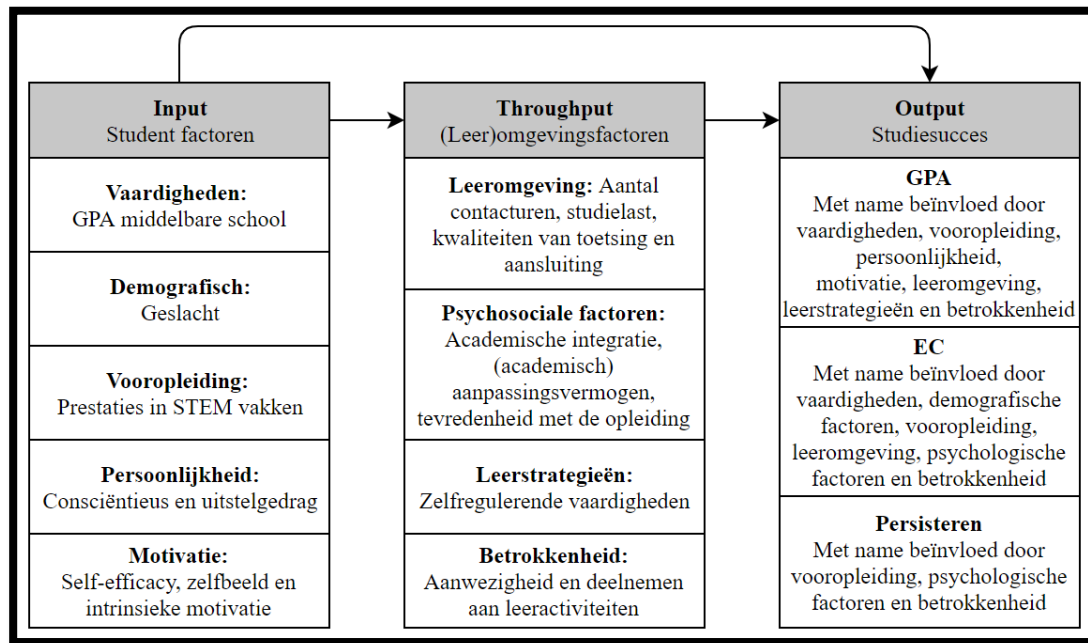
zoals het onderzoek van Van Rooij et al. (2017), hanteren naast GPA en persisteren ook het aantal studiepunten als maat voor studiesucces (EC). Studiepunten is een maat voor het uitdrukken van studielast, namelijk 28 uur per EC, en om deze punten te behalen dient de student een voldoende resultaat te behalen voor het betreffende vak, waarbij een hoger cijfer geen extra EC oplevert. Wat in het veld van EDM echter opvalt is dat GPA en persisteren (dropout of uitval) het meest gebruikt worden. Hierbij wordt GPA gebruikelijk verdeeld in categorieën, variërend van twee (Osmanbegović & Suljić, 2012) tot zeven classificaties (Ramaswami & Bhaskaran, 2010), waarbij de beste resultaten, in termen van accuratesse, behaald worden door te splitsen op binair niveau (Romero et al., 2010). Het aantal studiepunten kwam nergens voor.

2.3.1 Voorspellers van studiesucces

Studiesucces kan dus in diverse vormen worden uitgedrukt, zoals GPA en persisteren. Nu bekend is wat studiesucces inhoudt, kan nagegaan worden welke voorspellers ervan bekend zijn in de literatuur. Een veelgebruikt theoretisch model is het input-throughput-output model. Dit model is gebaseerd op Tinto's theorie van (student) uitval (Tinto, 1987), de herziene versie van Tinto's theorie door Braxton, Milem en Sullivan (2000) en het 3P-model van Biggs (Biggs, Kember, & Leung, 2001). Wat het input-throughput-output model zegt is dat studenten aan een opleiding beginnen met bepaalde eigenschappen (input), zoals vaardigheden en demografische kenmerken, welke ook wel student factoren worden genoemd. Vervolgens vindt tijdens het eerste jaar interactie met de onderwijsinstelling plaats (throughput). Tijdens deze fase wordt studentgedrag zichtbaar, evenals de impact die de onderwijsinstelling en opleiding daarop heeft. De fase wordt gekenmerkt door onder andere de betrokkenheid, zelfregulerende vaardigheden en psychosociale factoren, zoals academische integratie en aanpassingsvermogen. Tot slot zijn er drie output factoren, namelijk GPA, EC en persisteren.

Van Rooij et al. (2017) deden vervolgonderzoek naar dit model voor het vinden van voorspellers van studiesucces voor eerstejaars Nederlandse en Vlaamse studenten. Geconcludeerd werd dat studenten die hogere cijfers hebben en vakken volgen in de wetenschappen en wiskunde (STEM vakken) beter presteren in het hoger onderwijs en een grotere kans hebben het eerste jaar succesvol af te ronden (persisteren). Hoewel het geen nieuw inzicht is dat resultaten op de vooropleiding een voorspeller is, bestaat internationaal gezien geen consensus over het gegeven dat het volgen van wiskundige en wetenschappelijke vakken tijdens de vooropleiding als voorspellers gelden. Dit is namelijk afhankelijk van het onderwijssysteem in het land, waardoor het in sommige landen wel als voorspeller geldt (zoals

in Nederland) en in andere landen niet. Naast vooropleiding blijken consciëntieusheid, intrinsieke motivatie, academische aanpassingsvermogen, gebrek aan zelfregulerende vaardigheden, aanwezigheid en participeren in leeractiviteiten gerelateerd aan de drie outputfactoren (GPA, EC en persisteren). In figuur 2.4 is een overzicht te zien van de belangrijkste gevonden resultaten en voorspellers per output categorie.



*Figuur 2.4. Overzicht van de belangrijkste voorspellende factoren van studiesucces voor eerstejaars Nederlandse en Vlaamse studenten in het hoger onderwijs. Aangepast overgenomen uit “A Systematic Review of Factors related to First-Year Students’ Success in Dutch and Flemish Higher Education,” door E. Van Rooij et al., 2017, *Pedagogische Studiën*, p. 375.*

Hoewel dit type theoretische modellen onderbouwend kunnen zijn voor het selecteren van features, wordt binnen EDM soms andere resultaten dan verwacht gevonden (Dekker, Pecheniznky, & Vleeshouwers, 2009). Het ontdekken van onverwachte resultaten is dan ook juist de kracht van EDM, waardoor het ingezet wordt wanneer deze theorieën onvoldoende verklarend werken (Romero et al., 2010). Voorspellers worden binnen EDM daarom gekozen op basis van feature selection technieken (zie bijlage A) zoals Information Gain, wat in simpele termen gezien kan worden als meer zekerheid (informatie) krijgen over de waarde (classificatie) van de ene variabele als gevolg van informatie uit de andere (predictor). Feature selection komt dus neer op het toepassen van (statistische) technieken waarmee variabelen met de hoogste voorspellende waarde geselecteerd kunnen worden, ten einde het totaal aantal te

gebruiken features (de dimensionaliteit) te verminderen om overfitting te voorkomen. Bijlage A biedt een overzicht van deze begrippen.

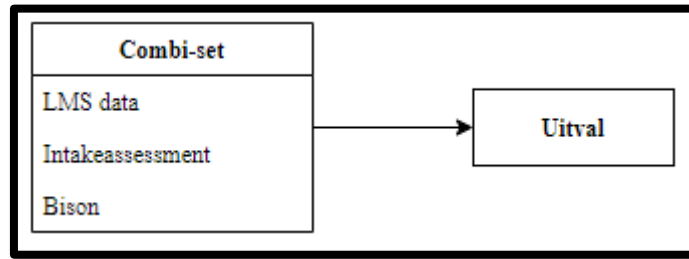
2.4 Variabelen en het conceptueel model

Dit hoofdstuk kan nu afgesloten worden door alle voorgaande informatie met elkaar te integreren om tot een raamwerk voor dit onderzoek te komen. In hoofdstuk 1 werd benoemd dat dit onderzoek de vraag wil beantwoorden in welke mate studiesucces voorspelt kan worden op basis van de beschikbare data. Deze data is afkomstig uit diverse bronnen binnen Saxion, namelijk Blackboard Learn, Bison en het intakeassessment, waarvan de laatste (intakeassessment) bestaat uit demografische gegevens, persoonlijkheid en capaciteiten. In paragraaf 2.2 en 2.3 werden vervolgens definities en voorspellers van studiesucces gepresenteerd. Deze informatie is leidend voor het definiëren van variabelen en construeren van het conceptueel model van dit onderzoek (figuur 2.5).

In sub-paragraaf 2.2.2 werd duidelijk dat data reeds aanwezig kan zijn, maar men kan er ook voor vragenlijsten uit te zetten. In het geval van dit onderzoek gaat het om reeds aanwezige data. Blackboard is een LMS systeem en data daaruit afkomstig wordt in het vervolg ook wel LMS data genoemd. Het intakeassessment bevat demografische gegevens en pre-enrollment data (persoonlijkheid en capaciteiten). De data afkomstig uit Bison bestaat uit de cijfers van toetsen, waarvan cijfers uit het eerste kwartiel (eerste toets-kans) worden gebruikt. Omdat feature selection (zie bijlage A) wordt gebruikt voor het selecteren van voorspellers, worden deze variabelen in het volgende hoofdstuk nader omschreven.

Studiesucces wordt gedefinieerd als uitval: het al dan niet doorstromen naar het tweede studiejaar. De eerste reden hiervoor is dat de accuratesse van modellen hoger is bij binaire classificatie (Romero et al., 2010). Ten tweede is de financiering (lumpsum; zie paragraaf 1.1) van hogescholen niet afhankelijk van het GPA maar het aantal voltooide bachelors waarvoor een diploma is verleend. Verder vindt uitval plaats wanneer onvoldoende studiepunten zijn behaald of de student besluit te stoppen met de opleiding. In deze criteriummaat is dus het aantal EC verworven, aangezien de student een minimaal aantal van 48 EC moet behalen in het eerste studiejaar om door te mogen.

In het volgende hoofdstuk bevindt zich de methodologische verantwoording, waarnaast het proces alsmede de analyses worden toegelicht.



Figuur 2.5. Het conceptueel model

Hoofdstuk 3. Onderzoeksdesign

Dit hoofdstuk legt uit hoe het onderzoek wordt uitgevoerd. Dit wordt gedaan door eerst de methode, aaneengesloten met het gekozen proces, toe te lichten in paragraaf 3.1. Dit wordt gevolgd door de onderzoeksdoelgroep (3.2), onderzoeksinstrumenten (3.3), procedure (3.4) en analyses (3.5).

3.1 Onderzoeksmethode

Om te onderzoeken in welke mate de beschikbare data studiesucces van TP-studenten kan voorspellen wordt een explorerend onderzoek toegepast. Het kenmerkt zich door het ontbreken van een samenhangend stelsel waarmee empirische regelmatigheden (studiesucces in dit geval) beschreven, verklaard of voorspelt kunnen worden. In andere woorden, er wordt een theorie gezocht in plaats van getoetst. Een volgend kenmerk is dat kennis ontbreekt over wat er met de beschikbare gegevens in de data bases en EDM gedaan kan worden. Tot slot onderscheidt het zich van een toetsend onderzoek aangezien er vooraf geen hypotheses geformuleerd zijn die methodologisch getoetst worden. Dit betekent dat de aanwezige variabelen in de data bestanden veranderd of verwijderd kan worden om een model te vinden waarmee studiesucces voorspelt kan worden (Hart, Boeijs, & Hox, 2007).

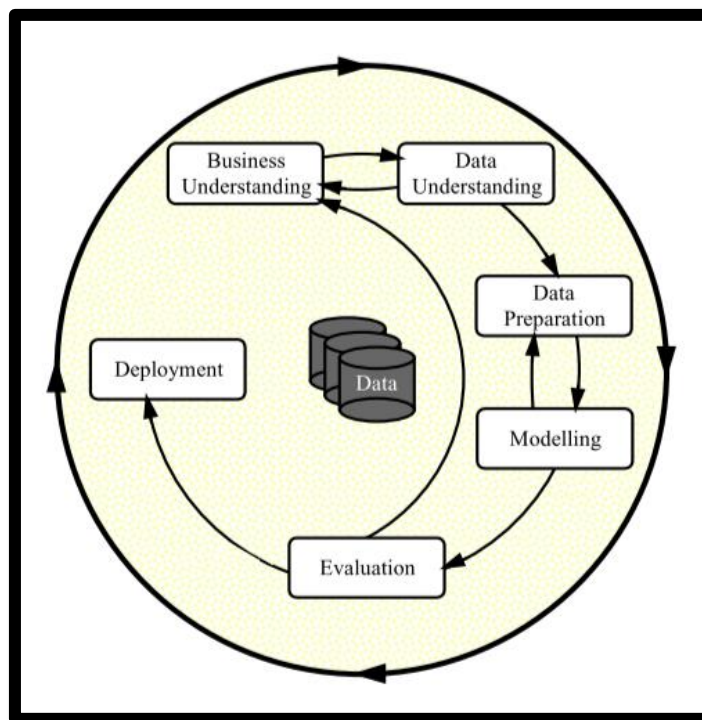
Voor wat betreft de keuze van het proces en de methodiek bij data mining wordt in de literatuur het meest gebruik gemaakt van de Knowledge Discovery in Databases (KDD), CRISP-DM en SEMMA methode. SEMMA valt af, aangezien deze verbonden is aan de SAS Enterprise Miner software waar Saxion geen licenties voor heeft. De KDD en CRISP-DM methode lijken sterk op elkaar (Azevedo & Santos, 2008). Uiteindelijk is gekozen om de CRISP-DM methode (Cross-Industry Standard Process for Data Mining) te hanteren (Chapman et al., 2000). Deze is gratis toegankelijk en niet gebonden aan specifieke software licenties en wordt veel ingezet in het kader van wetenschappelijk onderzoek (Kabakchieva, 2013). De methode bestaat uit zes stappen:

1. *Business Understanding* – Deze fase bestaat uit het begrijpen van het probleem vanuit het perspectief van de opdrachtgever (probleemstelling), de rol van data mining bepalen, doelenformulering en plan van aanpak.
2. *Data Understanding* – Hierbij draait het om het verzamelen en in kaart brengen van de variabelen waarover de data beschikt.
3. *Data Preparation* – Bestaat uit alle activiteiten die benodigd zijn om de uiteindelijk dataset te construeren. Deze fase wordt vaak meerdere malen herhaald. Taken bestaan

uit het selecteren van tabellen, rijen en kolommen (attributen), opschonen van data, nieuwe attributen construeren en het transformeren van data voor modellen.

4. *Modelling* – In deze fase worden diverse modellen toegepast op de data. Er zijn vaak meerdere modellen geschikt voor dezelfde type vraag- en probleemstellingen, waarvan sommige een specifieke opmaak vereisen. Hierdoor is er vaak sprake van een wisselwerking tussen deze en de voorgaande fase (data preparation). Men kan bijvoorbeeld realiseren dat een andere constructie van categorieën geschikter is voor het model.
5. *Evaluation* – In deze fase worden de modellen en het proces (welke stappen ondernomen zijn) geëvalueerd. Vaak wordt hierin bepaald in welke mate data mining effectief is gebleken voor het probleem.
6. *Deployment* – In het geval van dit onderzoek draait deze fase om rapporteren van gevonden resultaten.

De CRISP-DM methode is een cyclus en eindigt wanneer de Deployment fase succesvol kan worden afgerond. In figuur 3.1 is een schematisch overzicht te zien van de fases en waar deze in relatie tot elkaar staan.



Figuur 3.1. Fases van de CRISP-DM methode. Overgenomen uit “CRISP-DM: Towards a Standard Process Model for Data Mining,” door R. Wirth, & J. Hipp, 2013, Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, p. 65.

3.2 Onderzoeksdoelgroep

De onderzoeksdoelgroep bestaat uit alle eerstejaars voltijd studenten aan de opleiding toegepaste psychologie aan hogeschool Saxion te Deventer (TP-studenten; $n = 812$; zie tabel 3.1) van studiejaar 2016-2017 ($n = 266$), 2017-2018 ($n = 268$) en 2018-2019 ($n = 278$). Er is voor deze doelgroep gekozen aangezien, zoals in hoofdstuk 1 werd benoemd, slechts 55% van de studenten van de opleiding toegepaste psychologie doorstromen naar het tweede leerjaar. Daarnaast blijkt uit dat in 2016 het curriculum van de opleiding is veranderd, waardoor de data van oudere cohorten te veel afwijken en te veel informatie van huidige cohorten niet benut kunnen worden. Verder worden deeltijd studenten buiten beschouwing gehouden aangezien 1) van hen het intakeassessment afwijkt, 2) enkel (deeltijd) data van cohort 2016 beschikbaar is en 3) de omvang van deze groep zeer klein is ($n = 30$). Voor een uitgebreidere omschrijving en onderbouwing en het proces wordt verwezen naar bijlage B.

Tabel 3.1

Frequentietabel van de cohorten

Cohort	Frequentie	Percentage
2016-2017	266	32.8
2017-2018	268	33.0
2018-2019	278	34.2
Cumulatief	812	100.0

Er is geen toestemming vanuit de TP-studenten nodig om de data te mogen gebruiken en zij zijn niet ingelicht over het onderzoek. Saxion mag persoonsgegevens van studenten, studiekeizers, alumni en oud-studenten namelijk verstrekken ten behoeve van de bedrijfsvoering van Saxion. Een overzicht van de privacyverklaring is te vinden op de website van Saxion (<https://www.saxion.nl/over-saxion/privacyverklaring>). Om de privacy te waarborgen wordt het onderzoek binnen Saxion uitgevoerd en worden gegevens enkel op de servers van de hogeschool bewerkt en opgeslagen. Daarnaast heeft de onderzoeker een geheimhoudingsverklaring ondertekend, waarin verklaard wordt zorgvuldig met de informatie om te gaan en het niet voor andere doeleinden dan dit onderzoek te gebruiken.

3.3 Onderzoeksinstrumenten

Onder het onderzoeksinstrument vallen de data bases van Saxion, te weten: het intakeassessment, Blackboard Learn (Bb) en Bison. De student-gegevens beschikbaar vanuit het intakeassessment kunnen in drie categorieën worden: (1) algemene gegevens, (2) capaciteiten en (3) persoonlijkheid. De gegevens beschikbaar vanuit Bb bevatten het klikgedrag van de studenten in de gevolgde cursussen. Bison bevat informatie van de gemaakte toetsen van de studenten, waaruit de resultaten van eerste toets-kansen uit het eerste kwartiel zijn geselecteerd.

Uitval kan achterhaald worden door de studentlijsten per studiejaar met elkaar te vergelijken. Elk jaar worden namelijk lijsten opgesteld waarin de studenten staan die staan ingeschreven voor het betreffende jaar. Door de lijsten van eerstejaars studenten naast die van tweedejaars studenten van het daarop volgende jaar te leggen, kan achterhaald worden welke studenten doorgestroomd zijn. In bijlage B worden de databestanden uitgebreid omschreven.

3.4 Procedure

In paragraaf 3.1 werd CRISP-DM (Chapman et al., 2000) geïntroduceerd. Hieronder worden de belangrijkste stappen in het verwerken van de data toegelicht: data understanding, data preparation, modelling en deployment.

3.4.1 Data understanding

Data understanding bestaat uit het verzamelen en inventariseren van datasets en variabelen. Gegevens van het intakeassessment zijn al aanwezig in losse databestanden. De gegevens van Blackboard Learn dienen via een externe partij (Eesysoft) opgevraagd te worden aangezien slechts gegevens vanaf 2018 beschikbaar zijn. Tot slot kan de data van Bison via de Saxion servers gedownload worden.

Het inventariseren van variabelen wordt gedaan middels Excel, Python, SPSS en SPSS Modeler. Excel wordt gebruikt om data initieel in op te slaan. Deze data kan vervolgens met Python (programmeertaal) uitgelezen worden om bronbestanden van Bison en Bb met elkaar te combineren. De functionaliteiten van SPSS worden naderhand benut om inzicht te krijgen in frequenties en de verdelingen. Ter visualisatie biedt SPSS Modeler in het verlengde van SPSS een efficiënte tool om meerdere grafieken te genereren in functie van interpretatie. In bijlage D wordt een voorbeeld van dit proces met inzet van Python (in combinatie jupyter notebook) getoond.

3.4.2 Data preparation

Data preparation omvat het selecteren en opschonen van data en staat in het verlengde de voorgaande fase (data understanding). Na het bekijken van de data worden features geselecteerd op basis van relevantie, kwaliteit en (beperkingen in) omvang. Met Python kan met missing values worden gewerkt (invullen dan wel verwijderen). Studenten waarvan te veel data afwezig is worden verwijderd uit de dataset. De inzet van zowel Python als SPSS maakt feature selection op basis van statistische tests mogelijk (zie paragraaf 3.5 voor meer hierover).

3.4.3 Modelling en deployment

Na data preparation vindt modelling plaats, waarbij voor zowel het selecteren, genereren als evalueren van classificeerders plaatsvindt in SPSS Modeler (versie 18.2.1). Deployment omvat vervolgens het presenteren van de resultaten, wat wordt gedaan in de vorm van dit onderzoek en het presenteren van de meest accurate decision trees.

3.5 Analyses

Om te bepalen in hoeverre de verschillende datasets belangrijke features bevatten voor het voorspellen van uitval (deelvragen een tot en met drie) worden feature selection methodes gebruikt. Op correlatie en Chi-Kwadraat toets gebaseerde methodes blijken daartoe het best in staat (Ramaswami & Bhaskaran, 2009). Aangezien de responsvariabele dichotoom is (uitval of geen uitval) en er geen sprake is van onderliggende continuïteit, wordt in het kader van samenhang tussen deze en (continue) predictorvariabelen gesproken over point-biserial correlatie, wat berekend wordt met Pearson's correlatie coëfficiënt (Field, 2013). Voor het selecteren van features op nominaal niveau wordt de Chi-Kwadraat toets benut.

Om te toetsen in hoeverre met een combinatie van deze features studiesucces voorspelt kan worden (deelvraag vier), wordt gebruik gemaakt diverse decision tree algoritmes beschikbaar in SPSS Modeler (zie bijlage E voor een omschrijving de decision trees):

- C5.0
- CART
- CHAID
- Exhaustive CHAID
- QUEST

Om de decision trees te evalueren worden zij vergeleken met een baseline: ZeroR classificeerder. De ZeroR classificeerder negeert alle voorspellers en gaat enkel uit van de

respons variabele (uitval in dit geval). Het voorspelt op basis van de grootste groep (modus). Dit kan vervolgens uitgedrukt worden in een percentage van het totaal (Nasa & Suman, 2012). Dit maakt vergelijken met de accuratesse van de decision trees mogelijk. Accuratesse wordt als volgt bepaald:

$$Accuratesse = \frac{\text{Aantal correcte voorspelling}}{\text{Totaal aantal voorspelling}}$$

Vervolgens kan significantie worden bepaald met een binomiale test, waarbij grenswaardes voor significantie aangepast worden volgens de Bonferroni correctie, zodat de ratio Type I fouten (α) laag ($< 5\%$) blijft (Salzberg, 1997; Field, 2013). Dit wordt berekend door α te delen door het aantal vergelijkingen, k :

$$P_{criterion} = \frac{\alpha}{k}$$

Hoofdstuk 4. Onderzoeksresultaten

In dit hoofdstuk worden de resultaten getoond van het onderzoek. Het is dusdanig opgebouwd dat allereerst bijzondere handelingen worden vermeld (paragraaf 4.1), waarna in paragraaf 4.2 de resultaten per deelvraag worden getoond, waaronder het beschrijven van de variabelen.

4.1 Uitvoering

Intakeassessment

Het onderzoek werd uitgevoerd door eerst de data van het intakeassessment op te schonen. Deze dataset bestond uit zowel starters als niet-starters. Door de studenten in de dataset te vergelijken met de klassenlijsten van betreffende jaren, konden deze geïdentificeerd en verwijderd worden uit de dataset. Verder bleek er sprake van missing values op de items van dimensie 2 (persoonlijkheid) van het type *Not Missing at Random*, waardoor deze niet genegeerd mogen worden. De missing values werden van een unieke waarde voorzien, corresponderend met de missing values van de antwoordmogelijkheden op de items van dimensie 2, zoals “niet van toepassing” of “weet ik niet” (Kaiser, 2014).

Uitval

Vervolgens werd per student het criterium uitval achterhaald. Om te bepalen tot welke classificatie van uitval de student toebehoort, werden enrollment gegevens van voorgaande jaren gebruikt. Er wordt namelijk elk jaar aan de start van het studiejaar een bestand gemaakt met alle studenten die aan de opleiding starten. Het tweede leerjaar wordt eenzelfde soort lijst opgesteld, bestaande uit doorgestroomde studenten. Door te bekijken welke studenten wel en niet in het databestand voor het tweede studiejaar zitten kon bepaald worden welke studenten uitgevallen waren. Voor 46.8% (n = 380) van de studenten was er sprake van uitval en 53.2% (n = 432) stroomde door naar het tweede leerjaar.

Bison

Hierna werd data van Bison toegevoegd. Er werd geïnventariseerd welke toetsen per cohort in het eerste kwartiel gegeven werden door van de betreffende jaren de catalogi met toets-samenvatting te downloaden. Daarnaast werd voor elk vak de modulehandleiding gedownload vanuit de module omgevingen op Blackboard. In tabel 4.1 is te zien dat cohort 2016 afwijkt van cohort 2017 en 2018. Aangezien Professionele Gespreksvoering en Training uniek is voor cohort 2016, is deze niet meegenomen in de dataset. Na het bestuderen van de modulehandleidingen voor de vakken Presteren en Leren en Inleiding Arbeids- en

Organisatiepsychologie, bleek dat deze inhoudelijk overeen kwamen en samengevoegd konden worden onder dezelfde naam “Inleiding Arbeids- en Organiseatiepsychologie” (INLAOP). De dataset werd gefilterd op de eerste toetskans voor de betreffende toetskans, waarbij bekend was of de toets al dan niet behaald was en met welk resultaat. Een probleem aan deze dataset is dat niet exact achterhaald kon worden of de toets daadwerkelijk in het eerste kwartiel werd gemaakt door de student, aangezien de aan de toets-poging gekoppelde timestamp afhankelijk is van het *definitieve* invoer moment van de docent.

Tabel 4.1

Toetsen in het eerste kwartiel van de drie cohorten.

Toets	Toets	16-17	17-18	18-19	Gebruikt
Presteren en Leren	digitaal	x			ja
Diagnostisch Onderzoek	digitaal	x	x	x	ja
Professionele Gespreksvoering en Training	werkstuk	x			nee
Inleiding Psychologie (1)	digitaal	x	x	x	ja
Inleiding Psychologie (2)	werkstuk	x	x	x	ja
Practice Based Learning 1 (1)	schriftelijk	x	x	x	ja
Practice Based Learning 1 (2)	werkstuk	x	x	x	ja
Practice Based Learning 1 (3)	assessment	x	x	x	ja
Inleiding Arbeids- en Organisatiepsychologie	digitaal		x	x	ja

Blackboard

Als laatst werden gegevens van Blackboard gedownload. Het bleek dat data van studenten slechts beschikbaar was vanaf 1 januari 2018, waardoor gegevens ontbraken van het eerste kwartiel van cohorten 2016 en 2017. Om toegang te krijgen tot eerdere gegevens werd contact opgenomen met de externe beheerder van de databases (Eesysoft), waarna de Academie Mens & Arbeid toegang kreeg tot de benodigde data. Dit proces nam echter drie maanden in beslag, waardoor data kon niet eerder dan 29 november 2019 gedownload kon worden van de servers van Eesysoft. Doordat hierna beperkte tijd beschikbaar om data op te schonen, prepareren en analyseren, kon de volle potentie van deze gegevens niet worden benut.

De resulterende dataset bestaat uit het totaal aantal kliks (per week en cursus). Voor wat betreft de cursussen is dezelfde selectie gemaakt als voor Bison. Dit betekent dat klikgedrag in de cursussen verzameld is voor:

- Inleiding Arbeids- en Organisationspsychologie (INLOAP)
- Diagnostisch Onderzoek (DO)
- Inleiding Psychologie (INLPSY)
- Practice Based Learning (PBL)

Ook deze dataset bevatte missing values van de categorie *Not Missing At Random*. Studenten die geen enkele keer actief waren in de cursus, kwamen niet voor in de dataset met aantal kliks, ondanks dat zij bijvoorbeeld de toetsen wel gehaald hadden. Een missing value betekende daardoor dat de student geen enkele keer de cursus bezocht heeft, waardoor missing values vervangen kon worden met de waarde 0 (geen kliks).

Samenvoegen

Tot slot dient benoemd te worden dat het samenvoegen van bestanden per cohort gedaan werd. Om de correcte gegevens van de studenten uit cohort 2016 te verkrijgen vanuit de datasets van Bison en Blackboard, werd ten eerste een selectie gemaakt van de studenten uit het intakeassessment gecodeerd als cohort 2016. Hierna werden studenten die daar niet in deze lijst stonden verwijderd uit Bison en Blackboard. Daarna werden hun gegevens toegevoegd aan het complete databestand. Door op deze wijze gegevens samen te voegen wordt voorkomen dat de dataset *vervuild* raakt met gegevens uit andere jaren (bijvoorbeeld als gevolg van herkansing).

Na het samenvoegen werden de incomplete records uit de dataset verwijderd. Hoewel er verschillende methodes zijn om missing values in te vullen, wordt aangeraden records te verwijderen indien er daardoor niet te veel data verloren gaat. Hierdoor is de resterende data waarmee het model getraind wordt zo puur mogelijk (Kaiser, 2014).

Als laatst werd de data opgesplitst in een train- en testset. Om zoveel mogelijk een gelijke verdeling te behouden, werd eerst een onderscheid gemaakt tussen wel en niet uitgevallen studenten. Daarna werd 30% per groep toegewezen aan de testset en de overige 70% aan de trainset. Studenten werden van een willekeurige waarde voorzien van 0 tot 1 uit een uniforme distributie in lengte gelijk aan het aantal studenten in de dataset. Studenten met een waarde $< .3$ werden toegewezen aan de testset en de resterende aan de trainset. De testset werd volledig separaat gehouden van de trainset tijdens het trainen van het model om bias te voorkomen (Geurts, 2002; Hastie, Tibshirani, & Friedman, 2009).

Resulterende dataset en de ZeroR classificeerder

In tabel 4.2 is zijn de exacte aantallen te zien van de studenten in de trainset, testset en de totale dataset. Er is te zien dat de dataset na het verwijderen van missing values is gekrompen van 812 naar 693 studenten, waarvan 192 zijn toegewezen een de testset en 501 aan de training set. Bij 41.99% van deze 693 studenten was er sprake van uitval, wat betekent dat grootste groep uit 58.01% bestaat. Dit betekent dat de baseline, de ZeroR classificeerder, een accuratesse heeft van 58.01%. Met deze baseline worden de modellen vergeleken. Er is sprake van een hogere voorspellende waarde wanneer deze volgens een binomiale test significant hoger is.

Tabel 4.2

Verdeling van het aantal studenten en uitval

Dataset	Aantal	%-Aantal	Uitval	%-Uitval
Training	501	72.29%	217	43.31%
Testing	192	27.71%	74	38.54%
Totale data	693	100%	291	41.99%

4.2 Resultaten per deelvraag

Zoals in paragraaf 3.5 is benoemd worden deelvragen 1 tot en met 4 beantwoord met inzet van feature selection methodes. Voor variabelen op interval en ratio niveau wordt relevantie bepaald door de correlatie (point-biserial, r_{pb}) te berekenen tussen de features en uitval. Om de richting van de correlatiecoëfficiënten als resultaat van de point-biserial correlatie te kunnen interpreteren, is het belangrijk om te weten dat *Uitval* gecodeerd is met de waarde 1 voor “Uitval” en 0 voor “Geen uitval”. Verder, voor features van nominaal en ordinaal niveau, werd gebruik gemaakt van de Chi-Kwadrat toets met Yates’ continuïteitscorrectie om de kans op type I fouten te verlagen (Field, 2013). Hieronder worden in de sub-paragrafen de resultaten per deelvraag weergegeven.

4.2.1 Deelvraag 1

De eerste deelvraag van dit onderzoek was:

- *In hoeverre kan data van Blackboard Learn ingezet worden om studiesucces van TP-studenten te voorspellen?*

Om hier antwoord op te geven werden in totaal 121 features gegenereerd (zie bijlage B). Deze features kunnen grofweg verdeeld worden in twee groepen. De eerste groep features bestaat uit

features met totaal aan kliks. De tweede groep bevat features op binair niveau en bevatten informatie over het al dan niet actief zijn binnen bepaalde periodes.

Aantal kliks

In relatie tot het aantal kliks werden in totaal werden 34 significante correlaties gevonden, waarvan de variabelen met een met een significantieniveau $< .001$ (13 variabelen) in tabel 4.3 worden weergegeven. Alle correlaties zijn negatief, wat betekent dat minder kliks samenhangt met Uitval. In de tabel is te zien dat voor al de variabelen de correlatie zeer significant, negatief en zeer zwak is. De gevonden variabelen zijn met name afkomstig uit de vakken INLAOP (vier variabelen) en INLPSY (drie variabelen). DO komt één keer voor in de tabel en PBL helemaal niet. Ook is te zien dat de totale activiteit in alle cursussen bij elkaar opgeteld vijf keer voorkomt, waarbij de hoogste correlatie is gevonden met het aantal kliks tijdens week 8, ook wel de herfstvakantie ($r_{pb} = -0.20$, $p < .001$). Tot slot valt op dat 8 van 13 variabelen kliks betreffen vanaf week acht, oftewel de laatste vier weken van het kwartiel.

Tabel 4.3

Correlaties tussen Uitval en het totaal aantal kliks van alle vakken

Variabele	Beschrijving	correlatie*	significantie
kliks_totaal_wk_7	Totaal aantal kliks week 7	$r_{pb} = -0.17$	$p < .001$
kliks_totaal_wk_8	Totaal aantal kliks week 8	$r_{pb} = -0.20$	$p < .001$
kliks_totaal_wk_9	Totaal aantal kliks week 9	$r_{pb} = -0.18$	$p < .001$
kliks_totaal_wk_11	Totaal aantal kliks week 11	$r_{pb} = -0.18$	$p < .001$
kliks_totaal_bb	Totaal aantal kliks alle weken	$r_{pb} = -0.16$	$p < .001$
do_wk_8	Aantal kliks – DO – week 8	$r_{pb} = -0.17$	$p < .001$
inlaop_wk_3	Aantal kliks – INLAOP – week 3	$r_{pb} = -0.15$	$p < .001$
inlaop_wk_5	Aantal kliks – INLAOP – week 5	$r_{pb} = -0.16$	$p < .001$
inlaop_wk_9	Aantal kliks – INLAOP – week 9	$r_{pb} = -0.15$	$p < .001$
kliks_totaal_inlaop	Totaal aantal kliks INLAOP	$r_{pb} = -0.16$	$p < .001$
inlpsy_wk_8	Aantal kliks – INLPSY – week 8	$r_{pb} = -0.18$	$p < .001$
inlpsy_wk_9	Aantal kliks – INLPSY – week 9	$r_{pb} = -0.15$	$p < .001$
inlpsy_wk_11	Aantal kliks – INLPSY – week 11	$r_{pb} = -0.19$	$p < .001$

Opmerking. * r_{pb} , point-biserial correlatie.

Binaire variabelen

Er werden variabelen geconstrueerd op basis van het wel of niet actief zijn tijdens een periode, zoals les-, toets- en vakantieweken of een combinatie daarvan. Er zijn op basis van de Chi-kwadraat 13 variabelen gevonden waarbij een significant verschil in verhoudingen tussen het aantal wel en geen uitvallers bestaat. De variabelen worden in tabel 4.4 weergegeven. Wederom komt PBL niet voor. De variabelen laten zien dat er een aanzienlijk aantal studenten is wie niet actief is tijdens bepaalde periodes, waardoor de verhouding wel en geen uitvallers significant afwijkt van wat verwacht wordt.

Tabel 4.4

Chi-kwadraat toets voor verschillen in verhouding uitvallers op basis van het al dan niet actief zijn tijdens een periode

Vak	Actief in periode*	χ^2	significantie
INLPSY	Week 8 – vakantie	$\chi^2 = 6.58$	$p = .010$
INLPSY	Week 11	$\chi^2 = 8.27$	$p = .004$
INLPSY	Elke week actief	$\chi^2 = 7.80$	$p = .005$
INLPSY	Elke toets-week actief	$\chi^2 = 8.68$	$p = .003$
INLPSY	Elke les- en toets-week actief	$\chi^2 = 5.34$	$p = .021$
INLAOP	Week 5	$\chi^2 = 6.70$	$p = .010$
INLAOP	Week 7	$\chi^2 = 9.38$	$p = .002$
INLAOP	Week 9	$\chi^2 = 4.4$	$p = .036$
INLAOP	Week 11	$\chi^2 = 4.28$	$p = .039$
INLAOP	Alle lesweken	$\chi^2 = 9.17$	$p = .002$
DO	Week 8 – vakantie	$\chi^2 = 9.62$	$p = .002$
DO	Elke week	$\chi^2 = 4.07$	$p = .044$
DO	Elke toets-week	$\chi^2 = 5.09$	$p = .024$

Opmerking. *Lesweken = week 1 tot en met 7 + week 9; Toetsweken = week 10 + 11.

4.2.2 Deelvraag 2

De tweede deelvraag van dit onderzoek luidde:

- *In hoeverre kan data van het intakeassessment ingezet worden om studiesucces van TP-studenten te voorspellen?*

Het intakeassessment bevatte data zoals demografische gegevens, algemene gegevens (dimensie 0) en scores op capaciteiten (dimensie 1) en persoonlijkheidsschalen (dimensie 2). In totaal omvatte dit segment van de dataset 116 features. Er werden geen significantie voorspellers gevonden. Deze uitkomst is niet onverwachts, aangezien zoals in het input-through-output model (figuur 2.4) te zien is blijkt dat persoonlijkheid geen invloed heeft op persisteren of wel *geen* uitval (Van Rooij et al., 2017).

4.2.3 Deelvraag 3

Deelvraag 3 was:

- *In hoeverre kan de data van Bison ingezet worden om studiesucces van TP-studenten te voorspellen?*

De Bison dataset bevatte de resultaten van de eerste toets-kansen uit het eerste kwartiel. Ten eerste was per toets bekend of de student deze wel of niet behaald had (binair), of hij aanwezig was (binair) en welk cijfer behaald werd. Van de behaalde cijfers werden gemiddelden berekend en een feature met het totaal aantal behaalde toetsen werd aangemaakt. In totaal bestond deze dataset uit 23 features.

Toets cijfers

In tabel 4.5 worden de resultaten getoond van cijfer, gemiddelden en aantal behaalde toetsen, waarin te zien is dat met name de kennistoetsen, het gemiddelde cijfer en het aantal behaalde toetsen het sterkst correleren met uitval. Alle correlaties zijn negatief, wat betekent dat een lager cijfer samenhangt met uitval en een hoger cijfer met geen uitval. Alle correlaties zijn zeer zwak, met uitzondering van het gemiddelde cijfer ($r_{pb} = -0.34, p < .001$) en het aantal behaalde toetsen ($r_{pb} = -0.40, p < .001$), welke laag zijn en zeer significant ($p < .001$). De behaalde cijfers voor PBL blijken tevens, net als in sub-paragraaf 4.2.1, het minst van belang in relatie tot Uitval.

Tabel 4.5

Resultaten van feature selection op de toets-cijfers uit de Bison dataset

Variabele	Beschrijving	stat.*	sig.**
t_resultaat_pbl1_as	Cijfer assessment PBL	$r_{pb} = -0.15$	$p < .001$
t_resultaat_pbl1_sc	Cijfer schriftelijke communicatie PBL	$r_{pb} = -0.13$	$p = .005$
t_resultaat_inlpsy_dig	Cijfer kennistoets INLPSY	$r_{pb} = -0.30$	$p < .001$

t_resultaat_inlpsy_wk	Cijfer werkstuk INLPSY	$r_{pb} = -0.11$	$p = .011$
t_resultaat_inlaop_dig	Cijfer kennistoets INLAOP	$r_{pb} = -0.28$	$p < .001$
t_resultaat_do_dig	Cijfer kennistoets DO	$r_{pb} = -0.24$	$p < .001$
t_aantal_behaalde_toetsen	Aantal behaalde toetsen	$r_{pb} = -0.40$	$p < .001$
t_toets_mean	Gemiddelde resultaat	$r_{pb} = -0.34$	$p < .001$

Opmerking. *Test statistiek: r_{pb} , point-biserial correlatie. **significantie niveau.

Binaire variabelen uit Bison

Uit de dataset werden vervolgens nog vier significante features gevonden in relatie tot Uitval. Deze worden in tabel 4.6 getoond. Er is te zien dat het behalen van de kennistoetsen voor INLPSY, INLAOP en DO voor een onverwachte verhouding zorgt in het aantal uitvallers. Dit geldt tevens voor het behalen van de toets schriftelijke communicatie van PBL ($\chi^2 = 4.98$, $p = .026$). Het effect van deze toets is echter lager dan de eerder genoemde kennis toetsen.

Tabel 4.6

Resultaten van feature selection op basis van chi-kwadraat test voor binaire variabelen – het al dan niet behalen van toetsen – uit de Bison dataset

Vak	Toets	χ^2	significantie
INLPSY	Kennistoets	$\chi^2 = 19.06$	$p < .001$
INLAOP	Kennistoets	$\chi^2 = 17.08$	$p < .001$
DO	Kennistoets	$\chi^2 = 9.70$	$p = .002$
PBL	Schriftelijke communicatie	$\chi^2 = 4.98$	$p = .026$

4.2.4 Deelvraag 4

Om antwoord te geven op de vierde, en daarmee laatste, deelvraag van het onderzoek werden de resultaten van deelvragen een tot en met drie gecombineerd. Deelvraag 4 luidde als volgt:

- *In welke mate kan de data met elkaar gecombineerd worden om studiesucces van TP-studenten te voorspellen?*

In paragraaf 3.5 werd belicht dat de features als resultaat van feature selection (deelvraag een tot en met drie) gebruikt werden in de C5.0, CART, CHAID, Exhaustive CHAID en QUEST decision trees. In totaal werden 42 variabelen gevonden die gebruikt kunnen worden voor de decision trees. Dit aantal is echter zeer hoog en veel van deze features meten tot op zekere hoogte hetzelfde. Verder zijn er zowel binaire features als features op interval en ratio niveau

welke wederom tot op zekere hoogte hetzelfde meten. Er is daarom de keuze gemaakt twee groepen modellen te generen. De eerste groep modellen wordt gegenereerd op basis van de features op interval en ratio niveau. Bij de tweede groep worden de binaire features gebruikt.

De decision trees van een algoritme is valide wanneer de accuratesse op zowel de train- als testset ongeveer gelijk is, wat in dit onderzoek een maximaal verschil van 5% inhoudt. Vervolgens, om te bepalen in hoeverre er sprake was van een hogere accuratesse, werd de accuratesse per algoritme vergeleken met de ZeroR classificeerder door een binomiale toets toe te passen. Voor wat betreft de ZeroR classificeerder geldt dat bij 41.99% ($n = 291$) van de studenten sprake was van uitval, betekende dat de grootste groep 58.01% ($n = 402$) van de studenten bevat. De baseline heeft daarom een accuratesse van 58.01%.

In totaal werden 10 modellen genereerd, waardoor het significantie niveau (α) voor de binomiale test op basis van de Bonferroni correctie verlaagd wordt naar $\alpha = .005$. Er is dus sprake van een significante verbetering in accuratesse ten opzichte van de baseline bij $p < .005$.

Modellen op basis van features op interval en ratio niveau

De algoritmes van de eerste groep (features op interval en ratio niveau) werden met standaard instellingen gedraaid, met uitzondering van de C5.0. Voor de C5.0 werd de *Pruning Severity* aangepast tot het verschil in accuratesse minimaal was, aangezien dit algoritme als enige in de standaard settings geen rekening houdt met de maximale diepte van het model, waardoor overfitting plaatsvindt. Dit resulteert in een groot verschil tussen accuratesse op de train- en testset.

In tabel 4.7 worden de resultaten van het modelleren getoond. De CHAID en Exhaustive CHAID worden niet nader geanalyseerd aangezien de verschillen tussen de train- en testset voor beide meer dan 5% zijn. De accuratesse van de overige algoritme werd vergeleken met de accuratesse van de baseline op basis van een binomiale test. De CART presteert het best. De binomiale test laat zien dat het aantal correcte voorspelling door de CART met .7374 hoger is dan het verwachte aantal van .5801 ($p < .001$). In bijlage C, figuur C1, wordt de decision tree getoond.

Tabel 4.7

Accuratesse van de verschillende algoritmes op basis van de features op interval en ratio niveau

Algoritme	Validering accuratesse			TACC*	Base**	Sig***
	Train	Test	Verschil			
CART	73.85%	73.44%	0.41%	73.74%	58%	$p < .001$
C5.0	72.85%	71.88%	0.97%	72.58%	58%	$p < .001$
CHAID	73.45%	70.31%	5.14%	-	58%	-
E. CHAID	73.85%	68.75%	5.10%	-	58%	-
QUEST	72.06%	73.44%	1.38%	72.44%	58%	$p < .001$

Opmerking. *Accuratesse van definitieve model, enkel van toepassing bij verschil $< 5\%$.

Accuratesse van de baseline .*significantie.

Modellen op basis van binaire features

Voor het generen van decision trees op basis van de nominale features werden alle features van tabel 4.5 en tabel 4.6 gebruikt. In tabel 4.8 is te zien dat alle algoritmes in staat waren een model te generen. Het verschil tussen de accuratesse op de train- en testset is het grootst voor de CART (3.46%) en het laagst bij de C5.0 (0.39%). De QUEST presteert heeft de laagste accuratesse (69.12%). De hoogste accuratesse werd behaald door de C5.0. De binomiale test laat zien dat het aantal correcte voorspelling door de C5.0 met .7316 hoger is dan het verwachte aantal van .5801 ($p < .001$). In bijlage C, figuur C2, wordt de decision tree getoond.

Tabel 4.8

Accuratesse van de verschillende algoritmes op basis van binaire features

Algoritme	Validering accuratesse			TACC*	Base**	Sig***
	Train	Test	Verschil			
CART	73.25%	69.79%	3.46%	72.29%	58%	$p < .001$
C5.0	73.05%	73.44%	0.39%	73.16%	58%	$p < .001$
CHAID	73.05%	72.40%	0.65%	72.87%	58%	$p < .001$
E. CHAID	73.05%	72.40%	0.65%	72.87%	58%	$p < .001$
QUEST	68.66%	70.31%	1.65%	69.12%	58%	$p < .001$

Opmerking. *Accuratesse van definitieve model, enkel van toepassing bij Verschil $< 5\%$.

Accuratesse van de baseline .*significantie.

Gebruikte features

In tabel 4.9 is te zien welke features meegegeven werden aan de algoritmes voor de eerste groep modellen (zie tabel 4.7) en op basis van welke features vervolgens modellen genereerd werden. Ook is te zien welke features voor het betreffende algoritme het meest relevant waren om uitvallers en doorstromers van elkaar te kunnen onderscheiden. De relevantie wordt uitgedrukt in *predictor importance*, wat een maat (percentage) is om het belang van de feature in het onderscheidend vermogen uit te drukken in relatie tot de overige features. De CHAID en Exhaustive CHAID zijn weggelaten aangezien deze algoritme niet in staat waren decision trees te genereren die als valide beschouwd konden worden. Er is te zien dat de resultaten voor de kennistoetsen van INLPSY en INLAOP in elk algoritme zeer relevant waren. Daarnaast blijken het aantal kliks aan het begin (CART en C5.0) en eind (QUEST) van INLAOP van belang. Tabel 4.10 laat vervolgens een soortgelijk overzicht zien voor de tweede groep algoritmes (binaire features). Ook voor deze groep modellen is te zien dat de toetsen zeer relevant zijn voor het onderscheiden van wel en geen uitvallers.

Tabel 4.9

Gebruikte features van de eerste groep modellen

Feature	CART	C5.0	QUEST
t_resultaat_inlpsy_dig	.44	.43	.45
t_resultaat_inlaop_dig	.43	.18	.46
t_resultaat_do_dig	--	.21	--
do_wk_8	--	--	--
inlpsy_wk_8	--	--	--
inlpsy_wk_9	--	--	--
inlpsy_wk_11	--	--	--
inlaop_wk_3	.13	.18	--
inlaop_wk_5	--	--	--
inlaop_wk_9	--	--	.09

Tabel 4.10

Gebruikte features met predictor importance van de tweede groep modellen

Feature	CART	C5.0	CHAID	ECHAID	QUEST
t_behaald_inlpsy_dig	.33	.45	.37	.37	.87
t_behaald_inlaop_dig	.25	.17	.13	.13	.07
t_behaald_do_dig	.12	.06	.13	.13	.07
t_behaald_pbl1_sc	.06	.08	.02	.02	--
inlpsy_wk_8_actief	--	--	.06	.06	--
inlpsy_wk_11_actief	--	--	--	--	--
inlpsy_elke_week_actief	--	.06	--	--	--
inlpsy_alle_toetsweken_actief	--	--	.15	.15	--
inlpsy_alle_les_en_toetsweken_actief	--	--	.03	.03	--
inlaop_wk_5_actief	--	--	.06	.06	--
inlaop_wk_7_actief	.15	.14	.04	.04	--
inlaop_wk_9_actief	--	--	.02	.02	--
inlaop_wk_11_actief	--	--	--	--	--
do_wk_8_actief	.09	--	--	--	--
do_alle_toetsweken_actief	--	.03	--	--	--

Bij het interpreteren van de decision trees dient men echter bewust te zijn dat hier over accuratesse gesproken wordt, oftewel, het aantal correct voorspelde gevallen. Meer informatie over modellen kan gewonnen worden door naar de betreffende *confusion matrix* te kijken. In een confusion matrix is te zien hoeveel gevallen correct voorspeld zijn in de vorm van *true positives* (Tp) en *true negatives* (Tn). Daarnaast kan men hierin ook het aantal *false positives* (Fp) en *false negatives* (Fn) vinden. Accuratesse wordt vervolgens uitgedrukt als:

$$\frac{(Tp + Tn)}{(Tp + Fn + Tn + Fp)}$$

Andere maten zijn:

1. *Sensitivity*. Hoeveel relevante gevallen (uitvallers) correct geïdentificeerd worden:

$$\frac{Tp}{Tp + Fn}$$

2. *Specificity*. Hoe goed het model in staat is een vals alarm te vermijden:

$$\frac{Tn}{Tn + Fp}$$

3. *Precision*. Hoeveel van de classificaties Uitval correct zijn:

$$\frac{Tp}{Tp + Fp}$$

Met deze waardes kan men vervolgens optimale modellen kiezen (Nasa & Suman, 2012), waar in dit onderzoek geen rekening mee is gehouden en meer inzicht ontbreekt.

Hoofdstuk 5. Conclusie, discussie en aanbevelingen

Hoofdstuk 5 bestaat uit de conclusie, discussie en aanbevelingen. In paragraaf 5.1 worden de resultaten geïnterpreteerd en bediscussieerd, waarna wordt nagegaan in hoeverre het onderzoek betrouwbaar, valide en bruikbaar is. Op basis daarvan wordt in paragraaf 5.2 een aanbeveling gedaan aan de opdrachtgever over de vervolgstappen.

5.1 Conclusie en discussie

Dit onderzoek had het doel een model te ontwikkelen op basis van educational data mining, waarmee studiesucces (uitval in het eerste jaar) van TP-studenten voorspeld kon worden. Belangrijk hierbij was dat het voor diverse partijen, zowel bekend als onbekend met het vakgebied, gemakkelijk in gebruik te nemen is. De vraag ontstond in welke mate de beschikbare data van Saxion (Blackboard Learn, het intakeassessment en Bison) daartoe in staat was.

5.1.1 Antwoord op de hoofdvraag en deelvragen

Deelvraag 1:

- *In hoeverre kan data van Blackboard Learn ingezet worden om studiesucces van TP-studenten te voorspellen?*

De features vanuit Blackboard Learn bestonden puur uit het totaal aantal kliks en het al dan niet actief zijn in de diverse cursussen, met een nader onderscheid in de lesweken van het eerste kwartiel. Er is dus veel ruimte om deze features te verfijnen en meer informatie te winnen uit de blackboard data, zoals bijvoorbeeld het (op tijd) maken van oefentoetsen en inleveren van verslagen. De gebruikte features waren dus vrij elementair. Desondanks was meer dan een derde ($\frac{47}{121} = 38.84\%$) van de features relevant, waarvan $\frac{13}{47} = 27.66\%$ zelfs zeer relevant. Opvallend was dat de activiteiten aan het eind van het kwartiel en in de vakken INLAOP en INLPSY veruit het meest van belang zijn, hoewel de gevonden samenhang tussen dezen features en uitval echter zeer laag is. Er kunnen vervolgens een aantal conclusies worden getrokken. Ten eerste is er bij studenten die actiever zijn aan het eind van het kwartiel minder sprake van uitval. Ten tweede blijkt dat de Blackboard data voor de vakken PBL en DO vrijwel van geen toegevoegde. Ten derde wordt geconcludeerd dat vanwege het elementaire niveau van de features er nog veel meer ruimte is voor het winnen van informatie uit de data.

Deelvraag 2:

- *In hoeverre kan data van het intakeassessment ingezet worden om studiesucces van TP-studenten te voorspellen?*

Er werden geen relevantie voorspellers gevonden in de data vanuit het intakeassessment. Kijkend naar het model van Van Rooij et al. (2017), is dit echter geen onverwacht resultaat. Persisteren, oftewel *geen* uitval, is met name afhankelijk van through-put factoren: factoren tijdens de studie. Zowel de uitgevallen als doorgestroomde studenten laten dus een gelijk beeld zien in relatie tot de capaciteiten, persoonlijkheid en algemene gegevens. De conclusie luidt dat het intakeassessment niet relevant is om studiesucces van de studenten te voorspellen.

Deelvraag 3:

- *In hoeverre kan de data van Bison ingezet worden om studiesucces van TP-studenten te voorspellen?*

Uit de analyse van de Bison dataset blijken de kennis toetsen veruit het meeste van belang in het kader van studiesucces voorspellen. Vier van de zeven toetsen waren zeer significante voorspellers, waarin alle (drie) kennistoetsen voorkwamen. Verder blijkt het ook aantal behaalde toetsen tijdens van het eerste kwartiel (eerste poging) en het gemiddelde cijfer van deze eerste pogingen samen te hangen met uitval. Dit resultaat was te verwachten. Voor zowel Dekker, Pecheniznky en Vleeshouwers (2009), Mueen, Zafar en Manzoor (2016) en Conijn et al. (2016) waren tussentijdse toets-cijfers de belangrijkste voorspellers van studiesucces. Er wordt daarom geconcludeerd dat de gevonden features relevant zijn en Bison over waardevolle informatie beschikt voor het voorspellen van studiesucces.

Deelvraag 4:

- *In welke mate kan de data met elkaar gecombineerd worden om studiesucces van TP-studenten te voorspellen?*

In totaal konden zeven besluitbomen gegeneerd worden. Er werd gevonden dat de kennistoetsen van INLPSY en INLAOP het belangrijkste zijn om onderscheid te maken tussen wel en geen uitval. Op basis van enkel de toetsen en één (CART decision tree met features p[interval en ratio niveau) tot drie (C5.0 decision tree met binaire features) features uit de Blackboard data werd al een aanzienlijke toename van meer dan 15% behaald in het aantal correcte voorspellingen. De toets resultaten waren daarbij echter veruit het meest van belang,

terwijl de Blackboard data over zeer weinig onderscheidend vermogen beschikte. De data kan dus in beperkte mate met elkaar gecombineerd worden.

Hoofdvraag:

- *In welke mate kan de beschikbare data studiesucces van TP-studenten voorspellen?*

Er kan geconcludeerd worden dat op basis van de beschikbare data het mogelijk is om tot wel ruim 73% van de TP-studenten correct te kunnen classificeren als wel of niet uitvallers. De meeste informatie werd gewonnen uit Bison (toets-cijfers eerste toets-kans tijdens het eerste kwartiel) en Blackboard Learn (Bb). Het intakeassessment bevat geen relevante informatie waar studiesucces (uitval) mee voorspelt kan worden.

5.1.2 Discussie

Opvallend is dat ondanks het hoge aantal gevonden features in de Blackboard data slechts een klein aantal daarvan in de decision trees als belangrijke voorspellers werden gevonden. Zoals echter ook in de beantwoording van de deelvraag werd benoemd kan dit wellicht verklaard worden vanuit het elementaire niveau van de features. Er was beperkte tijd beschikbaar de data op te schonen, waardoor ook minder inzicht in de gedragingen van studenten gevonden kon worden in de diverse vakken. Deze web-data is echter vaak van groot belang in educational data mining, in die zin dat daarin de meeste informatie wordt gevonden, zoals het gebruik maken van resources, leergedrag en prestaties (Romero et al., 2010), welke vanuit het input-throughput-output model van Van Rooij et al. (2017) kunnen deze gedefinieerd worden als de throughput factoren, de belangrijkste voorspellers van persisteren, oftewel *geen* uitval. Hierdoor kan afgevraagd worden of de Blackboard van DO en PBL wellicht toch over meer relevante informatie beschikt.

Verder kan opgemerkt worden dat voor het ontwikkelen van het model enkel gebruik gemaakt werd van een train- en testset, waarnaast feature selection werd gedaan op basis van correlaties en Chi-Kwadraat toetsen. Een andere optie is echter om *k*-Fold Cross-Validation toe te passen. Hierbij verdeelt men de data nog steeds in een train- en testset, waarna de trainset vervolgens in *k* groepen (folds) verdeeld wordt. Vervolgens wordt een model ontwikkeld met $k - 1$ folds en wordt deze geëvalueerd op de separaat gehouden fold. Dit wordt voor alle folds gedaan. Op deze wijze kunnen features gerangschikt en geëvalueerd worden om het uiteindelijke model te optimaliseren naar de beste set met features. Deze methode biedt over het algemeen de voorkeur wanneer men over kleinere datasets beschikt, zoals in dit onderzoek

het geval was. Het leidt in de meeste gevallen namelijk tot een meer valide en betrouwbaarder model (Hastie, Tibshirani, & Friedman, 2009).

5.1.3 Betrouwbaarheid, validiteit en bruikbaarheid

Betrouwbaarheid

Van de 812 studenten waren van 693 studenten (85.34%) alle gegevens beschikbaar. Vervolgens werd onder deelvraag 4 in dit hoofdstuk benadrukt dat de omvang van de dataset voldoende was voor het ontwikkelen van een model. Verder, omdat de gegevens opgeslagen zijn in de databases van Saxion, zijn deze altijd in zuivere vorm beschikbaar, zonder dat zij aangetast kunnen worden. Het hele proces is daarnaast uitgebreid gedocumenteerd en te vinden op de servers van Saxion, waardoor het volledig navolgbaar is. Echter, omdat er weinig tijd besteed is aan de data understanding en preparation fases van de CRISP-DM methode (Chapman et al., 2000) is het onduidelijk in hoeverre de data van Blackboard voldoende opgeschoond is in het kader van bias. Het onderzoek is vanwege de navolgbaarheid betrouwbaar, waarbij kanttekeningen gezet moeten worden in de kwaliteit van de data (Verhoeven, 2014).

Validiteit

In het kader van interne validiteit is zorgvuldig omgegaan met de data. Studenten werden volgens randomisatie toegevoegd en in de groepen verdeeld (train- en testset), waarna deze van elkaar gescheiden werden gehouden tot het eind van het onderzoek. Daarnaast werd strikt aan voorgeschreven methodes gehouden (CRISP-DM), waardoor de interne validiteit hoog is. Voor wat betreft de externe validiteit doet de omvang en representativiteit ervan in eerste instantie vermoeden dat de resultaten te generaliseren zijn naar alle TP-studenten en de externe validiteit daarom hoog is (Verhoeven, 2014). Dit valt echter in twijfel te trekken. Ten eerste door de beperkte mate waarin het resulterende model geëvalueerd is. Het model is slechts op basis van accuratesse geëvalueerd, terwijl meer informatie over het model te winnen valt door andere evaluatie methodes zoals sensitivity en precision (Nasa & Suman, 2012). Dit gaat dus ten koste aan de externe validiteit. Ten tweede had de externe validiteit verhoogd kunnen worden door inzet van *k*-fold Cross-Validation (Hastie, Tibshirani, & Friedman, 2009).

Bruikbaarheid

De resterende vraag is vervolgens hoe Saxion de uitkomsten van dit onderzoek kan gebruiken. Wat het onderzoek heeft laten zien is dat het met Educational Data Mining mogelijk is een

gemakkelijk te interpreteren model te ontwikkelen op basis van beschikbare gegevens in de databases. De potentie van de data is echter niet vol benut en er is veel ruimte beschikbaar ter optimalisatie. Dit had voor een grotendeels te maken met het feit dat de data gedurende een periode van drie maanden niet beschikbaar was, aangezien een externe partij deze data beheert waar Saxion niet direct toegang tot heeft. Het is daarom de moeite waard te investeren in het verbeteren van de beschikbaarheid en kwaliteit van data om de volle potentie van Educational Data Mining te benutten. Kortom, het advies luidt het huidige model door te ontwikkelen alvorens deze in gebruik te nemen, waarover hieronder in de aanbevelingen wordt geschreven.

5.2 Aanbevelingen

Zoals in hoofdstuk 1 werd beschreven bestaat bij Saxion het probleem dat zij niet in staat is het studiesucces van TP-studenten te kunnen voorspellen. De data waarover Saxion beschikt leek daartoe onvoldoende in staat (Ambagts, 2018; Wilmer, 2019; Jansen, 2019). Het vermoeden was ontstaan dat EDM daar mogelijk een uitkomst voor kon bieden, aldus de aanleiding voor dit onderzoek. Zoals gebleken in de voorgaande paragraaf beschikt de hogeschool wel degelijk over waardevolle informatie in haar databases, waarbij de belangrijkste informatie te vinden is in de tussentijdse prestaties op toetsen. Er werden modellen mee ontwikkeld welke uitval van TP-studenten kunnen voorspellen. De vraag die overblijft is hoe Saxion in de toekomst verder kan.

Het door ontwikkelen van het model

In de resultaten (hoofdstuk 4) werd, ten eerste, vermeld dat in verband met de late ter beschikking stellen van de data afkomstig uit Blackboard Learn, deze zeer beperkt benut kon worden. Dit terwijl, zoals blijkt uit Romero et al. (2010), juist deze (LMS-)data de hogeschool van waardevolle informatie kan voorzien. Ten tweede kon ook het intakeassessment zeer beperkt in verband gebracht worden met studiesucces. Een derde beperking ligt in de ontwikkeling van het model en de mate waarin het geëvalueerd werd. Deze drie punten met elkaar gecombineerd betekenen dat er veel ruimte beschikbaar is voor verbetering. Hieronder worden daarom drie (samenhangende) onderzoeken voorgesteld om tot een verbeterd model (hogere accuratesse) te komen die in gebruik genomen kan worden om risico studenten (studenten die een hoge kans hebben uit te vallen) te identificeren.

Onderzoek 1

Een eerste onderzoek richt zich op de vraag in hoeverre meer relevante gegevens geïnventariseerd kunnen worden tijdens het intakeassessment in kader van voorspellen van studiesucces met EDM. Wat zowel uit dit onderzoek als eerdere (Ambagts, 2018; Wilmer, 2019) blijkt, is dat het intakeassessment zeer beperkt in staat is studiesucces te voorspellen. Denkend aan het input-throughput-output model van Van Rooij et al. (2017), wordt aangeraden te focussen op input factoren. In sub-paragraaf 2.2.3 werd gezien dat veel studies gebruik maken van factoren vanuit vooropleiding en de omgeving van aspirant studenten. Hierbij kan bijvoorbeeld gedacht worden aan socio-demografische gegevens, prestaties tijdens de vooropleiding en opleidingsachtergrond van de ouders. De centrale vraag die beantwoord dient te beantwoorden is hoe en welke (input) factoren de hogeschool het best in kaart kan brengen om studiesucces van TP-studenten te kunnen voorspellen met EDM. Een vragenlijst lijkt daar een goed instrument voor (Osmanbegović & Suljić, 2012; Ramesh, Parkavi, & Ramar, 2013). Dit onderzoek kan in een half jaar worden uitgevoerd.

Onderzoek 2

Parallel lopend aan *onderzoek 1* kan onderzocht in hoeverre diverse feature selection en evaluatie methodes verschillende (of betere) resultaten behalen voor het ontwikkelen van een voorspellend model. In dit onderzoek werd feature selection gedaan op basis van correlaties en Chi-Kwadraat toetsen, terwijl een mogelijkheid bestaat dit op basis van bijvoorbeeld Cross Validation te doen, wat op basis van wat in de literatuur gevonden is de voorkeur biedt zie sub-paragraaf 5.1.2). Een onderzoek waarin dus verschillende feature selection technieken vergeleken worden kan helpen inzicht te krijgen hoe data optimaal benut kan worden. Daarnaast leidt meer inzicht in het evaluatieproces van modellen (zoals de ROC curve) ertoe dat modellen beter op hun validiteit en betrouwbaarheid beoordeeld en geselecteerd kunnen worden (Nasa & Suman, 2012). Het eindresultaat van dit onderzoek moet laten zien *hoe* men deze processen in de praktijk uitvoert, zodat dit in het laatste onderzoek gedaan kan worden. Het onderzoek kan in een periode van zes maanden worden uitgevoerd.

Onderzoek 3

Vervolgens kan in een derde onderzoek een nieuw model ontwikkeld worden. Men gaat eerst na in hoeverre meer informatie gewonnen kan worden uit de databases van Blackboard Learn, aangezien daar in dit onderzoek beperkt aandacht aan besteed kon worden. Voorbeelden van deze informatie zijn prestaties op proeftoetsen en het op tijd inleveren van opdrachten. Want,

zoals uit dit onderzoek blijkt, *wanneer* een student actief is in de modules is een belangrijke voorspeller, waardoor een interessant gegeven kan zijn *welke* activiteiten dat precies betreft. Om deze data te verwerken wordt aangeraden Python te benutten, aangezien daarmee handelingen geautomatiseerd kunnen worden die anders zeer tijdrovend zijn (zie bijlage E). Deze data kan dan gecombineerd worden met de gegevens van Bison en de gegevens van het hiervoor genoemde *onderzoek 1*. Voor het ontwikkelen van het model kan men de resultaten van *onderzoek 2* gebruiken voor een optimaal feature selection en evaluatie proces. Dit kan in een periode van 6 maanden worden uitgevoerd. Dit betekent dat binnen een periode van één jaar de drie onderzoeken uitgevoerd kunnen worden, waarna men over een verbeterd model beschikt.

Literatuurlijst

- Abu Tair, H. M., El-Halees, A. M. (2012). Mining Educational Data to Improve Students' Performance: A Case Study. *International Journal of Information and Communication Technology Research*, 2(2), 140-146.
- Alharbi, Z., Cornford, J., Dolder, L., & De La Iglesia, B. (2016, July). Using Data Mining Techniques to Predict Students at Risk of Poor Performance. In *2016 SAI Computing Conference (SAI)* (pp. 523-531). IEEE.
- Ambagts, J. (2018). *Een Onderzoek naar de Determinanten van Studiesucces en een Vergelijking van de Uitkomsten verkregen met een Klassieke Methode voor Data Analyse en Data Mining* (Thesis).
- Azevedo, A., Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: A Parallel Overview. In *Proceedings of the IADIS European Conference Data Mining* (pp. 182-185).
- Baars, G. J., Stijnen, T., & Splinter, T. A. (2017). A Model to Predict Student Failure in the First Year of the Undergraduate Medical Curriculum. *Health Professions Education*, 3(1), 5-14.
- Baker, R. S., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1(1), 3-17.
- Baradwaj, B. K., & Pal, S. (2011). Mining Educational Data to Analyze Students' Performance. *International Journal of Advanced Computer Science and Applications*, 2(6), 63-69.
- Biggs, J. B., Kember, D., & Leung, D. Y. P. (2001). The Revised Two Factor Study Process Questionnaire: R-SPQ-2F. *British Journal of Educational Psychology*, 71, 133-149.
- Braxton, J. M., Millem, J. F., & Sullivan, A. S. (2000). The Influence of Active Learning on the College Student Departure Process: Toward a Revision of Tinto's Theory. *The Journal of Higher Education*, 71(5), 569-590.
- Brookshear, J. (2007). *Computer Science: An Overview* (9th ed.). Boston, MA: Addison-Wesley Publishing Company.
- Bussemaker, J. (2016, 17 november). Kamerbrief over eindbeoordeling prestatieafspraken hoger onderwijs [Kamerbrief]. Geraadpleegd van <https://www.rijksoverheid.nl/documenten/kamerstukken/2016/11/17/kamerbrief-over-eindbeoordeling-prestatieafspraken-hoger-onderwijs>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-Step Data Mining Guide*. Opgehaald van <https://the-modeling-agency.com/crisp-dm.pdf>

- Conijn, R., Kleingeld, A., Matzat, U., Snijders, C., Van Zaanen, M. (2016). Influence of Course Characteristics, Student Characteristics, and Behavior in Learning Management Systems on Student Performance. In *Neural Information Processing Systems (NIPS) Workshop on Machine Learning for Education 2016*.
- Dekker, G. W., Pecheniskiy, M., & Vleeshouwers, J. W. (2009). Predicting Students Drop Out: A Case Study. *International Working Group on Educational Data Mining*, 41-50.
- Fernandes, C. M., Da Silva, W. T., & Ladeira, M. (2004). An Algorithm to Handle Structural Uncertainties in Learning Bayesian Network. In *Proceedings of Ibero-American Symposium on Software Engineering and Knowledge Engineering*.
- Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational Data Mining: Predictive Analysis of Academic Performance of Public School Students in the capital of Brazil. *Journal of Business Research*, 94, 335-343.
- Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. London, England: Sage.
- Garrison, D. R., Kanuka, H. (2004). Blended Learning: Uncovering its Transformative Potential in Higher Education. *The Internet and Higher Education*, 7(2), 95-105.
- Geurts, P. (2002). *Contributions to decision tree induction: bias/variance tradeoff and time series classification* (Dissertatie, Universiteit van Luik).
- Hart, H., Boeije, H., & Hox, J. (2007). *Onderzoeksmethoden*. Amsterdam, Nederland: Boom Uitgevers.
- Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2e ed.). New York, USA: Springer-Verslag New York.
- Heppen, J. B., Theriault, S. B. (2008). Developing Early Warning Systems to Identify Potential High School Dropouts. *National High School Center, American Institutes for Research*, 1-13.
- Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., & Murray, D. J. (2019). Identifying Key Factors of Student Academic Performance by Subgroup Discovery. *International Journal of Data Science and Analytics*, 7(3), 227-245.
- IBM (2019). *IBM SPSS Modeler 18.2.1 Modeling Nodes: Instruction Manual*. Chicago, IL: IBM.
- Jansen, R. (2019). *Afstudeeronderzoek naar het Gebruik van Blackboard Learn en het Effect op Studieresultaten* (Thesis).
- Jonkman, A. (z.d.). Onderwijs. Geraadpleegd van <https://www.vereniginghogescholen.nl/themas/onderwijs>

- Kabakchieva, D. (2013). Predicting Student Performance by Using Data Mining Methods for Classification. *Cybernetics and Information Technologies*, 13(1), 61-72.
- Kaiser, J. (2014). Dealing with Missing Values in Data. *Journal of Systems Integration*, 1, 42-51.
- Lenz, P. H., McCallister, J. W., Luks, A. M., Le, T. T., & Fessler, H. E. (2015). Practical Strategies for Effective Lectures. *Annals of the American Thoracic Society*, 12(4), 561-566.
- Lizzio, A., & Wilson, K. (2013). Early Intervention to Support the Academic Recovery of First-Year Students at Risk of Non-Continuation. *Innovations in Education and Teaching International*, 50(2), 109-120.
- Lopez, M. I., Luna, J. M., Romero, C., Ventura, S. (2012). Classification via Clustering for Predicting Final Marks based on Student Participation in Forums. In *Proceedings of the 5th International Conference on Educational Data Mining* (pp. 148-151).
- Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early Dropout Prediction using Data Mining: A Case Study with High School Students. *Expert Systems*, 33(1), 107-124.
- Mueen, A., Zafar, B., Manzoor, U. (2016). Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. *International Journal of Education and Computer Science*, 11, 36-42.
- Nasa, C., Suman, S. (2012). Evaluation of Different Classification Techniques for WEB Data. *International Journal of Computer Applications*, 52(9), 34-40.
- Onderwijsinspectie. (2019). *De staat van het hoger onderwijs 2019*. Geraadpleegd van <https://www.onderwijsinspectie.nl/onderwerpen/staat-van-het-onderwijs/documenten/rapporten/2019/04/10/deelrapport-hoger-onderwijs>
- Osmanbegović, E., Suljić, M. (2012). Data Mining Approach for Predicting Student Performance. *Economic Review: Journal of Economics and Business*, 10(1), 3-12.
- Ramaswami, M., Bhaskaran, R. (2009). A Study on Feature Selection Techniques in Educational Data Mining. *Journal of Computing*, 1(1), 7-11.
- Ramaswami, M., Bhaskaran, R. (2010). A CHAID Based Performance Prediction Model in Educational Data Mining. *International Journal of Computer Science Issues*, 7(1), 10-18.
- Ramesh, V., Parkavi, P., Ramar, K. (2013). Predicting Student Performance: A Statistical and Data Mining Approach. *International Journal of Computer Applications*, 63(8), 35-39.

- Romero, C., & Ventura, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. *Expert System with Applications*, 33(1), 135-146.
- Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. J. (Red.). (2010). *Handbook of educational data mining*. Boca Raton, FL: CRC Press.
- Saa, A. A. (2016). Educational Data Mining & Students' Performance Prediction. *International Journal of Advanced Computer Science and Applications*, 7(5), 212-220.
- Salzberg, S. L. (1997). On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. *Data Mining and Knowledge Discovery*, 1, 317-327.
- Sathya, R., & Abraham, A. (2013). Comparison of Supervised and Unsupervised Learning Algorithms for Patter Classification. *International Journal of Advanced Research in Artificial Intelligence*, 2(2), 34-38.
- Saxion. (z.d.-a). Studie-inhoud. Geraadpleegd van <https://www.saxion.nl/opleidingen/voltijd/bachelor/toegepaste-psychologie/studie-inhoud>
- Saxion. (z.d.-b). *Toegepaste Psychologie: Het brein als domein*. Geraadpleegd van https://www.saxion.nl/binaries/content/assets/opleidingen/voltijd/bachelor/toegepaste-psychologie/003-2027-brochure-toegepaste-psychologie_lr-def_2.pdf
- Saxion. (2016). *Saxion Jaarverslag 2016*. Geraadpleegd van <https://www.saxion.nl/binaries/content/assets/over-saxion/organisatie/anbi/jaarverslag-saxion-2016.pdf>
- Saxion. (2018). *Saxion Jaarverslag 2018*. Geraadpleegd van <https://viewer.wepublish.com/jaarverslag-2018>
- Saxion. (2019). *Onderwijs- en Examenregeling voor bacheloropleiding Human Resource Management (voltijd en deeltijd) en Toegepaste Psychologie (voltijd en deeltijd) van de Academie Mens & Arbeid (AMA) van Saxion Hogeschool*. Geraadpleegd van: https://resolver.saxion.nl/serve_oer/2019_2020/ama-hrm-tp/bachelor_nl/oer.pdf
- Seifert, K. (2011). *Educational Psychology*. OpenStax CNX. Geraadpleegd van <https://cnx.org/contents/zmxetoTT>
- Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J., & Abreu, R. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. *International Educational Data Mining Society*.
- Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Tree, Bagging and Random Forests. *Psychological Methods*, 14(4), 323-348.

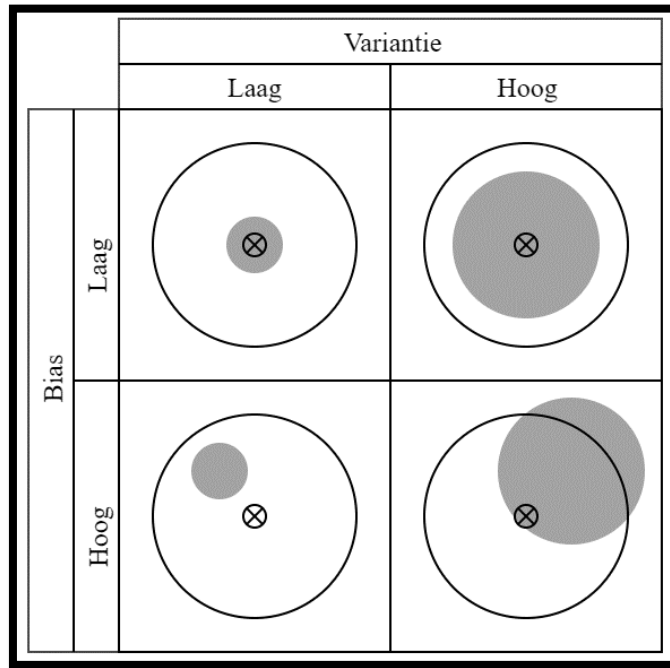
- Superby, J. F., Vandamme, J., & Meskens, N. (2006). Determination of Factors Influencing the Achievement of the First-Year University Students using Data Mining Methods. In *Proceedings of the Workshop on Educational Data Mining at ITS'06* (pp. 37-44).
- Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In Search for the Most Informative Data for Feedback Generation: Learning Analytics in a Data-Rich Context. *Computers in Human Behavior*, 47, 157-167.
- Thai Nghe, N., Janecek, P., & Haddaway, P. (2007, Oktober). A Comparative Analysis of Techniques for Predicting Academic Performance. In *Proceedings of the 37th Conference on ASEE/IEEE Frontiers in Education* (pp. T2G-7-T2G-12).
- Therneau, T. M., & Atkinson, E. J. (1997). An Introduction to Recursive Partitioning Using the RPART Routines. Rochester, MN: Mayo Foundation.
- Tinto, V. (1987). *Leaving college: Rethinking the causes and cures of student attrition* (2e ed.). Chicago, IL: The University of Chicago Press.
- Van Rooij, E., Brouwer, J., Fokken-Buinsma, M., Jansen, E., Donche, V., & Noyens, D. (2017). A Systematic Review of Factors related to First-Year Students' success in Dutch and Flemish Higher Education. *Pedagogische Studiën*, 94(5) 360-405.
- Vereniging Hogescholen. (2018). *Factsheet Studiesucces en Uitval 2018*. Geraadpleegd van https://www.verenighogescholen.nl/system/knowledge_base/attachments/files/000/001/001/original/factsheet_studiesucces_uitval_2018_definitief.pdf
- Verhoeven, N. (2014). *Wat is onderzoek?* (5^e ed.). Amsterdam, Nederland: Boom Lemma.
- Wilmer, S. (2019). *Studierendement Optimaliseren door middel van het Verklaren en Bevorderen van Studiesucces* (Thesis).
- Wilson, K., Korn, J. H. (2007). Attention During Lectures: Beyond Ten Minutes. *Teaching of Psychology*, 34(2), 85-89.
- Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a Standard Process Model for Data Mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (pp. 29-39). Citeseer.
- York, T. T., Gibson, C., & Rankin, S. (2015). Defining and Measuring Academic Success. *Practical Assessment, Research & Evaluation*, 20(5).

Bijlage A: Feature selection

In paragraaf 2.2.1.1 werd gesproken over overfitting en feature selection. Dit zijn factoren waarmee rekening gehouden dient te worden wanneer met een model tracht te ontwikkelen middels Educational Data Mining. Een daarvan is dus overfitting. Dit houdt in dat het model te sterk naar de training-set vormt. Het resulteert erin dat het model de meest zeldzame en ongewone voorspellingen kan doen op de train-set, maar niet te generaliseren is. Bij overfitting onthoudt het model als het ware elke rij van de training data. Hierdoor wordt het een model die gespecialiseerd is op de records uit de train-set. Overfitting vindt plaats wanneer het model relatief aan de omvang van de data-set te complex is. Er worden te veel features (attributen) geselecteerd. Lijnrecht tegenover overfitting staat underfitting, waarbij het model te simpel is (te weinig features), waardoor het onvoldoende in staat is correct te voorspellen. Het daarom belangrijk zowel over- als underfitting te kunnen detecteren. Dit kan gedaan worden door de error te bestuderen.

Error is een combinatie van variantie, bias en complexiteit. Een optimaal model heeft zowel een lage bias als variantie. Het optimaliseren van een model op basis van de bias en variantie wordt de bias-variantie tradeoff genoemd (Geurts, 2002). Bias houdt in dat het algoritme consistent foutief voorspelt. Analooq aan een hoge bias staat een darter die consistent in het vak 20 raakt, terwijl op de bull's-eye gericht wordt. Variantie reflecteert de gevoeligheid voor het model voor kleine veranderingen in de training set. Voortbouwend op de analogie van de darter die op de bull's-eye richt, zullen de dartpijlen (bij hoge variantie) met een hoge spreiding in (of buiten) het dartbord landen. Figuur A1 laat een schematische weergave zien van bias en variantie. Om variantie te verlagen en een hogere bias te voorkomen, kan men bagging of boosting technieken inzetten. Het idee hierachter is dat diverse modellen gecreëerd worden waarbij (lichte) overfitting plaatsvindt, waarna op basis van gemiddelden een eindmodel geconstrueerd wordt.

Tot slot, om overfitting te voorkomen, wordt feature selection ingezet. Het doel van feature selection is het selecteren van een subset van (input) variabelen door het verwijderen van features die geen tot weinig toegevoegde (voorspellende) waarde bieden. Feature selection verhoogd het lerend vermogen, de accuratesse en complexiteit van het model (Ramaswami & Bhaskaran, 2009).



Figuur A1. Schematische weergave van de effecten van hoge en lage bias en variantie op accuratesse van een darter. De kruizen zijn het richtpunt van de darter en de grijze cirkels de landingsgebieden van de dartpijlen.

Bijlage B: Data preparation

Intakeassessment en doelgroep

Het databestand van het intake assessment van eerdere cohorten was al aanwezig. De cohorten in het bestand waren:

- Cohort 2014-2015;
- Cohort 2015-2016;
- Cohort 2016-2017;
- Cohort 2017-2018.

Cohort 2018-2019 ontbrak van deze dataset en werd toegevoegd. In tabel B1 zijn de cohorten en de omvang ervan weergegeven. Er is te zien dat 5.6% van de data bestaat uit TP-studenten die de deeltijd route (n = 114) of vwo-route (n = 11) volgen. Gezien de geringe omvang en bias te voorkomen zijn zij verwijderd uit de dataset, waardoor de omvang kromp naar n = 2095. Verder komt een aantal studenten meerdere malen voor. Dit zijn her-inschrijvers (n = 18). Zij werden verwijderd uit de dataset, omdat het koppelen van gegevens vanuit andere databases aan deze student gecompliceerd en het aantal her-inschrijvers laag was.

Tabel B1

Frequenties van de cohorten uit het databestand van het intakeassessment.

Cohort	Deeltijd	Voltijd	vwo-route	Totaal
2014-2015	32	530	--	562
2015-2016	52	450	11	513
2016-2017	30	437	--	467
2017-2018	--	376	--	376
2018-2019	--	302	--	302
<i>Cumulatief</i>	<i>114</i>	<i>2095</i>	<i>11</i>	<i>2220</i>
<i>Percentage</i>	<i>5.1</i>	<i>94.4</i>	<i>0.5</i>	<i>100%</i>

Vervolgens werden cohorten 2014-2015 (n = 562) en 2015-2016 (n = 513) verwijderd uit de dataset, aangezien de opleiding in 2016 van een nieuwe curriculum is voorzien. De toetsen, structuur en beroeps-specifieke competenties komen niet meer overeen. Verder wijkt ook de structuur van het intakeassessment dermate af dat veel informatie verloren gaat.

Uit nadere analyses bleek dat, op cohort 2018-2019 (n = 302), de dataset bestond uit starters en niet-starters (aangemeld, maar niet begonnen). Om deze studenten te identificeren

werd de lijst met studenten vergeleken met de klassenlijsten van het betreffende jaar. Daarin staan enkel studenten die als gestart worden beschouwd. Het komt echter wel eens voor dat studenten ingedeeld zijn in een klas, maar alsnog niet deelnemen aan de opleiding. Er kon niet exact achterhaald worden welke studenten dit betrof. Ook bleek dat het databestand enkel scores van het capaciteiten onderdeel, persoonlijkheid en demografische gegevens bevatte. Het onderdeel met algemene vragen ontbrak en werd toegevoegd aan de dataset voor elk cohort. Open vragen en niet-overeenkomende variabelen werden niet meegenomen. Na het samenvoegen van de gegevens bleek van 38 studenten data te ontbreken. Er kon niet achterhaald worden waar deze missende waarden vandaan kwamen. Omdat deze missende waardes beschrijvend waren voor de persoon (zoals participatie vóór inschrijven), konden waardes niet vervangen worden met de populatie gemiddeldes of modus (Kaiser, 2014). Daarnaast zijn veel decision trees niet in staat met missende waardes te werken (IBM, 2019). De keuze is gemaakt hen te verwijderen uit de dataset. De totale omvang van de doelgroep bestaat na dit data-cleaning proces uit 812 studenten (zie tabel B2).

Tabel B2

Frequenties per cohort van de doelgroep.

Cohort	Frequentie	Percentage	Complete records
2016-2017	266	32.8	100%
2017-2018	268	33.0	100%
2018-2019	278	34.2	100%
<i>Cumulatief</i>	<i>812</i>	<i>100.0</i>	<i>--</i>

Hierna werden de elementen uit het intakeassessment bekeken. De dataset bestond uit demografische gegevens, dimensie 0 (algemene vragen), dimensie 1 (capaciteiten) en dimensie 2 (persoonlijkheid). De demografische gegevens bestonden uit woonplaats en postcode, geboortjaar en geslacht. De volgende van deze variabelen werden omgevormd (zie tabel B3):

- Geboortjaar: leeftijd tijdens intake (in jaren).
- Woonplaats en postcode: omgevormd naar reistijd met de trein van de woonplaats naar het dichtstbijzijnde station van de hogeschool (in minuten). Berekend via <https://www.9292ov.nl/>.

Voor wat betreft dimensie 2 waren gemiddelde scores op de diverse persoonlijkheidsschalen berekend en opgeslagen als variabelen. De individuele antwoorden waren ook beschikbaar en bewaard in het kader van feature selection (zie tabel B4). Dit segment van de data bevatte

echter missing values. Aangezien deze missing values onder de categorie *Not Missing at Random* vallen, kunnen deze niet genegeerd worden (Kaiser, 2014). Zoals in hoofdstuk 4 vermeld zijn zij verwijderd uit de dataset.

Tabel B3

Demografische gegevens vóór en na transformeren.

Variabele	Beschrijving	Voorbeeld
<i>Voor</i>		
studnr	Student nummer (ID)	11111
inflowDate	Datum van inschrijving	01-Sep-16
voornaam	Voornaam van de student	Jan
achternaam	Achternaam van de student	Janssen
geslacht	Geslacht	{200, man} {201, vrouw}
geboortejaar	Jaar van geboorte	1993
postcode	Postcode huidig woonadres	1234 AB
woonplaats	Woonplaats huidig woonadres	DEVENTER
email	Emailadres van de student	email@adres.com
cohort	Cohort	[1-9]
PrimaryLast	Mest recente intakegegevens	{1, ja} {0, nee}
<i>Na</i>		
studnr	Student nummer (ID)	11111
geslacht	Geslacht	{1, man} {2, vrouw}
cohort	Cohort	[7-9]
leeftijd_intake	Leeftijd in jaren tijdens het assessment	19
reistijd_in_minuten	Reistijd met het openbaar vervoer vanaf de woonplaats naar Station Deventer	55

Opmerking. Speciale symbolen: [1-9], neemt één van de waardes aan van 1 tot en met 9; {,}, neemt een van de waardes aan links van de komma (bijvoorbeeld 200), rechts staat de betekenis.

Tabel B4

Intakestructuur op basis van de variabelen na het samenvoegen van de cohorten

#	Vraag	Codering	Waardes
<i>Dimensie 0: algemene vragen</i>			
1	Beeld van de opleiding	vr15021	Cijfer [1-10]
2	Beeld van beroep en werkveld	vr15022	Cijfer [1-10]
3	Aansluiting vooropleiding	vr15023	Ja/Nee/Twijfel
4	Tevreden met studiekeuze	vr15024	Cijfer [1-10]
5	[hoe] de studielast	vr15027	Likert [1-5]
6	[hoe] het niveau van de opleiding	vr15028	Likert [1-5]
7	[hoe] het zelfstandig studeren	vr15029	Likert [1-5]
8	[hoe] het wennen aan het hbo	vr15030	Likert [1-5]
9	[hoe] de combinatie studie en op kamers	vr15031	Likert [1-5]
10	Kans op propedeuse in 1 jaar	vr15033	Likert [1-5]
11	Eerder hbo- of wo-opleiding gevolgd	vr15034	Ja/Nee
12	Tevens aangemeld bij andere opleiding	vr15036	Ja/Nee
13	Aansluiting eindopdracht vorige opleiding	vr15037	Ja/Nee
14	[act] open dag	vr15014_01	Ja/Nee
15	[act] meeloop dag	vr15014_02	Ja/Nee
16	[act] try-out	vr15014_03	Ja/Nee
17	[act] voorlichting	vr15014_04	Ja/Nee
18	[act] geen	vr15014_05	Ja/Nee
19	[act] anders	vr15014_06	Ja/Nee
20	[sax] dichtbij	vr15016_01	Ja/Nee
21	[sax] vrienden gaan ook	vr15016_02	Ja/Nee
22	[sax] Saxion heeft een reputatie	vr15016_03	Ja/Nee
23	[sax] beoordelingen in keuzegids	vr15016_04	Ja/Nee
24	[sax] hogeschool past beter dan andere	vr15016_05	Ja/Nee
25	[sax] opleiding is exclusief bij Saxion	vr15016_06	Ja/Nee
26	[sax] studiekeuzetest	vr15016_07	Ja/Nee
27	[sax] advies vanuit studiekeuzebegeleiding	vr15016_08	Ja/Nee
28	[sax] anders	vr15016_09	Ja/Nee
29	[alt] dezelfde opleiding elders	vr15020_01	Ja/Nee
30	[alt] opleiding(en) in dezelfde sector	vr15020_02	Ja/Nee
31	[alt] opleiding(en) in andere sector	vr15020_03	Ja/Nee
32	[alt] opleidingen op ander niveau	vr15020_04	Ja/Nee
33	[alt] geen	vr15020_05	Ja/Nee
34	[bep] functiebeperking	vr15025_01	Ja/Nee
35	[bep] ziekte	vr15025_02	Ja/Nee
36	[bep] (top)sport	vr15025_03	Ja/Nee
37	[bep] werk	vr15025_04	Ja/Nee
38	[bep] zorgtaken	vr15025_05	Ja/Nee
39	[bep] anders	vr15025_06	Ja/Nee
40	[bep] vertel ik liever niet	vr15025_07	Ja/Nee
41	[bep] niet van toepassing	vr15025_08	Ja/Nee
<i>Dimensie 1: capaciteiten totaalscores</i>			
42	Taal: Woordbeoordeling & spelling	gem_d1_wb	Score [0-10]
43	Redeneren: Rekenvaardigheid	gem_d1_rv	Score [0-10]
44	Redeneren: Logische plaatjes reeksen	gem_d1_lp	Score [0-10]
45	Redeneren: Analogieën	gem_d1_lr	Score [0-10]
<i>Dimensie 2: persoonlijkheid totaalscores</i>			
46	Ambitie	gem_d2_amb	Score [1-5]
47	Bedachtzaamheid	gem_d2_be	Score [1-5]

48	Doelmatigheid	gem_d2_do	Score [1-5]
49	Studiemotivatie	gem_d2_mot	Score [1-5]
50	Ordelijkheid	gem_d2_or	Score [1-5]
51	Zelfdiscipline	gem_d2_zd	Score [1-5]
52	Zelfvertrouwen	gem_d2_zv	Score [1-5]
<i>Dimensie 2: persoonlijkheid items</i>			
53	Ik ga hard werken om mijn studiepunten te behalen.	d2_am_vr10353	Likert [1-5]
54	Ik heb weinig motivatie om hogerop te komen.	d2_am_vr10425	Likert [1-5]
55	Ik wil mijn mogelijkheden optimaal benutten.	d2_am_vr10426	Likert [1-5]
56	Ik hoef niet zo nodig uit te blinken.	d2_am_vr10427	Likert [1-5]
57	Ik omschrijf mezelf als een workaholic.	d2_am_vr10428	Likert [1-5]
58	Om mijn doelen te behalen, maak ik plannen en houd ik mij daaraan.	d2_am_vr10429	Likert [1-5]
59	Ik heb een duidelijk beeld van wat ik met mijn studie wil bereiken.	d2_am_vr10430	Likert [1-5]
60	Ik heb in mijn leven de dingen niet altijd even goed aangepakt.	d2_be_vr10431	Likert [1-5]
61	Voordat ik een beslissing neem, weeg ik alle opties goed af.	d2_be_vr10432	Likert [1-5]
62	Ik doe soms dingen spontaan zonder erover na te denken.	d2_be_vr10433	Likert [1-5]
63	Voordat ik aan de slag ga bedenk ik altijd wat de consequenties zijn.	d2_be_vr10434	Likert [1-5]
64	Ik handel vaak impulsief.	d2_be_vr10435	Likert [1-5]
65	Ik handel zelden in een opwelling.	d2_be_vr10436	Likert [1-5]
66	Anderen vinden mij een welbewust en bedachtzaam persoon.	d2_do_vr10437	Likert [1-5]
67	Mijn schoolzaken heb ik op orde.	d2_do_vr10438	Likert [1-5]
68	Ik zorg dat ik mijzelf goed informeer en kom tot weloverwogen besluiten.	d2_do_vr10439	Likert [1-5]
69	Ik kom regelmatig in situaties terecht waarop ik mij niet heb voorbereid.	d2_do_vr10440	Likert [1-5]
70	Ik ben tevreden over mijn goed doordachte standpunten.	d2_do_vr10441	Likert [1-5]
71	Ik weet mijn kennis toe te passen.	d2_do_vr10442	Likert [1-5]
72	Ik werk doelgericht en efficiënt.	d2_do_vr10443	Likert [1-5]
73	Ik verspil veel tijd, voordat ik aan de slag ga met een schoolopdracht.	d2_mot_vr10458	Likert [1-5]
74	Mijn schoolopdrachten weet ik op een productieve wijze af te ronden.	d2_mot_vr10459	Likert [1-5]
74	Ik maak af, waaraan ik begonnen ben.	d2_mot_vr10460	Likert [1-5]
75	[...] een opdracht lastig wordt, ga ik mij met andere zaken bezig houden.	d2_mot_vr10461	Likert [1-5]
76	Mijn aandacht dwaalt snel weg van de schooltaken.	d2_mot_vr10462	Likert [1-5]
77	Wat mijn school betreft, ben ik een echte doorzetter.	d2_mot_vr10463	Likert [1-5]
78	Ik geef doelen vaak op, voordat ik ze bereikt heb.	d2_mot_vr10464	Likert [1-5]
79	Ik maak thuis vaak schoon.	d2_mot_vr10465	Likert [1-5]
80	[...] maak een studieplanning, om te voorkomen [...] laatste moment.	d2_mot_vr10466	Likert [1-5]
81	Ik kan goed structuur aanbrengen in de manier waarop ik werk of leer.	d2_mot_vr10467	Likert [1-5]
82	Het werk dat ik inlever ziet er niet zo netjes uit.	d2_mot_vr10468	Likert [1-5]
83	Ik maak vaak slordigheidsfouten.	d2_mot_vr10469	Likert [1-5]

84	De mensen uit mijn omgeving vinden mij ordelijk.	d2_or_vr10451	Likert [1-5]
85	Ik ben nooit te laat voor mijn afspraken.	d2_or_vr10452	Likert [1-5]
86	Ik doe meer dan er van mij gevraagd wordt.	d2_or_vr10453	Likert [1-5]
87	Ik eis veel van mijzelf.	d2_or_vr10454	Likert [1-5]
88	Ik behaal hele goede resultaten.	d2_or_vr10455	Likert [1-5]
89	In mijn bezigheden lever ik minimale inspanning.	d2_or_vr10456	Likert [1-5]
90	Ik wil de beste van de klas zijn.	d2_or_vr10457	Likert [1-5]
91	Ik doe altijd mijn best.	d2_zd_vr10444	Likert [1-5]
92	Ik ben al blij met een voldoende.	d2_zd_vr10445	Likert [1-5]
93	Andere mensen vinden mij gedreven.	d2_zd_vr10446	Likert [1-5]
94	Het voelt voor mij goed om er alles uit te halen wat er in zit.	d2_zd_vr10447	Likert [1-5]
95	Ik ben zeer ambitieus.	d2_zd_vr10448	Likert [1-5]
96	Ik ben een toegewijd persoon.	d2_zd_vr10449	Likert [1-5]
97	[...] alleen voldoening als ik een beter resultaat behaal dan verwacht.	d2_zd_vr10450	Likert [1-5]
98	Over het algemeen ben ik tevreden met mijzelf.	d2_zv_vr10470	Likert [1-5]
99	Soms denk ik dat ik niets waard ben.	d2_zv_vr10471	Likert [1-5]
100	Ik vind dat ik een aantal goede kwaliteiten heb.	d2_zv_vr10472	Likert [1-5]
101	Ik doe dingen net zo goed als de meeste andere mensen.	d2_zv_vr10473	Likert [1-5]
102	Ik denk dat ik niet veel heb om echt trots op te zijn.	d2_zv_vr10474	Likert [1-5]
103	Ik voel me weleens nutteloos of niet belangrijk.	d2_zv_vr10475	Likert [1-5]
104	Ik vind dat ik als persoon de moeite waard ben.	d2_zv_vr10476	Likert [1-5]
105	Ik zou willen dat ik wat meer respect voor mezelf kon opbrengen.	d2_zv_vr10477	Likert [1-5]
106	Ik denk weleens dat ik een mislukkeling ben.	d2_zv_vr10478	Likert [1-5]
107	Ik denk positief over mezelf.	d2_zv_vr10479	Likert [1-5]

Opmerking. Voor numerieke antwoordmogelijkheden geldt dat een hoger cijfer positiever/beter/gunstiger is. Antwoorden met het type likertschaal bevatte een zesde optie: niet van toepassing, weet ik niet of geen mening (missing). Afkortingen: [hoe], hoe denk jij dat het straks gaat met?; [act], aan welke activiteiten heb jij meegenomen aangeboden door Saxion?; [sax], waarom koos jij voor Saxion?; [alt], welke alternatieven heb je onderzocht?; [bep], speelt er iets wat effect kan hebben op je studie en het studeren?.

Uitval

Om te bepalen de student tot de classificatie uitval behoorde, werden enrollment gegevens van voorgaande jaren gebruikt. Elk jaar wordt aan de start een bestand gemaakt met alle studenten die aan de opleiding starten. Het tweede leerjaar wordt hetzelfde gedaan. Door te bekijken welke studenten wel en niet in het databestand voor het tweede studiejaar zitten kon bepaald worden welke studenten uitgevallen waren. Dit werd gedaan in Excel. Hierna werd deze variabele toegevoegd aan het databestand met het intakeassessment.

Bison

De cijfers en gemaakte toetsen van de studenten waren in losse databestanden te vinden, waarin elke toetskans te vinden is (eerste, tweede, derde, etc.). In het bestand waren de volgende gegevens te vinden:

- LoginID: student nummer
- Naam: voor- en achternaam
- Klas: klas waar de student in zit
- Toetscode: intern referentienummer naar de toets
- Toetsomschrijving: naam van de toets
- Ecs: aantal studiepunten die het behalen van de toets oplevert
- Toetsdatum: datum waarop de toets gemaakt werd
- Toetsresultaat: het behaalde cijfer
- Toetsbehaald: of de toets behaald is
- Opleidingsvariant: gevolgde opleiding van de student

De data moest eerst opgeschoond worden om enkel de eerste toets-kans van de studenten te bevatten. Toetsen die in kwartiel twee, drie en vier werden gegeven zijn niet meegenomen, aangezien het doel is dropout zo vroeg mogelijk, bij voorkeur vanaf het eerste semester, te kunnen voorspellen, zodat interventies ingezet kunnen worden. De resultaten van de toetsen gemaakt in kwartiel twee zijn pas het tweede semester bekend, waardoor deze resultaten onbruikbaar zijn.

Toetsen die in het eerste kwartiel gegeven werden weken af voor de cohorten. Cohort 2016-2017 kregen in het eerste kwartiel de toets Presteren en Leren aangeboden, terwijl deze voor de andere cohorten vervangen werd door Inleiding Arbeids- en Organisatie psychologie. Na het bestuderen van de curricula van de betreffende toetsen, bleken deze toetsen dezelfde competenties te meten en bevorderen. Ook kwamen de toetsvormen overeen en konden daarom samengevoegd worden toe dezelfde variabele. In tabel B5 worden de toetsen, toetsvormen en de cohorten waarin zij gegeven werden getoond. Tabel B6 laat zien hoe deze verwerkt zijn in de Bison dataset.

Tabel B5

Toetsen in het eerste kwartiel van de drie cohorten

Toets	Toets-vorm	16-17	17-18	18-19
Presteren en Leren	digitaal	x		
Diagnostisch Onderzoek	digitaal	x	x	x
Professionele Gespreksvoering en Training	werkstuk	x		
Inleiding Psychologie (1)	digitaal	x	x	x
Inleiding Psychologie (2)	werkstuk	x	x	x
Practice Based Learning 1 (1)	schriftelijk	x	x	x
Practice Based Learning 1 (2)	werkstuk	x	x	x
Practice Based Learning 1 (3)	assessment	x	x	x
Inleiding arbeids- en organisatiepsychologie	digitaal		x	x

Tabel B6

Variabelen in de Bison dataset (met uitzondering van student nummer)

Variabele	Type	Beschrijving
t_behaald_do_dig	ja/nee	Toets behaald Diagnostisch Onderzoek
t_behaald_inlaop_dig	ja/nee	Toets behaald Inleiding Arbeids- en Organisatiepsychologie of Presteren en Leren
t_behaald_inlpsy_dig	ja/nee	Toets behaald Inleiding Psychologie (1)
t_behaald_inlpsy_wk	ja/nee	Toets behaald Inleiding Psychologie (2)
t_behaald_pbl1_as	ja/nee	Toets behaald Practice Based Learning (1)
t_behaald_pbl1_sc	ja/nee	Toets behaald Practice Based Learning (2)
t_behaald_pbl1_wk	ja/nee	Toets behaald Practice Based Learning (3)
t_resultaat_do_dig	cijfer [1-10]	Toets-cijfer Diagnostisch onderzoek
t_resultaat_inlaop_dig	cijfer [1-10]	Toets-cijfer Inleiding Arbeids- en Organisatiepsychologie of Presteren en Leren
t_resultaat_inlpsy_dig	cijfer [1-10]	Toets-cijfer Inleiding Psychologie (1)
t_resultaat_inlpsy_wk	cijfer [1-10]	Toets-cijfer Inleiding Psychologie (2)
t_resultaat_pbl1_as	cijfer [1-10]	Toets-cijfer Practice Based Learning (1)
t_resultaat_pbl1_sc	cijfer [1-10]	Toets-cijfer Practice Based Learning (2)
t_resultaat_pbl1_wk	cijfer [1-10]	Toets-cijfer Practice Based Learning (3)
t_aanwezig_do_dig	ja/nee	Toets gemaakt Diagnostisch onderzoek

t_aanwezig_inlaop_dig	ja/nee	Toets gemaakt Inleiding Arbeids- en Organisatiepsychologie of Presteren en Leren
t_aanwezig_inlpsy_dig	ja/nee	Toets gemaakt Inleiding Psychologie (1)
t_aanwezig_inlpsy_wk	ja/nee	Toets gemaakt Inleiding Psychologie (2)
t_aanwezig_pbl1_as	ja/nee	Toets gemaakt Practice Based Learning (1)
t_aanwezig_pbl1_sc	ja/nee	Toets gemaakt Practice Based Learning (2)
t_aanwezig_pbl1_wk	ja/nee	Toets gemaakt Practice Based Learning (3)
t_aantal_behaalde_toetsen	cijfer [1-10]	Aantal behaalde toetsen
t_toets_mean	cijfer [1-10]	Gemiddeld cijfer van de gemaakte toetsen
t_mean_behaalde_toetsen	cijfer [1-10]	Gemiddeld cijfer van de behaalde toetsen

Blackboard

Allereerst werd data gedownload van de server van Eesyssoft. In bijlage D wordt een overzicht getoond van het eerste proces in de data-cleaning. Na deze cleaning bleef per cursus een databestand over met de volgens informatie (zie figuur B1):

- user_id: het student nummer
- tool_group: deze leek in eerste instantie wellicht interessant, maar er kon niet achterhaald worden wat de informatie betekende
- activity_date: tijd en datum waarop van de klik
- enrollment_date: datum van inschrijving in de cursus
- minor_id: inter referentie nummer naar monitor_name
- monitor_name: naam van de minor_id
- monitor_description: omschrijving van de monitor_name
- course_name: naam van de Blackboard cursus
- activity_week: enrollment_data omgezet naar kalenderweek

	A	B	C	D	E	F	G	H	I
	user_id	tool_group	activity_date	enrollment_date	minor_id	monitor_name	monitor_description	course_name	activity_week
1		1 Course Management	10/5/2016 13:13	8/31/2016 10:46	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	40
2		1 Course Management	10/30/2016 12:30	8/31/2016 10:46	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	43
3		2 Course Management	9/30/2016 20:31	8/23/2016 14:36	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	39
4		3 Course Management	9/28/2016 11:13	8/31/2016 10:46	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	39
5		4 Course Management	9/30/2016 13:51	8/28/2016 8:39	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	39
6		4 Course Management	10/29/2016 16:49	8/28/2016 8:39	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	43
7		5 Course Management	10/2/2016 11:50	8/29/2016 14:04	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	39
8		5 Course Management	10/6/2016 9:05	8/29/2016 14:04	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	40
9		5 Course Management	10/12/2016 14:26	8/29/2016 14:04	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	41
10		5 Course Management	11/8/2016 8:58	8/23/2016 17:25	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	45
11		6 Course Management	10/5/2016 10:19	8/30/2016 9:19	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	40
12		6 Course Management	10/9/2016 15:56	8/30/2016 9:19	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	40
13		6 Course Management	10/10/2016 15:38	8/30/2016 9:19	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	41
14		6 Course Management	10/11/2016 15:25	8/30/2016 9:19	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	41
15		6 Course Management	10/15/2016 21:45	8/30/2016 9:19	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	41

Figuur B1. Voorbeeld van een sectie uit een van de databestanden van Blackboard.

Tot op dit punt was de data al gefilterd om enkel activiteiten van de studenten in het eerste kwartiel te bevatten. Hiervoor werd eerst uitgezocht welke kalenderweek gelijk stond aan de lesweken van de betreffende cohorten. Vervolgens werd voor de studenten (per cohort) berekend hoe veel kliks zijn maakten per vak. Dit resulteerde in drie databestanden met elk vier variabelen (totaal aantal kliks per cursus, 2 tot en met 5). Deze bestanden werden samengevoegd tot een bestand.

Hierna werd voor elke cursus voor elke lesweek berekend hoe actief de studenten waren. Het eerste kwartiel duurde 11 weken, waarvan de achtste week de herfstvakantie betrof. Dit proces leverde $11 \text{ (weken)} \times 4 \text{ (cursussen)} = 44$ variabelen op. Hierna kon per lesweek bekeken worden hoe vaak elke student in de cursussen actief waren door het totaal van elke lesweek te berekenen. Dit leverde 11 (weken) variabelen op.

De hiervoor genoemde variabelen waren op interval niveau. Aangezien binaire variabelen in het kader van decision trees goede voorspellers lijken (Romero et al., 2010), werden een aantal variabelen van nominaal niveau gecreëerd. Voor elke lesweek werd per student bekeken of zij al dan niet actief waren in de respectievelijke lesweek (per vak). Dit leverde 44 variabelen op. Ook werd per vak achterhaald of zij actief waren tijdens alle weken, lesweken, toets-weken en les- & toets-weken, waardoor 17 variabelen toegevoegd konden worden aan de data set. Zie tabel B7 voor een overzicht van alle variabelen verkregen uit Blackboard Learn.

Tabel B7

Variabelen in de Blackboard Learn dataset (met uitzondering van studentnummers)

Variabele	Beschrijving
kliks_totaal_do	Totaal aantal kliks DO
kliks_totaal_inlaop	Totaal aantal kliks INLAOP
kliks_totaal_inlpsy	Totaal aantal kliks INLPSY
kliks_totaal_pbl	Totaal aantal kliks PBL
do_wk_1	Totaal aantal kliks lesweek 1 DO
do_wk_2	Totaal aantal kliks lesweek 2 DO
do_wk_3	Totaal aantal kliks lesweek 3 DO
do_wk_4	Totaal aantal kliks lesweek 4 DO
do_wk_5	Totaal aantal kliks lesweek 5 DO
do_wk_6	Totaal aantal kliks lesweek 6 DO

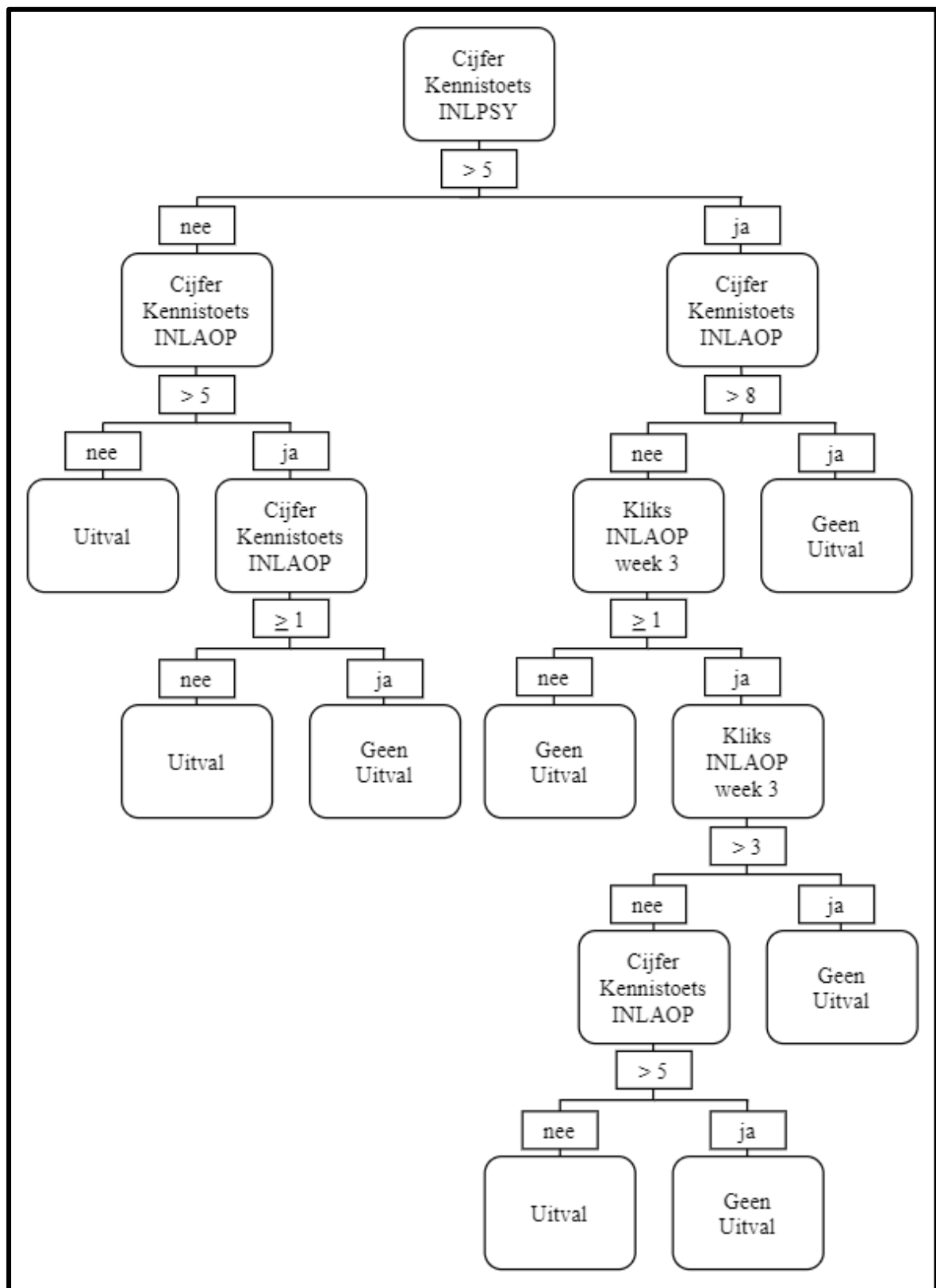
do_wk_7	Totaal aantal kliks lesweek 7 DO
do_wk_8	Totaal aantal kliks lesvakantie DO
do_wk_9	Totaal aantal kliks lesweek 8 DO
do_wk_10	Totaal aantal kliks lesweek 9 DO
do_wk_11	Totaal aantal kliks lesweek 10 DO
inlaop_wk_1	Totaal aantal kliks lesweek 1 INLAOP
inlaop_wk_2	Totaal aantal kliks lesweek 2 INLAOP
inlaop_wk_3	Totaal aantal kliks lesweek 3 INLAOP
inlaop_wk_4	Totaal aantal kliks lesweek 4 INLAOP
inlaop_wk_5	Totaal aantal kliks lesweek 5 INLAOP
inlaop_wk_6	Totaal aantal kliks lesweek 6 INLAOP
inlaop_wk_7	Totaal aantal kliks lesweek 7 INLAOP
inlaop_wk_8	Totaal aantal kliks lesvakantie INLAOP
inlaop_wk_9	Totaal aantal kliks lesweek 8 INLAOP
inlaop_wk_10	Totaal aantal kliks lesweek 9 INLAOP
inlaop_wk_11	Totaal aantal kliks lesweek 10 INLAOP
inlpsy_wk_1	Totaal aantal kliks lesweek 1 INLPSY
inlpsy_wk_2	Totaal aantal kliks lesweek 2 INLPSY
inlpsy_wk_3	Totaal aantal kliks lesweek 3 INLPSY
inlpsy_wk_4	Totaal aantal kliks lesweek 4 INLPSY
inlpsy_wk_5	Totaal aantal kliks lesweek 5 INLPSY
inlpsy_wk_6	Totaal aantal kliks lesweek 6 INLPSY
inlpsy_wk_7	Totaal aantal kliks lesweek 7 INLPSY
inlpsy_wk_8	Totaal aantal kliks lesvakantie INLPSY
inlpsy_wk_9	Totaal aantal kliks lesweek 8 INLPSY
inlpsy_wk_10	Totaal aantal kliks lesweek 9 INLPSY
inlpsy_wk_11	Totaal aantal kliks lesweek 10 INLPSY
pbl_wk_1	Totaal aantal kliks lesweek 1 PBL
pbl_wk_2	Totaal aantal kliks lesweek 2 PBL
pbl_wk_3	Totaal aantal kliks lesweek 3 PBL
pbl_wk_4	Totaal aantal kliks lesweek 4 PBL
pbl_wk_5	Totaal aantal kliks lesweek 5 PBL
pbl_wk_6	Totaal aantal kliks lesweek 6 PBL

pbl_wk_7	Totaal aantal kliks lesweek 7 PBL
pbl_wk_8	Totaal aantal kliks lesvakantie PBL
pbl_wk_9	Totaal aantal kliks lesweek 8 PBL
pbl_wk_10	Totaal aantal kliks lesweek 9 PBL
pbl_wk_11	Totaal aantal kliks lesweek 10 PBL
kliks_totaal_wk_1	Som van totaal aantal kliks lesweek 1
kliks_totaal_wk_2	Som van totaal aantal kliks lesweek 2
kliks_totaal_wk_3	Som van totaal aantal kliks lesweek 3
kliks_totaal_wk_4	Som van totaal aantal kliks lesweek 4
kliks_totaal_wk_5	Som van totaal aantal kliks lesweek 5
kliks_totaal_wk_6	Som van totaal aantal kliks lesweek 6
kliks_totaal_wk_7	Som van totaal aantal kliks lesweek 7
kliks_totaal_wk_8	Som van totaal aantal kliks lesweek 8
kliks_totaal_wk_9	Som van totaal aantal kliks herfstvakantie
kliks_totaal_wk_10	Som van totaal aantal kliks lesweek 9
kliks_totaal_wk_11	Som van totaal aantal kliks lesweek 10
do_wk_1_actief	Actief tijdens lesweek 1 DO
do_wk_2_actief	Actief tijdens lesweek 2 DO
do_wk_3_actief	Actief tijdens lesweek 3 DO
do_wk_4_actief	Actief tijdens lesweek 4 DO
do_wk_5_actief	Actief tijdens lesweek 5 DO
do_wk_6_actief	Actief tijdens lesweek 6 DO
do_wk_7_actief	Actief tijdens lesweek 7 DO
do_wk_8_actief	Actief tijdens lesweek 8 DO
do_wk_9_actief	Actief tijdens herfstvakantie DO
do_wk_10_actief	Actief tijdens lesweek 9 DO
do_wk_11_actief	Actief tijdens lesweek 10 DO
inlaop_wk_1_actief	Actief tijdens lesweek 1 INLAOP
inlaop_wk_2_actief	Actief tijdens lesweek 2 INLAOP
inlaop_wk_3_actief	Actief tijdens lesweek 3 INLAOP
inlaop_wk_4_actief	Actief tijdens lesweek 4 INLAOP
inlaop_wk_5_actief	Actief tijdens lesweek 5 INLAOP
inlaop_wk_6_actief	Actief tijdens lesweek 6 INLAOP

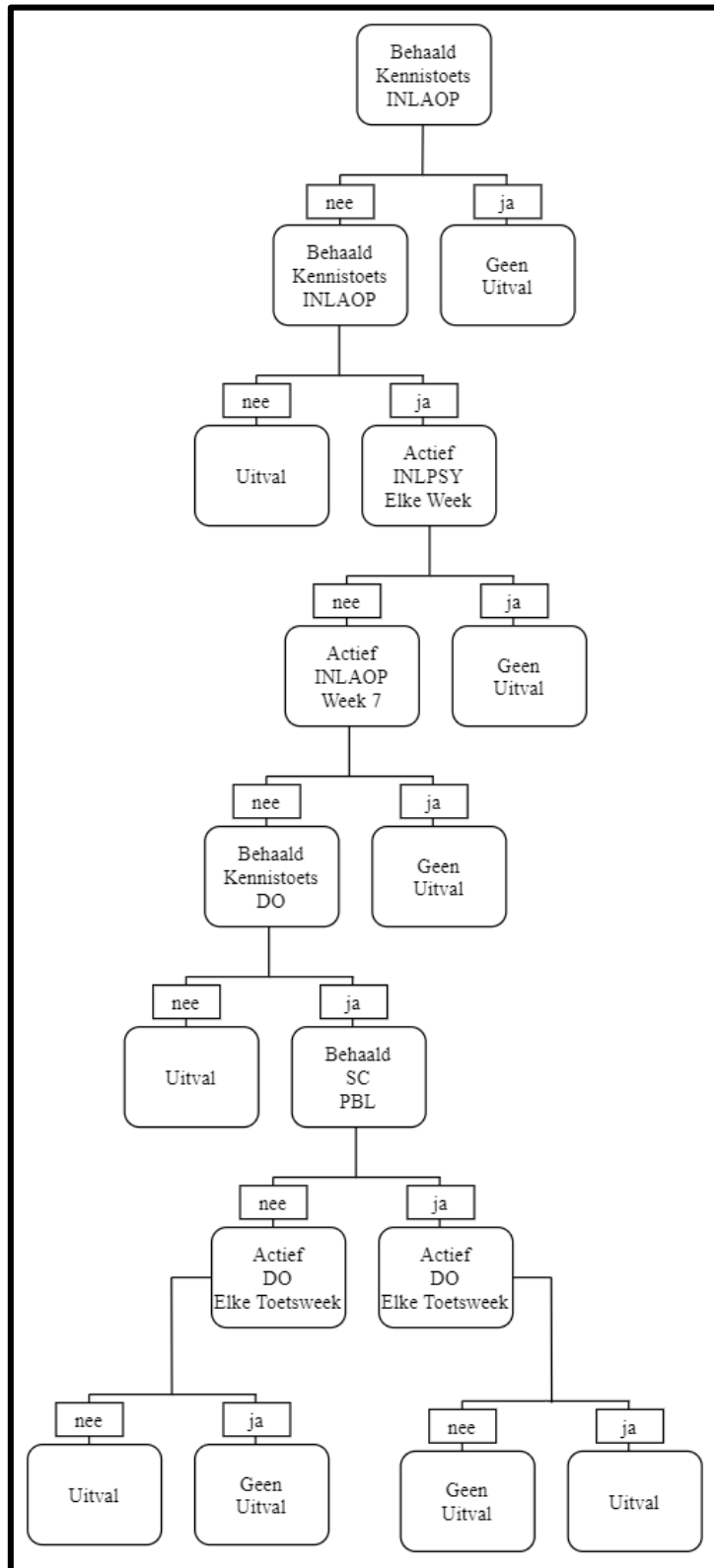
inlaop_wk_7_actief	Actief tijdens lesweek 7 INLAOP
inlaop_wk_8_actief	Actief tijdens lesweek 8 INLAOP
inlaop_wk_9_actief	Actief tijdens herfstvakantie INLAOP
inlaop_wk_10_actief	Actief tijdens lesweek 9 INLAOP
inlaop_wk_11_actief	Actief tijdens lesweek 10 INLAOP
inlpsy_wk_1_actief	Actief tijdens lesweek 1 INLPSY
inlpsy_wk_2_actief	Actief tijdens lesweek 2 INLPSY
inlpsy_wk_3_actief	Actief tijdens lesweek 3 INLPSY
inlpsy_wk_4_actief	Actief tijdens lesweek 4 INLPSY
inlpsy_wk_5_actief	Actief tijdens lesweek 5 INLPSY
inlpsy_wk_6_actief	Actief tijdens lesweek 6 INLPSY
inlpsy_wk_7_actief	Actief tijdens lesweek 7 INLPSY
inlpsy_wk_8_actief	Actief tijdens lesweek 8 INLPSY
inlpsy_wk_9_actief	Actief tijdens herfstvakantie INLPSY
inlpsy_wk_10_actief	Actief tijdens lesweek 9 INLPSY
inlpsy_wk_11_actief	Actief tijdens lesweek 10 INLPSY
pbl_wk_1_actief	Actief tijdens lesweek 1 PBL
pbl_wk_2_actief	Actief tijdens lesweek 2 PBL
pbl_wk_3_actief	Actief tijdens lesweek 3 PBL
pbl_wk_4_actief	Actief tijdens lesweek 4 PBL
pbl_wk_5_actief	Actief tijdens lesweek 5 PBL
pbl_wk_6_actief	Actief tijdens lesweek 6 PBL
pbl_wk_7_actief	Actief tijdens lesweek 7 PBL
pbl_wk_8_actief	Actief tijdens lesweek 8 PBL
pbl_wk_9_actief	Actief tijdens herfstvakantie PBL
pbl_wk_10_actief	Actief tijdens lesweek 9 PBL
pbl_wk_11_actief	Actief tijdens lesweek 10 PBL
do_elke_week_actief	Elke week actief bij DO
inlaop_elke_week_actief	Elke week actief bij INLAOP
inlpsy_elke_week_actief	Elke week actief bij INLPSY
pbl_elke_week_actief	Elke week actief bij PBL
all_bb_elke_week_actief	Elke week actief bij alle cursussen
do_alle_lesweken_actief	Alle lesweken actief bij DO

inlaop_alle_lesweken_actief	Alle lesweken actief bij INLAOP
inlpsy_alle_lesweken_actief	Alle lesweken actief bij INLPSY
pbl_alle_lesweken_actief	Alle lesweken actief bij PBL
do_alle_toetsweken_actief	Alle toets-weken actief bij DO
inlaop_alle_toetsweken_actief	Alle toets-weken actief bij INLAOP
inlpsy_alle_toetsweken_actief	Alle toets-weken actief bij INLPSY
pbl_alle_toetsweken_actief	Alle toets-weken actief bij PBL
do_alle_les_en_toetsweken_actief	Alle les- en toets-weken actief bij DO
inlaop_alle_les_en_toetsweken_actief	Alle les- en toets-weken actief bij INLAOP
inlpsy_alle_les_en_toetsweken_actief	Alle les- en toets-weken actief bij INLPSY
pbl_alle_les_en_toetsweken_actief	Alle les- en toets-weken actief bij PBL

Bijlage C: Resultaten modellering



Figuur C1. Decision Tree van op basis van de CART.



Figuur C2. Decision Tree op basis van de C5.0.

Bijlage D: Dataverwerking met Python

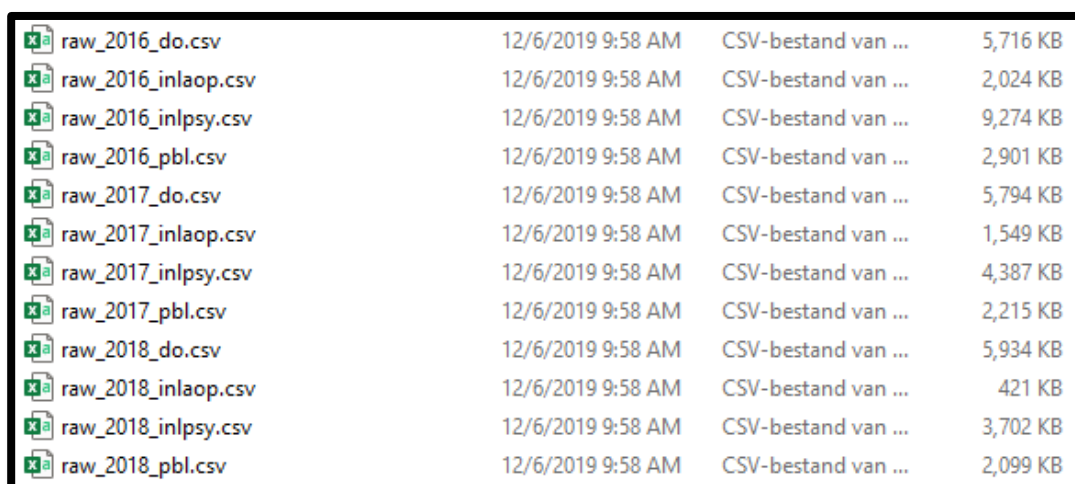
Hieronder wordt een voorbeeld getoond van de dataverwerking van Blackboard met inzet van Python. Eerst moest van elke cursus de ruwe data gedownload worden via Eesysoft. Het betreft in totaal 12 cursussen, zoals te zien in Tabel D1.

Tabel D1.

Course ID's per vak en cohort van Blackboard Learn.

Vak	Course ID		
	Cohort 16-17	Cohort 17-18	Cohort 18-19
INLAOP	AMA_TP_PRESLEREN_D_1617_3	AMA_TP_INLAO_D_1718_1	AMA_TP_INLAO_D_1819_1
DO	AMA_TP_DO_D_1617_34	AMA_TP_DO_D_1718_12	AMA_TP_DO_D_1819_12
INLPSY	AMA_TP_ALGPSY_D_1617_1	AMA_TP_ALGPSY_D_1718_1	AMA_TP_ALGPSY_D_1819_1
PBL	AMA_TP_PBLTAKEN_D_1617_23	AMA_TP_PBLTAKEN_D_1718_1	AMA_TP_PBLTAKEN_D_1819_1

Elke cursus moest handmatig opgezocht worden door de cursusnaam te selecteren in Eesysoft en vervolgens de correcte filters toe te passen (zoals tijdsperiode). Het resultaat is bijvoorbeeld een databestand met 39 kolommen en 21536 rijen (“raw_2016_inlpsy.csv”; zie figuur D1). Elke rij stelt één klik voor. Figuur D1 toont een overzicht van alle bestanden.



raw_2016_do.csv	12/6/2019 9:58 AM	CSV-bestand van ...	5,716 KB
raw_2016_inlaop.csv	12/6/2019 9:58 AM	CSV-bestand van ...	2,024 KB
raw_2016_inlpsy.csv	12/6/2019 9:58 AM	CSV-bestand van ...	9,274 KB
raw_2016_pbl.csv	12/6/2019 9:58 AM	CSV-bestand van ...	2,901 KB
raw_2017_do.csv	12/6/2019 9:58 AM	CSV-bestand van ...	5,794 KB
raw_2017_inlaop.csv	12/6/2019 9:58 AM	CSV-bestand van ...	1,549 KB
raw_2017_inlpsy.csv	12/6/2019 9:58 AM	CSV-bestand van ...	4,387 KB
raw_2017_pbl.csv	12/6/2019 9:58 AM	CSV-bestand van ...	2,215 KB
raw_2018_do.csv	12/6/2019 9:58 AM	CSV-bestand van ...	5,934 KB
raw_2018_inlaop.csv	12/6/2019 9:58 AM	CSV-bestand van ...	421 KB
raw_2018_inlpsy.csv	12/6/2019 9:58 AM	CSV-bestand van ...	3,702 KB
raw_2018_pbl.csv	12/6/2019 9:58 AM	CSV-bestand van ...	2,099 KB

Figuur D1. Overzicht van de ruwe data bestanden met kliks in Blackboard Learn.

De volgende stap is het in kaart brengen van de informatie in de ruwe data bestanden. Dit werd met een combinatie van Python en Excel gedaan. De databestanden werden eerst gecontroleerd of het downloaden succesvol was (*face validity*). Vervolgens werd gefocust op een bestand (“raw_2016_inlpsy.csv”) om een eerste indruk te krijgen van de aanwezige informatie. Na het bekijken van de kolommen en inhoud (in Excel) werd, zoals hierboven vermeld, geconstateerd dat “raw_2016_inlpsy.csv” 39 kolommen bevatte. Om te controleren

of alle ruwe databestanden soortgelijke informatie bevatten, werd een simpel script in Python (jupyter notebook) geschreven. Zou je dit handmatig in Excel willen doen, ontkom je er haast niet aan elk bestand een voor een te openen, de kolomnamen te kopiëren en plakken, om vervolgens na te gaan of ze overeen komen. In figuur D2 is echter te zien hoe met een aantal regels hetzelfde bereikt kan worden met Python. In de figuur is onderaan naast “Out[4]” het woord “True” te zien, wat betekent dat zowel de kolommen als hun respectievelijke positie in de bestanden gelijk is.

```

Importeer benodigde modules

In [1]: import pandas as pd
import numpy as np
import os

Maak een lijst met de naam van de databestanden en controleer of deze aanwezig zijn in de directory

In [2]: # Python Lists ter referentie
courses = ['do', 'inlaop', 'inlpsy', 'pbl']
years = ['2016', '2017', '2018']

# Lege Lijst om bestandsnamen in op te slaan
files = []

# Voeg bestandsnamen toe aan de Lege Lijst
for year in years:
    for crs in courses:
        file = 'raw_{0}_{1}.csv'.format(year, crs)
        files.append(file)
        print(file, '\tin directory:\t', os.path.exists(file))

raw_2016_do.csv      in directory:  True
raw_2016_inlaop.csv  in directory:  True
raw_2016_inlpsy.csv  in directory:  True
raw_2016_pbl.csv     in directory:  True
raw_2017_do.csv      in directory:  True
raw_2017_inlaop.csv  in directory:  True
raw_2017_inlpsy.csv  in directory:  True
raw_2017_pbl.csv     in directory:  True
raw_2018_do.csv      in directory:  True
raw_2018_inlaop.csv  in directory:  True
raw_2018_inlpsy.csv  in directory:  True
raw_2018_pbl.csv     in directory:  True

Open de bestanden

In [3]: frames = []
for file in files:
    df_temp = pd.read_csv(file, sep=';')
    frames.append(df_temp)

Controleer of kolommen overeen komen

In [4]: np.all([np.all(frames[i].columns == frames[i+1].columns) for i in range(len(frames)-1)])
Out[4]: True

```

Figuur D2. Voorbeeld van een sectie uit een Python script om te controleren of alle kolommen in de ruwe data van Blackboard met elkaar overeen komen.

Aangezien de kolommen met elkaar overeenkomen (naam en positie), is het voldoende om slechts één bestand ter referentie open te houden in Excel. Er werd hierin opgemerkt dat de kolomnamen zowel hoofdletters, kleine letters als spaties bevatten. Om te voorkomen dat dit later problemen opleverde, werden de hoofdletters omgevormd tot kleine letters, terwijl spaties vervangen werden door lage streepjes (“_”). Dit werd wederom via Python gedaan.

Daarnaast bleek dat de databestanden tevens kliks van docenten bevatte wie tijdelijke de rol “student” werd toegewezen in het systeem. Deze moesten ook verwijderd worden. Om dit handmatig te doen kost enorm veel tijd, waardoor gezocht werd naar een simpele methode om hen te identificeren in de data. Het bleek vervolgens dat er een kolom was met emailadressen gekoppeld aan de identificatiecodes van studenten (student nummers) en docenten (*short names*). Elke student krijgt een emailadres waarin in de extensie het woord “*student*”. Bij docenten komt dit niet voor. Door dus enkel de records te selecteren en bewaren in de data die in het emailadres het woord student hebben staan, worden niet-studenten verwijderd uit de data set.

Bij het verder bestuderen van de kolommen, bleek veel informatie niet relevant te zijn. Er werd geconcludeerd dat enkel de volgende kolommen relevante informatie bevatten:

- user_id: het student nummer
- tool_group: deze leek in eerste instantie wellicht interessant, maar er kon niet achterhaald worden wat de informatie betekende
- activity_date: tijd en datum waarop van de klik
- renrollment_date: datum van inschrijving in de cursus
- minotor_id: inter referentie nummer naar monitor_name
- monitor_name: naam van de minotor_id
- monitor_description: omschrijving van de monitor_name
- course_name: naam van de Blackboard cursus

Om al deze handelingen (kolomnamen aanpassen en records en kolommen selecteren op basis van criteria) uit te kunnen voeren, is het beheersen van een programmeertaal geen vereiste. De mogelijkheid bestaat namelijk elk databestand in Excel te openen om vervolgens, bijvoorbeeld door te “zoeken en vervangen” of het toevoegen van filters aan de kolommen, dezelfde resultaten te behalen. Dit heeft echter het nadeel dat het veel tijd kan kosten, waarnaast men het risico loopt (type-)fouten te maken. Dit proces kan echter met slechts een aantal regels in Python bereikt worden, zoals in figuur D3 te zien is, wat de kans op fouten aanzienlijk verminderd.

```
Transformeer kolomnamen

In [5]: #vervang hoofdletters en spaties
for i in range(len(frames)):
    frames[i].columns = frames[i].columns.str.lower() # kleine Letters
    frames[i].columns = frames[i].columns.str.replace(' ', '_') # spaties

#check of kolommen nog overeenkomen
np.all([np.all(frames[i].columns == frames[i+1].columns) for i in range(len(frames)-1)])

Out[5]: True

Selecter enkel de studenten in de databestanden

In [6]: for i in range(len(frames)):
    frames[i] = frames[i][frames[i]['email'].str.contains('student')]

Selecteer alleen relevante kolommen

In [7]: for i in range(len(frames)):
    columns_to_keep = ['user_id', 'tool_group', 'activity_date', 'enrollment_date',
                      'monitor_id', 'monitor_name', 'monitor_description', 'course_name']
    frames[i] = frames[i][columns_to_keep]

print('Kolommen:')
[print(' ', i) for i in frames[0].columns]
np.all([np.all(frames[i].columns == frames[i+1].columns) for i in range(len(frames)-1)])

Kolommen:
user_id
tool_group
activity_date
enrollment_date
monitor_id
monitor_name
monitor_description
course_name

Out[7]: True
```

Figuur D3. Voorbeeld van een sectie uit een Python script om kolomnamen aan te passen en studenten en relevante kolommen te selecteren.

Het laatste wat moest gebeuren was de kliks die na het eind van kwartiel 1 plaatsvonden verwijderen uit de dataset, zodat enkel gegevens van het eerste kwartiel in het databestand staan. Elk leerjaar kan echter een afwijkende startdatum hebben, waarnaast het per klas verschilt op welke dag zij les krijgen. Er werd daarom een nieuwe kolom gecreëerd, waarin ten eerste de datum getransformeerd werd naar kalenderweeknummer. Vervolgens werd deze omgezet naar de weeknummers die de hogeschool hanteert. Een studiejaar kent twee semester, welke beide opgedeeld zijn in twee kwartielen van elk 10 lesweken. Voor het eerste kwartiel (eerste semester) werd per cohort de kalenderweek gelijkstaand aan de eerste lesweek opgezocht, waarna hetzelfde werd gedaan voor de laatste lesweek van het kwartiel. Op basis hiervan kunnen kalenderweeknummers later getransformeerd worden naar lesweken. Het selecteren van de gegevens uit kwartiel 1 was de laatste stap van de eerste fase van het opschonen van Blackboard data. De aangebrachte veranderingen in het script werden opgeslagen in nieuwe bestanden. Figuur D4 laat een overzicht van dit proces zien. Figuur D5 toont een sectie uit een van databestanden, waarbij studentnummers geanonimiseerd zijn.

```

Selecteer gegevens van kwartiel 1

1. Converteer kolommen met tijd en datum naar een format die Python (pandas library) herkent als datum

In [8]: #converteer kolommen naar datetime
for i in range(len(frames)):
    frames[i]['activity_date'] = pd.to_datetime(frames[i]['activity_date'], dayfirst=True)
    frames[i]['enrollment_date'] = pd.to_datetime(frames[i]['enrollment_date'], dayfirst=True)

# creëer nieuwe kolom met weeknummers
for i in range(len(frames)):
    frames[i]['activity_week'] = frames[i]['activity_date'].dt.week

2. Converteer kalenderweken naar kwartiel-weeknummers

In [9]: # kalenderweken per cohort die gelijk staat aan week 1 van kwartiel 1
start_weken = [35, 36, 36] # [cohort 2016, cohort 2017, cohort 2017]

# kalenderweken per cohort die gelijk staat aan week 10 van kwartiel 1
eind_weken = [45, 46, 46] # [cohort 2016, cohort 2017, cohort 2017]

# hulp variabele
counter = 0

# selecteer daterange
# years = ['2016', '2017', '2018'] <-- zie: In [2] (Figuur C2)
for i, year in enumerate(years):
    sw = start_weken[i]
    ew = eind_weken[i]
    for j in range(4):
        frames[counter] = frames[counter].loc[(frames[counter]['activity_week']>=sw) & (frames[counter]['activity_week']<=ew)]
        counter+=1

 Sla bestanden op

In [10]: # hulp variabele
count = 0

# courses = ['do', 'inlaop', 'inlpsy', 'pbl'] <-- zie: In [2] (Figuur C2)
# years = ['2016', '2017', '2018'] <-- zie: In [2] (Figuur C2)
for i, year in enumerate(years):
    for j, crs in enumerate(courses):
        filename = 'scr_02_{0}_{1}.csv'.format(year, crs)
        frames[count].to_csv(filename, index=False)
        print('Succes: saved frames[{0}] as {1}'.format(count, filename))
        count+=1

Succes: saved frames[0] as scr_02_2016_do.csv
Succes: saved frames[1] as scr_02_2016_inlaop.csv
Succes: saved frames[2] as scr_02_2016_inlpsy.csv
Succes: saved frames[3] as scr_02_2016_pbl.csv
Succes: saved frames[4] as scr_02_2017_do.csv
Succes: saved frames[5] as scr_02_2017_inlaop.csv
Succes: saved frames[6] as scr_02_2017_inlpsy.csv
Succes: saved frames[7] as scr_02_2017_pbl.csv
Succes: saved frames[8] as scr_02_2018_do.csv
Succes: saved frames[9] as scr_02_2018_inlaop.csv
Succes: saved frames[10] as scr_02_2018_inlpsy.csv
Succes: saved frames[11] as scr_02_2018_pbl.csv

```

Figuur D4. Voorbeeld van een sectie uit een Python script om gegevens van kwartiel 1 te selecteren en op te slaan.

	A	B	C	D	E	F	G	H	I
1	user_id	tool_group	activity_date	enrollment_date	monitor_id	monitor_name	monitor_description	course_name	activity_week
2	1	Course Management	10/5/2016 13:13	8/31/2016 10:46	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	40
3	1	Course Management	10/30/2016 12:30	8/31/2016 10:46	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	43
4	2	Course Management	9/30/2016 20:31	8/23/2016 14:36	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	39
5	3	Course Management	9/28/2016 11:13	8/31/2016 10:46	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	39
6	4	Course Management	9/30/2016 13:51	8/28/2016 8:39	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	39
7	4	Course Management	10/29/2016 16:49	8/28/2016 8:39	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	43
8	5	Course Management	10/2/2016 11:50	8/29/2016 14:04	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	39
9	5	Course Management	10/6/2016 9:05	8/29/2016 14:04	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	40
10	5	Course Management	10/12/2016 14:26	8/29/2016 14:04	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	41
11	5	Course Management	11/8/2016 8:58	8/23/2016 17:25	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	45
12	6	Course Management	10/5/2016 10:19	8/30/2016 9:19	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	40
13	6	Course Management	10/9/2016 15:56	8/30/2016 9:19	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	40
14	6	Course Management	10/10/2016 15:38	8/30/2016 9:19	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	41
15	6	Course Management	10/11/2016 15:25	8/30/2016 9:19	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	41
16	6	Course Management	10/15/2016 21:45	8/30/2016 9:19	129	Enter My Grades	The monitor measures that a student enters his/her My Grades	Algemene psychologie Deventer(2016-2017)	41

Figuur D5. Voorbeeld van een sectie uit een van de databestanden na data cleaning.

Bijlage E: Decision trees

C5.0

Met het algoritme van de C5.0 decision tree kan een decision tree (of een set met regels in de vorm van ALS-DAN) gegenereerd worden. Bij elke splitsing berekent het algoritme welke feature de meeste informatie oplevert over de responsvariabele (*information gain*). De C5.0 kan alleen ingezet worden om classificaties te bepalen van nominaal niveau (IBM, 2019).

CART

De Classification and Regression Tree (CART) genereert een decision tree waarmee voorspellingen gedaan kunnen worden over continue en nominale responsvariabelen. Dit CART splitst enkel op binair niveau, in tegenstelling tot bijvoorbeeld de CHAID decision tree (IBM, 2019). CART trees worden gegenereerd op basis van *recursive partitioning* (het herhaalde splitsen van de data in groepen), evenals de C5.0 decision trees, door de mate van *impurity* te minimaliseren (Therneau & Atkinson, 1997; Strobl, Malley & Tutz, 2009). Dit houdt in dat gezocht wordt naar een splitsing met de laagste kans op type I en II fouten.

CHAID en Exhaustive CHAID

Decision trees op basis van het CHAID algoritme maken gebruik van Chi-Kwadraat toetsen om splitsingen te generen. In tegenstelling tot de eerder genoemde decision tree algoritmes, kan met de CHAID meerdere splitsingen per niveau plaatsvinden. In andere woorden, *non-binary* decision trees kunnen genereerd worden. De Exhaustive CHAID is een aanpassing op de CHAID. Er wordt namelijk grondiger gezocht naar alle mogelijke splitsingen (IBM, 2019).

QUEST

Met het algoritme van de QUEST decision tree kunnen enkel binaire classificaties voorspelt worden. Het algoritme werd ontwikkeld om de rekentijd van CART te verminderen. Daarnaast biedt die algoritme een oplossing voor de neiging die veel algoritmes hebben om de voorkeur te geven aan variabelen waarmee meerdere splitsingen gemakkelijk plaatsvinden (IBM, 2019). Predictor variabelen kunnen op interval en ratio niveau zijn, maar de responsvariabelen moet nominaal (meer specifiek, binair) zijn (IBM, 2019).

Meer informatie over de algoritmes kan gevonden worden in de documentatie van SPSS Modeler 18.2.1, te vinden op de website van IBM (<ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.2.1/en/>).