2012•2013
FACULTEIT BEDRIJFSECONOMISCHE WETENSCHAPPEN
master in de toegepaste economische wetenschappen:
handelsingenieur in de beleidsinformatica

Masterproef

Analyse van de slaagresultaten m.b.v. data-miningtechnieken

Promotor :
Prof. dr. Benoit DEPAIRE

Copromotor :
Prof. dr. Christiane MASUI

Jeroen Lemmens

Masterproef voorgedragen tot het bekomen van de graad van master in de toegepaste economische wetenschappen: handelsingenieur in de beleidsinformatica

universiteit hasselt
KNOWLEDGE IN ACTION

Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

universiteit hasselt
KNOWLEDGE IN ACTION

2012•2013
FACULTEIT BEDRIJFSECONOMISCHE WETENSCHAPPEN
master in de toegepaste economische wetenschappen: handelsingenieur in de beleidsinformatica

Masterproef

Analyse van de slaagresultaten m.b.v. data-miningtechnieken

Promotor :
Prof. dr. Benoit DEPAIRE

Copromotor :
Prof. dr. Christiane MASUI

Jeroen Lemmens

Masterproef voorgedragen tot het bekomen van de graad van master in de toegepaste economische wetenschappen: handelsingenieur in de beleidsinformatica

universiteit hasselt
KNOWLEDGE IN ACTION

Voorwoord

Met deze masterproef rond ik mijn opleiding handelsingenieur in de beleidsinformatica aan de Uhasselt af. Het onderwerp handelt over de actuele probleemstelling van de lage slaagkans van studenten aan de Vlaamse universiteiten en hoe deze tegen te gaan. Ik heb geprobeerd hiervoor opzoek te gaan naar geschikte dataminingmodellen om de slaagkans van een student te voorspellen. Op deze manier krijgt de student een beter inzicht in het feit of hij deze richting succesvol zal kunnen volbrengen. Deze datamining aanpak past perfect in het kader van mijn opleiding handelsingenieur in de beleidsinformatica en vormt dus een mooi sluitstuk deze opleiding.

Het schrijven van deze masterproef was niet mogelijk geweest zonder de hulp van een heel aantal personen. Langs deze weg wil ik hen dan ook oprecht bedankten.

Allereerst wens ik mijn promotor dr. Benoit Depaire te bedanken voor de vele goede ideeën, zeer nuttige feedback en broodnodige hulp. Zonder hem was deze masterproef nooit kunnen worden wat hij nu is. Zeer veel dank ook voor mijn co-promotor dr. Chris Masui, die mij bijstond voor taalkundig advies en vragen in verband met haar vakdomein. Ook een welgemeende dankuwel aan de voltallige onderwijsstaf, voor mij al de nodig kennis bij te brengen die nodig was voor de totstandkoming van deze masterproef.

Ook een speciaal woordje van dank aan mijn ouders, mijn zus en mijn broer voor de ondersteuning gedurende mijn hele opleiding. Natuurlijk bedank ik ook al mijn mede studenten voor de leuke tijd die ik met hen beleefd heb aan de Uhasselt.

Tot slot nog een laatste dankuwel aan iedereen die ook nog geholpen heeft tijdens het tot stand komen en nalezen van deze masterproef. Met een bijzondere woord van dank voor Michaël Vreys, Ineke Deneyer, Pieter Beerten en Lendert Vanderwaeren voor hun deskundige hulp en aan Eva Buttiëns voor het finaal nalezen van deze masterproef.

Jeroen Lemmens

Samenvatting

Een recent onderzoek wees uit dat in de periode 2001-2007 nauwelijks de helft van de 70 000 startende studenten aan een Vlaamse universiteit volledig slaagde voor zijn eerste bachelorjaar. Dit probleem brengt een aanzienlijke sociale kost met zicht mee, ook voor de maatschappij. Ter oplossing van dit probleem worden er verschillende mogelijkheden naar voor geschoven: invoering van een toegangsproef voor alle richtingen, screening op basis van de resultaten in het middelbaar, hervorming van het secundair onderwijs zodat het beter aansluit op de hogere studies, invoering van outputfinanciering in het eerste jaar voor de instellingen, invoeren van een niet-bindende oriëntatieproef. Al deze voorstellen hebben wel hun voor- en nadelen.

In deze masterproef gaan we onderzoeken of we met behulp van datamining en het knowledge discovery process een voorspelling kunnen maken van de slaagkans van een student. We proberen aan de hand van kenmerken en resultaten van een student een model te bouwen dat zijn slaagkans kan voorspellen. Dit model zou dan gebruikt kunnen worden om het probleem van de lage slaagkans op te lossen door studenten te informeren over hun potentiële slaagkans. Op deze manier kunnen studenten beter inschatten of de gekozen richting binnen hun mogelijkheden ligt.

Aan de hand van een literatuurstudie gaan we op zoek naar de factoren die een relatie hebben met de slaagkans van een student. Deze zijn op te delen in vier categorieën: persoonskenmerken (leeftijd, geslacht, etniciteit, taal, cognitieve intelligentie, persoonlijkheid, zelfvertrouwen, sociale responslabiliteit, flexibiliteit), persoonlijk verleden (eerdere studies, eerder blijven zitten, werkervaring, andere factoren), familiale achtergrond (diploma ouders, ondersteuning van het gezin, gescheiden ouders, socio-economische status, financiële hulp) en factoren tijdens de studie (werken tijdens studies, studeergedrag, tijd investeren, faculteit- student contact, peer interactie, extra-curriculaire activiteiten, studententevredenheid over de instelling, leven op de campus, eerste jaar seminaries, gebruik academische support services, factoren buiten de student om).

Helaas zijn niet al deze factoren opgenomen in de studentendatabase BEW van de UHasselt. Maar toch bevat deze database een heel deel nuttige factoren waarmee we aan de slag gaan om te proberen een model te bouwen om de slaagkans van een student te voorspellen. Voor het bouwen van de modellen gebruiken we verschillende datamining algoritmes:

- OneR classifier
- Naive Bayes classifier
- Beslissingsboom
- k-Nearest neighbor algoritme
- Neurale netwerken

We voeren verschillende analyses uit, telkens met alle vijf deze algoritmes en op drie momenten in de tijd:

- Voor aanvang van de studies, deze modellen bevatten volgende variabelen

- Geslacht
- Nationaliteit
- Gemeente
- Land
- Op kot of niet
- Secundaire school
- Gemeente secundaire school
- Studierichting secundaire school
- Aantal uren wiskunde secundaire school
- Aantal uren natuurkunde secundaire school
- Aantal uren scheikunde secundaire school
- Aantal uren Frans secundaire school
- Aantal uren Engels secundaire school
- Aantal uren Duits secundaire school
- Beursstudent of niet
- Studievertraging opgelopen
- Generatiestudent of niet
- Totaal resultaat leestest
- Totaal resultaat op denktest
- Deelresultaat denktest verbaal
- Deelresultaat denktest numeriek
- Deelresultaat denktest grafisch

- Na 1ste trimester (na kerstvakantie) worden volgende variabelen nog bijgevoegd

- Resultaat financieel boekhouden
- Resultaat geo-economie en economische geschiedenis
- Resultaat macro-economie
- Resultaat sociologie en demografie
- Resultaat wiskunde
- Studietijdmeting (totaal)

- Op het einde van het jaar worden volgende variabelen nog toegevoegd

- Resultaten op het einde van het jaar voor alle vakken
- Aantal herexamens
- Aantal onvoldoendes op het einde van het jaar
- Aantal studiepunten behaald
- Aantal studiepunten behaald na deliberatie
- Resultaat 1ste jaar (%)
- Eindbeoordeling 1ste jaar
- Eerste zit of tweede zit

Na uitvoering van de analyses, zien we dat we met de gegevens voor aanvang van de studies, met een minimum van 70% correct geclassificeerde cases, kunnen gaan voorspellen of een student :

- Slaagt in zijn eerste jaar
- Zijn bachelordiploma zal behalen
- Welk resultaat hij zal behalen op statistiek

We kunnen niet met een voldoende groot percentage voorspellen hoeveel herexamens een student zal behalen of in welke categorie van eindpercentage hij zal vallen. Ook de resultaten voor wiskunde en boekhouden zijn niet te voorspellen met een zekerheid groter dan 70 % .

Verder zien we dat hoe meer inputgegevens (aanvang studies/ na resultaten 1ste trimester / einde van het jaar) we toevoegen, hoe beter de resultaten zijn. Dit is een logisch gegeven, maar het toont aan hoe belangrijk een evaluatie van een student na het 1ste trimester is. Na de toevoeging van de resultaten van het eerste trimester zien we in elk van de analyses een grote toename in het percentage correct geclassificeerde cases.

Als we dan gaan kijken naar de analyses die plaatsvinden na de resultaten van het 1ste trimester, zien we dat we voor elke doelvariabele een model hebben dat beter presteert als 70%. We kunnen dus na het 1ste trimester voorspellen, met een zekerheid groter als 70%, of een student :

- Slaagt in zijn eerste jaar
- Hoeveel herexamens hij zal behalen
- In welke categorie van totaal percentage hij zal vallen
- Welke resultaat hij zal gaan behalen op wiskunde / statistiek / boekhouden
- Zijn bachelordiploma zal behalen en of hij dit zal doen volgens het modeltraject of niet.

Met toevoeging van de resultaten en gegevens op het einde van het jaar kunnen we zelfs nog betere voorspellingen gaan doen in verband met behalen van het bachelordiploma. Hier halen we zelfs een percentage correct geclassificeerde cases van 89%.

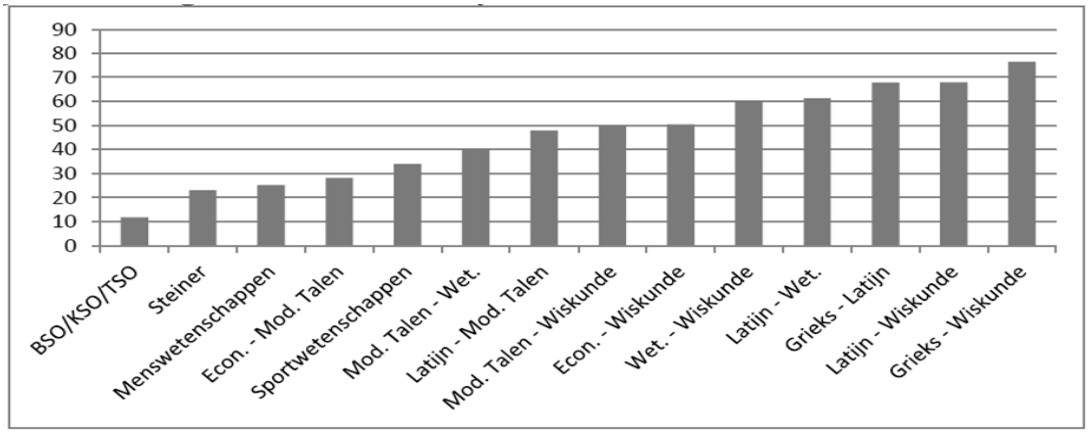
Inhoudsopgave

Hoofdstuk 1: Inleiding

1.1. Praktijkprobleem

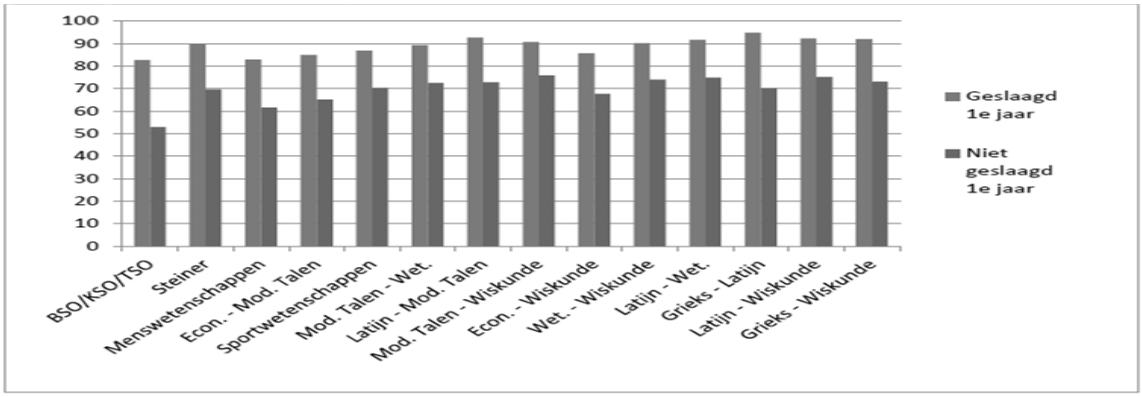
Een recent onderzoek van de Katholieke Universiteit Leuven (Declercq, K., Verboven, F. "Slaagkansen aan Vlaamse universiteiten: tijd om het beleid bij te sturen? (2010)) wees uit dat de slaagkansen van de startende studenten in het eerste bachelorjaar aan de universiteit zeer laag zijn. In de periode van 2001 tot 2007 slaagde nauwelijks de helft van de 70 000 startende studenten volledig voor zijn eerste bachelorjaar.

De studie toonde tevens aan dat meisjes een 10% hogere slaagkans hebben dan jongens, dat leerlingen uit een vrije school 20% meer slaagkans hebben dan leerlingen uit het gemeenschaps- of het officieel onderwijs en dat studenten die minstens één jaar moeten dubbelen in het middelbaar een tot 30% lagere slaagkans hebben. Ook het diploma dat behaald werd in het middelbaar heeft een effect op de slaagkans (zie Figuur 1). Voornoemde elementen zijn algemene trends die voor elke universitaire richting ongeveer gelijkwaardig zijn.



Figuur 1: Slaagkansen van eerstejaarsstudenten aan de Vlaamse universiteiten (per richting van het secundair onderwijs) - Declercq, Verboven (2010)

Na het bekijken van onderstaande figuur kunnen we verder uit het onderzoek besluiten dat de selectie van studenten voornamelijk in het eerste jaar gebeurt. Daarnaast zien we ook dat de problemen van niet-geslaagde studenten in het eerste jaar blijven doorwerken in het volgende jaar. Een student die in zijn eerste bachelorjaar dus niet volledig slaagt, zal in zijn tweede bachelorjaar een lagere slaagkans hebben. Als we naar Figuur 2 kijken, kunnen we bijvoorbeeld afleiden dat een student, die de richting Grieks-Wiskunde volgde in het secundair, in zijn 2de bachelorjaar een slaagkans heeft van meer dan 90% indien hij geslaagd was in zijn eerste jaar. Was hij echter niet geslaagd in zijn eerste bachelorjaar dan heeft hij slechts een slaagkans van 72%.



Figuur 2: Slaagkansen van tweedejaarsstudenten geslaagd versus niet geslaagd in het eerste jaar - Declercq, Verboven (2010)

De onderzoekers stelden ook een econometrisch model op. Dit legt de verschillen in slaagkansen tussen de verschillende studentengroepen bloot. We bespreken hier enkele conclusies uit dat econometrisch model. Zo blijkt dat studenten uit het gemeenschaps- of officiële onderwijs een significant lagere slaagkans hebben dan studenten uit het vrij onderwijs: respectievelijk -14,5% en -10,7%. Daarnaast hebben ook studenten die één of twee jaren overdeden in het secundair onderwijs een significant lagere slaagkans: respectievelijk -22,6% en -31,2%. Verder bleek uit het model dat vrouwen een hogere kans hebben op slagen, namelijk +9,8%. Buitenlandse studenten hebben dan weer een lagere slaagkans: -5,4%. Studenten die 100 km verwijderd zijn van de universitaire instelling hebben een beperkte maar significant hogere slaagkans van 3,1%. Dit kunnen we toeschrijven aan een selectie-effect: sterkere studenten willen een langere afstand afleggen om naar de universiteit te gaan. Opmerkelijk genoeg vinden we sterke verschillen voor de universitaire opleidingen. Ten opzichte van het vergelijkingspunt Economische en Toegepaste Economische Wetenschappen hebben de exacte opleidingen meestal een sterk lagere slaagkans en de humane opleidingen een sterk hogere slaagkans. We geven hier enkele voorbeelden: Diergeneeskunde: -14,8%, Toegepaste biomedische Wetenschappen: -6,2%, Politieke en Sociale Wetenschappen: +3,5%, Geschiedenis: +9,4%, Tandheelkunde en Geneeskunde: +18,1 en 29,5% (te verklaren door eerdere selectie via toelatingsproef).

Natuurlijk is er wel een grote voorzichtigheid geboden bij het interpreteren van deze resultaten. Dit aangezien de slaagkans het resultaat is van een samenloop van meerdere kenmerken, zowel intrinsiek als extrinsiek. Dat heeft als gevolg dat we de slaagkans dus niet enkel door een optelling van voorgaande factoren kunnen berekenen. Zo spelen onder meer de sociale omgeving, de motivatie en willekeurige gebeurtenissen ook een rol in de slaagkans van een student.

Het huidige beleid van screening op het einde van het eerste jaar in de plaats van op het begin van het eerste jaar heeft als gevolg dat iedere leerling met een diploma secundaire onderwijs zonder beperking toegang heeft tot bijna alle universitaire opleidingen. Dit feit brengt dus een lage slaagkans met zich mee zoals gestaafd door bovenstaande gegevens.

Het grote probleem met deze lage slaagkans is nu dat die een hoge sociale kost heeft voor de maatschappij. Rechtstreekse studiekosten, betaald door de overheid, variëren afhankelijk van de studierichting tussen €5.000 en €10.000 per jaar (Declercq, Verboven (2010)) en daarbovenop bedraagt de kost voor de student zelf ook nog een vergelijkbaar bedrag. Daarnaast is er dan nog de onrechtstreekse kost, volgend uit het uitstel tot toetreding tot de arbeidsmarkt, bestaande uit een niet-inkomen voor de student en niet-inning van belastingen voor de overheid.

Als we dit bekijken voor alle studenten kunnen we besluiten dat we wel degelijk mogen spreken van een aanzienlijke kost. Het is daarom zeker de moeite om zich actief op te stellen ten aanzien van dit probleem van lage slaagkansen van eerstejaarsstudenten. De financiering voor het hoger onderwijs gebeurt sinds 2008 weliswaar op basis van het aantal studenten dat effectief slaagt (outputfinanciering). Maar dit geldt niet voor het eerste jaar dat nog steeds volgens inputfinanciering werkt. Deze uitzondering is het resultaat van een compromis en had de bedoeling de universiteiten toch aan te zetten studenten uit alle sociale groepen te rekruteren. Toch blijft het feit dat de studieachtergrond van vele studenten ruim onvoldoende is om te slagen aan de universiteit, en daarnaast deze toegang zonder beperkingen tot universitaire opleidingen, geen ideaal scenario.

Als mogelijke oplossing voor dit probleem kan er, zoals in andere landen reeds gebeurt, gescreend worden op basis van de resultaten in het secundair onderwijs of kan men werken met een toegangsproef. Met deze laatste optie hebben we al enkele ervaringen in Vlaanderen. Zo zagen de onderzoekers dat na de afschaffing van de ingangsproef voor Burgerlijk Ingenieur, de slaagkans voor eerstejaarsstudenten daalde van 62% tot 46%. Bij Geneeskunde is de toegangsproef er nog steeds. Deze richting heeft dan ook een slaagpercentage van 87%. Hieruit kunnen we dus opmaken dat deze ingangsproef tot op zekere hoogte in staat is de bekwaamheid van studenten te screenen.

Andere mogelijke oplossingen die Declercq en Verboven in hun onderzoek aanhalen, zijn het inputfinancieringssysteem van het eerste jaar omzetten naar het outputsysteem in combinatie met begeleidende maatregelen of een hervorming van het secundair onderwijs, waardoor dit dan meer aansluit met de potentiële hogere studies.

Een maatregel die momenteel de aandacht trekt, is die van de invoering van een niet-bindende oriëntatieproef. Deze is louter informatief maar zou wel leiden tot een beter overwogen

studiekeuze. De meeste rectoren pleiten voor de invoering van zulke proef (De Standaard, 26 augustus, 2011). Ook de Minister van Onderwijs blijkt te vinden voor deze maatregel (Deredactie.be, 27 september, 2010).

Maar als we dan een kritische blik werpen op alle mogelijke oplossingen blijken er aan elk alternatief ook nadelen verbonden. Zo lijkt de toegangsproef, zoals toegepast bij Geneeskunde, aan de hand van de cijfers een goede oplossing. Toch is een van de nadelen hierbij dat er ook studenten zullen worden tegengehouden, die uiteindelijk wel zouden slagen voor deze richting. Het zou kunnen dat studenten, die juist niet slagen op deze ingangsproef, te weinig voorbereidingen hebben gedaan voor deze proef, niet genoeg gemotiveerd zijn voor de proef of zelfs een gewone tegenslag hebben op de dag van de proef. Deze studenten zouden dan onterecht worden tegengehouden om zicht in te schrijven tot de betreffende richting.

Op de screening op basis van resultaten in het secundair is dan weer als tegenargument in te brengen dat sommige studenten niet genoeg gemotiveerd zijn om al van in het secundair bezig te zijn met hun verdere studies. Dit kan ertoe leiden dat ze dus niet de benodigde resultaten zullen behalen om te kunnen starten met hun hogere studies, hoewel zij hierop normaal wel zouden kunnen slagen.

De bemerking op het alternatief om ook het systeem van inputfinanciering in het eerste jaar om te zetten naar outputfinanciering is vooral dat men zo de onderwijsinstelling aanzet naar het rekruteren van enkel de sterke studenten. Zo zouden een heel deel studenten uit andere sociale groepen uit de boot vallen en al op voorhand worden opgegeven, terwijl ook zij een eerlijke kans verdienen.

Een hervorming van het secundair onderwijs, zodat dit meer aansluit op de hogere studies, heeft dan weer als nadeel dat de studenten nog vroeger dan nu al gedwongen zouden worden een vorm van studiekeuze te maken, terwijl de meeste studenten zelfs op hun achttiende nog geen idee hebben welke richting ze uit willen.

Het laatste alternatief onder de vorm van een niet-bindende oriëntatieproef heeft dan weer als nadeel dat het toch vrijblijvend blijf. Het kan de studenten wel ondersteunen maar niet verplichten tot een keuze. Het uiteindelijke effect van zo'n proef kan daarom niet met zekerheid worden bepaald.

1.2. Wat willen we gaan onderzoeken?

De bedoeling van deze masterproef is het opstellen van een voorspellend model door middel van knowledge discovery. Dit model zal op basis van verscheidende kenmerken van een student een voorspelling maken van zijn slaagkans. Zo wordt de student geïnformeerd over zijn of haar mogelijkheden binnen een richting en eventueel ook gewaarschuwd indien een richting te zwaar lijkt. Dit model biedt dan tevens een mogelijk alternatief om de dalende slaagkans tegen te gaan door studenten te waarschuwen aan het begin van hun opleiding of na de resultaten van het eerste trimester, indien zij zich bevinden in een risicozone voor het niet slagen in hun opleiding. Op deze manier zouden we dan de grote maatschappelijke kost, die het niet slagen van vele studenten in hun eerste jaar met zich meebrengt, kunnen verminderen.

Dit model kan een hele waaier van variabelen omvatten, gaande van de kenmerken van de student tot de resultaten van taaltesten of studietijdmeting. Zelfs het resultaat van een oriëntatieproef kan hierin worden opgenomen. Welke variabelen we gebruiken, zal afhangen van de variabelen die te onzer beschikking zijn. Het doel van het model is daarnaast het op zoek gaan naar de onderliggende factoren die een gerelateerd zijn aan de slaagkans van een student. Dit moeten we echter differentiëren van factoren die een oorzakelijk verband hebben met de slaagkans. Ze zijn namelijk niet hetzelfde. Dit onderscheid balanceert op een dunne koord, maar met behulp van volgend voorbeeld trachten we het te verduidelijken. Stel dat uit een onderzoek blijkt dat studenten, die een zomercursus wiskunde volgen, een hoger slagingspercentage hebben voor het uiteindelijke examen van wiskunde. Dan kunnen we stellen dat er een relatie is tussen het volgen van de zomercursus wiskunde en het slagen op het examen wiskunde. We kunnen echter niet stellen dat er een oorzakelijk verband is tussen beide elementen. Want de relatie tussen het volgen van de zomercursus en het slagen op het examen zou te verklaren zijn door een onderliggende factor. Bijvoorbeeld omdat studenten die meer gemotiveerd zijn, deze zomercursus volgen en ditzelfde type studenten daardoor ook meer gemotiveerd is voor het studeren van het examen en

daardoor een beter resultaat zal behalen. Binnen deze masterproef gaan we dus op zoek naar relaties tussen factoren en niet naar oorzakelijke verbanden.

Naar de factoren die een relatie hebben met de slaagkans van een student zijn al eerder verschillende statistische onderzoeken gevoerd, zoals ook het voornoemde econometrische model. Doch lijkt het ons zeer interessant om dit onderzoek vanuit de invalshoek van knowledge discovery te voeren. Deze meer recente onderzoeksmethode, die ondermeer gebruik maakt van datamining, verschilt op een heel aantal punten van de statische aanpak. Zo start men bij statistiek eerder vanuit een vastgelegd model gebaseerd op onderliggende theorieën en probeert men via de data de theorie te bevestigen. Bij datamining vertrekt men echter eerder vanuit de data zelf en probeert men de geheimen te ontdekken die deze data bevatten (Kuonen *(2004))*. Daarnaast is het gebruik van datamining meer toepasselijk indien men een zeer grote database met veel variabelen wil onderzoeken (Hand (1999)). De bedoeling binnen deze masterproef is om het model op te stellen aan de hand van zoveel mogelijk gegevens over studenten en dit impliceert dus een grote database. Ook willen we het onderzoek zonder oogkleppen voeren, en dus niet starten met een vooropgesteld model maar de data zelf laten spreken. Om deze voorgaande redenen ligt de keuze voor het gebruik van het knowledge discovery process dus voor de hand.

1.3. Onderzoeksvragen

1.3.1. Centrale onderzoeksvraag

"Kunnen we, met behulp van het knowledge discovery process, voorspellingen gaan doen over de slaagkans van een student aan de UHasselt?"

1.3.2. Deelvragen

1) "Welke factoren hebben een relatie met de slaagkans van een student in het algemeen?"

2) "Welke datamining algoritmen zijn het meest geschikt voor het bouwen van een model om de slaagkans te voorspellen?"

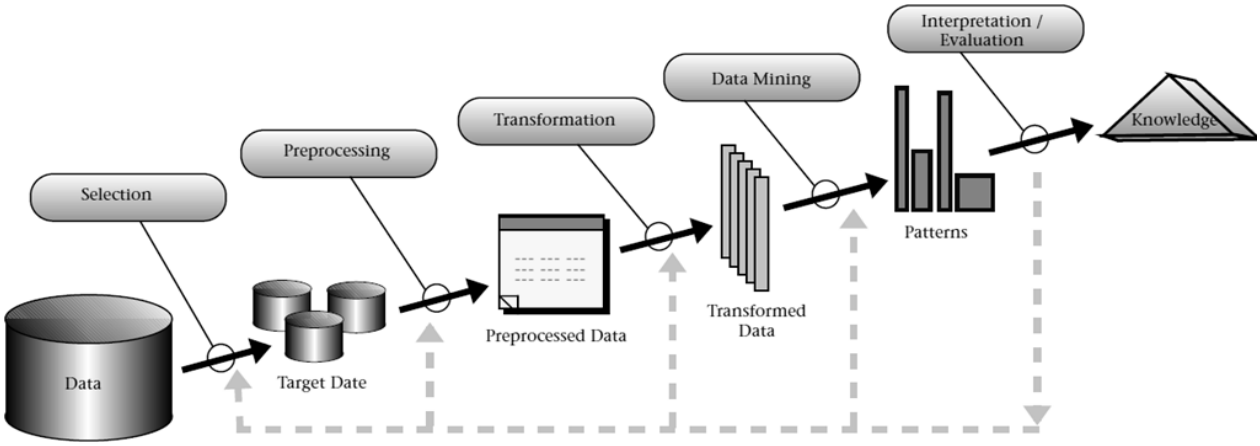
1.4. Opzet van het onderzoek

Stap één die we in deze masterproef moeten gaan ondernemen is natuurlijk de literatuurstudie. Deze stap kunnen we in twee delen opsplitsen. Namelijk een eerste deel over de mogelijke factoren die allemaal een rol kunnen spelen in het bepalen van de slaagkans van een student. Daarnaast een volgend deel over het knowledge discovery process.

In het eerste deel gaan we op zoek naar factoren die een mogelijke rol kunnen spelen in het bepalen van de slaagkans. We gaan op zoek naar artikels via Ebscohost, Web of Science en Google Scholar met het gebruik van de trefwoorden: Student success, higher education, factors of education success, deteminants of educational succes en success rate of education. Vervolgens proberen we deze factoren te groeperen in verschillende deelgroepen om een overzichtelijke lijst met factoren te krijgen.

In het deel over het knowledge discovery process gaan we vooral op zoek naar welke datamining algoritmes we het best kunnen gebruiken in ons onderzoek. We gaan op zoek naar de voor- en nadelen van elke methode en eerder gebruik van deze methodes in gelijkaardige onderzoeken. Hiervoor consulteren we allereerst het boek Data Mining: Practical Machine Learning Tools and Techniques (Frank, Witten (2005)). Verder gaan we ook op zoek naar artikels over de toepassing van datamining technieken via Ebscohost, Web of Science en Google Scholar. We kijken tevens naar de referenties van de al gevonden artikels en naar de voornaamste auteurs. Nadat we een overzicht hebben gevormd van de verschillende methodes, onderzoeken we welke methodes het beste zijn voor ons onderzoek en zullen we dieper ingaan op deze methodes.

Na de literatuurstudie gaan we over naar de toepassing van het knowledge discovery process. Hierin volgen we de stappen zoals beschreven in onderstaande figuur.



Figuur 3: Het knowledge discovery process - Fayyad 1996.

In het knowledge discovery process start alles vanuit de data, gevormd door de studentendatabase van de UHasselt in ons geval. Na de voorbereidende stappen van het knowledge discovery process gaan we over naar het gebruik van de gekozen datamining algoritmes in WEKA en zo hopen we te komen tot een model dat de slaagkans van studenten kan voorspellen aan de hand van hun kenmerken.

Hoofdstuk 2 : Literatuurstudie

Deze literatuurstudie zal handelen over welke factoren nu precies een relatie hebben met de slaagkans van een student. Zo kunnen we op zoek gaan naar zoveel mogelijk variabelen om eventueel op te nemen in ons model.

2.1. Welke factoren hebben een relatie met de slaagkans van een student ?

We halen hier nogmaals het eerder besproken onderzoek van Declercq en Verboven aan. In hun econometrisch logit model vermeldden ze al een groot aantal factoren die een relatie hebben met de slaagkans van een student. Zo zien we in Tabel 1 dat de gevolgde studierichting in het secundair onderwijs een grote rol speelt. Ook eerder opgelopen studievertraging, geslacht, nationaliteit en zelfs afstand van de woonplaats tot aan de onderwijsinstelling spelen een rol. Verder toont het onderzoek aan dat er ook verschillen zijn tussen de universitaire richtingen zelf en tussen de instellingen. De onderzoekers vonden tevens verschillen tussen verschillende academiejaren.

Determinanten van slaagkansen van eerstejaarsstudenten aan universiteiten: Logit model

Variabele	Parameter	Standaardfout
	Secundaire opleiding (referentiegroep = Econ. - Mod. Talen)	
BSO/KSO	-17,9%	(2,8%)
TSO	-14,8%	(1,4%)
Menswetenschappen	-0,2%	(1,1%)
Steiner	+0,4%	(4,1%)
Mod. Talen - Wet.	+15,1%	(1,0%)
Sportwetenschappen	+17,2%	(1,9%)
Latijn - Mod. Talen	+19,0%	(0,9%)
Mod. Talen - Wiskunde	+21,7%	(1,1%)
Econ. - Wiskunde	+24,5%	(0,8%)
Latijn - Wet.	+30,8%	(0,8%)
Grieks - Latijn	+33,7%	(0,9%)
Grieks - Wetenschappen	+35,9%	(1,9%)
Wet. - Wiskunde	+36,1%	(0,7%)
Latijn - Wiskunde	+38,3%	(0,6%)
Grieks - Wiskunde	+40,7%	(0,9%)
	Net secundaire school (referentiegroep = Vrije school)	
Gemeenschapsonderwijs	-14,5%	(0,6%)
Officieel onderwijs	-10,7%	(1,4%)
	Jaren vertraging in secundair ond. (referentiegroep = geen vertraging)	
-2 jaar in SO	-3,9%	(10,9%)
-1 jaar in SO	+0,8%	(1,4%)
+ 1 jaar in SO	-22,6%	(0,7%)
+ 2 jaar in SO	-31,2%	(1,4%)
+ 3 jaar in SO	-30,6%	(3,0%)
	Andere kenmerken student	
Vrouw	+9,8%	(0,4%)
Buitenlandse nationaliteit	-5,4%	(2,1%)
afstand tot univ. (per 100 km)	+3,1%	(0,7%)

Variabele	Parameter	Standaardfout
	Opleiding aan universiteit (referentiegroep = Econ. en Toeg. Econ. Wet.)	
Verkeerswet.	-16,4%	(6,9%)
Diergeneeskunde	-14,8%	(1,5%)
Biomedische	-13,3%	(1,1%)
Wetenschappen	-10,5%	(0,9%)
Toegepaste wet.	-6,6%	(1,0%)
Toegepaste bio.	-6,2%	(1,2%)
Farmaceutische	-4,5%	(1,4%)
L.O.	-3,3%	(1,6%)
Soc. Gezondheidswet.	-1,6%	(1,3%)
Taal- en letteren	+0,9%	(1,0%)
Rechten	+2,2%	(0,8%)
Pol. En Soc.	+3,5%	(0,9%)
Archeologie	+4,5%	(1,7%)
Psychologie	+4,8%	(0,9%)
Wijsbegeerte	+5,4%	(2,2%)
Gecombineerde	+8,9%	(1,8%)
Geschiedenis	+9,4%	(1,2%)
Tandheelkunde	+18,1%	(3,1%)
Godgeleerdheid	+20,9%	(4,0%)
Geneeskunde	+29,5%	(1,1%)
	Instelling universiteit (referentiegroep = V.U.B. en overige univ.)	
Universiteit Antwerpen	-4,9%	(0,8%)
K.U.Leuven	-10,1%	(0,7%)
Universiteit Gent	-11,7%	(0,7%)
	Academiejaar (basis = 2001-2002)	
2002-2003	+2,0%	(0,7%)
2003-2004	+2,9%	(0,7%)
2004-2005	+1,7%	(0,7%)
2005-2006	-0,5%	(0,7%)
2006-2007	-4,2%	(0,7%)

Tabel 1 : Determinanten van slaagkansen van eerstejaarsstudenten aan universiteiten: Logit model - Declercq, Verboven (2010)

Dit zijn slechts enkele factoren die een rol spelen in de slaagkans van een student. Daarom gaan we verder op zoek in de literatuur naar zulke factoren. We maken zelf een onderverdeling in verschillende categorieën, namelijk: de persoonskenmerken, het persoonlijk verleden van een persoon, de familiale achtergrond en de factoren die een rol spelen tijdens de studie. Factoren die eigen zijn aan de onderwijsinstelling en factoren die gezamenlijk zijn voor alle studenten, zoals bijvoorbeeld een overheidsbeslissing, laten we buiten beschouwing. Verder moeten we er ook van uitgaan dat zeker niet alle factoren die de slaagkans beïnvloeden, al bekend zijn.

2.2. Persoonskenmerken

Persoonskenmerken liggen min of meer vast van bij de geboorte. Ze bevatten een deel vaststaande kenmerken zoals leeftijd en geslacht en daarnaast een deel persoonlijkheids- en emotionele kenmerken die eigen zijn aan een persoon. Deze kunnen echter wel veranderen doorheen de tijd.

2.2.1. Leeftijd

Over de rol van leeftijd is er veel discussie. Graham (1991), Paolillo (1982) en Sulaiman en Mohezar (2006) besloten in hun onderzoek dat leeftijd niet significant gecorreleerd was aan de resultaten van studenten.

Peiperl en Trevelyan (1997), Ekpenyong (2000) en Ahmadi, Raiszadeh en Helm's (1997) vonden dan weer dat leeftijd wel een predicatieve waarde had, namelijk dat jongere studenten beter presteerden dan oudere studenten.

2.2.2. Geslacht

Deckro en Woudernberg (1997) vonden dat vrouwelijke studenten beter presteren dan hun mannelijke tegenhangers. Dit wordt bevestigd door Cheung en Kan (2002). Launius (1997) verklaarde dit verschil door te zeggen dat vrouwelijke studenten meer tijd en moeite steken in hun studies. Sulaiman en Mohezar (2006) weerlegden deze hypothese dan weer.

2.2.3. Etniciteit

Ook omtrent deze variabele bestaan er verschillende meningen. Clayton en Cate (2004) merken op dat studenten van verschillende etnische achtergrond verschillen in prestaties op academisch gebied. Johnson (2005) en Sulaiman en Mohezar (2006) spreken deze resultaten dan weer tegen.

2.2.4. Taal

Het onderzoek van Yang and Lu (2001) haalt aan dat taal een predictor is voor de resultaten van een student, moedertaalsprekers presteren beter dan andere studenten.

2.2.5. Cognitieve intelligentie

Een factor die een belangrijke rol speelt in de slaagkans van een student is zijn cognitieve intelligentie of zijn denkvaardigheid. Dit aspect van de intelligentie meten we aan de hand van testen en is onder meer een onderdeel van gestandaardiseerde toelatingsproeven voor hogescholen en universiteiten in de Verenigde Staten. Ook in Vlaanderen werd in het verleden heel wat onderzoek gedaan naar de samenhang tussen algemene intelligentie en studieresultaten in het hoger onderwijs (Masui (2002)).

2.2.6. Persoonlijkheid

Extroverte studenten zouden tegenwoordig meer succesvol zijn. Dit is te verklaren doordat de curricula momenteel meer nadruk leggen op vaardigheden zoals communicatie en teamwork en er is meer en meer behoefte aan verbale rapportering over het werk. Dit zijn vaardigheden waar introverte personen het moeilijker mee hebben (Burton en Dowling (2005)).

2.2.7. Zelfvertrouwen

Zelfvertrouwen helpt studenten om hun leeromgeving te managen. Het helpt hen ook om beter te presteren voor vakken waar samenwerking met andere studenten vereist is. Dit alles zal dan leiden tot een hogere slaagkans (Burton en Dowling (2005)).

2.2.8. Sociale responsabiliteit

Sociale responsabiliteit staat voor de bekwaamheid om op een meewerkende en constructieve manier deel uit te maken van een sociale groep. Sociale responsabiliteit is een belangrijke predictor voor de slaagkans (Sparkman, Maulding en Roberts (2012)).

2.2.9. Flexibiliteit

Flexibiliteit, dat we definiëren als de mogelijkheid om jezelf te kunnen aanpassen aan situaties, is negatief gerelateerd met afstudeerstatus. Dit verklaren we ondermeer doordat flexibele studenten sneller van richting gaan veranderen (Sparkman, Maulding en Roberts (2012)).

2.3. Persoonlijk verleden

De in het verleden gemaakte keuzes van de student, zowel op studie- als op persoonlijk gebied, spelen ook een rol in zijn of haar slaagkans.

2.3.1. Eerdere studies

De studenten die het best voorbereid uit het secundair onderwijs komen, zijn het best geplaatst om goed te presteren in hogere studies (Florida Department of Education (2005); Gladieux and Swail (1998); Horn en Kojaku (2001); Martinez en Klopott (2003); Warburton, Bugarin en Nunez (2001)). Volgens Pike en Saupe (2002) zijn de resultaten uit het middelbaar een sterke voorspeller van de slaagkansen in het eerste jaar van hogere studies. Zo verklaren ze ongeveer 25 procent van de variantie.

Bepaalde vakken kunnen daarnaast ook een belangrijke invloed op de slaagkans in de hogere studies hebben. Zo zouden studenten die wiskunde of wetenschappen hebben gevolgd een 25% hogere slaagkans hebben dan zij die deze vakken niet hebben gevolgd (Adelman (1999); (Warburton, Bugarin en Nuñez (2001)). Ely and Hittle (1990) vonden in hun onderzoek dat het eerder gevolgd hebben van boekhoudlessen een duidelijke determinant was van studentensucces in economische studies. Anderson en Benjamin (1994) en Didia en Hasnat (1998) besloten dat studenten die eerder wiskunde en calculus gevolgd hebben beter presteren in economische en financiële vakken dan studenten die deze ervaring niet hebben.

2.3.2. Eerder blijven zitten

Eerdere studievertraging kan leiden tot lager zelfgeloof en een minder engagement (Martin, 2009) en kan op die manier een negatieve invloed uitoefenen op academisch prestaties (Masui et al. (2012)).

2.3.3. Werkervaring

Hierover zijn er verschillende meningen, McClure, Wells en Bowerman (1986) indiceerden een positief verband tussen werk ervaring en de slaagkans. Dit omdat de werkervaring die de student opdoet hem aan een breder beeld over het bedrijfsleven helpt en dit een voordeel biedt tegenover studenten die dit niet hebben. Deze studenten zouden ook beter de relevantie en potentiële toepassing van het aangeleerde materiaal zien volgens Dreher en Ryan (2000).

Dugan, Grady, Payn, Baydar,& Johnson(1996) Graham (1991) Peiperl & Trevelyan (1997) Fisher & Resrick (1990) Sulaiman & Mohezar (2006) besloten dan weer dat werkervaring niet gerelateerd was aan de resultaten van een student.

2.3.4. Andere factoren die slaagkans negatief beïnvloeden

Binnen verschillende onderzoeken vonden we nog andere factoren die een negatieve invloed uitoefenen op de kans om te slagen. Zo zagen we onder meer: het uitstellen van hogere studies naar een later tijdstip (Adelman 2006), het parttime volgen van hogere studies (Community College Survey of Student Engagement (2005)) en het zorg moeten dragen voor een kind (Community College Survey of Student Engagement (2005)). Daarnaast bemerkten we ook dat het op je eigen inkomen moeten betrouwen voor het betalen van je studies, de slaagkans niet ten goede komt (Community College Survey of Student Engagement (2005)). Dit is eveneens zo voor interrupties in je studies (Pascarella and Terenzini (2005)).

2.4. Familiale achtergrond

De slaagkans van een student is ook onderhevig aan factoren van zijn of haar familiale achtergrond. Merk op dat al deze factoren een grote overlap hebben, zoals het bijvoorbeeld het feit dat studenten met een studiebeurs een familie met een lagere socio-economische status en waarschijnlijk ook ouders met een lagere diploma zullen hebben

2.4.1. Diploma van ouders

Studenten wiens ouders in het bezit zijn van een bachelordiploma hebben vijf maal meer slaagkans dan studenten waar dit niet het geval is (Pascarella and Terenzini 2005). Indien de ouders geen hogere diploma's bezitten, zou er een gebrek aan ervaring zijn om de student te ondersteunen. Dit zowel emotioneel als in het begrijpen van engagement nodig om te slagen in hogere studies (Sparkman, Maulding en Roberts (2012)).

Verder hebben studenten wiens ouders geen bacherlordiploma hebben een minder ontwikkeld time management en ook andere vaardigheden zijn minder ontwikkeld. Daarnaast bezitten ze minder kennis over hogere studies en hebben ze minder ervaring met bureaucratische instellingen (Attinasi 1989); London, 1989; Nuñez en Cuccaro-Alamin, 1998; Terenzini et al. (1996); York-Anderson en Bowman (1991)).

2.4.2 Ondersteuning van het gezin

Studenten zullen beter presteren wanneer hun familie hun keuzes ondersteunt en de student aanmoedigt om te blijven volharden (Gutierrez, 2000; Pathways to College Network (2004); Tierney, Corwin, en Colyar (2005)).

2.4.3 Gescheiden ouders

Studenten van wie de ouders gescheiden zijn voor hun 18^{e} verjaardag hebben 45% minder kans om te slagen in het hoger onderwijs. Dit komt ondermeer doordat er bij dergelijke gezinnen minder financiële middelen zijn en er minder contact tussen kind en ouder is (Havermans , Vanassche (2013)).

2.4.4 Socio-economische status

Met een stijging van het familie inkomen stijgt ook de slaagkans van de student. Dit komt doordat bij hogere familie-inkomens er een betere academische voorbereiding, een hogere verwachting en meer familieondersteuning bestaat (Coleman 1988).

2.4.5 Financiële hulp

Studenten die financiële hulp ontvangen zoals een beurs of lening hebben een betere slaagkans (The Pell Institute (2004)).

2.5. Tijdens de studie

Wat een student doet of laat tijdens zijn studie speelt ook een rol in het wel of niet slagen.

2.5.1 Werken tijdens studies

Werken tijdens de studies heeft pas een negatief effect vanaf dat er meer dan 15 uur per week gewerkt wordt (Choy 1999) (Pascarella 2001). Meer nog, werken op de campus heeft zelfs een positief effect op de slaagkans van de student, daar het een communicatiekanaal voorziet voor de student en hem helpt het onderwijssysteem doeltreffend te gebruiken (Institute for Higher

Education Policy (2001); Kuh et al. (2005)). Ook zal werken zorgen voor een beter zelfbeeld (Gleason1993), integratie met de campus omgeving (Murdock 1990) en een betere volharding (Heller, Pascarella et al. (1998)). Teveel werken heeft echter wel een negatief effect daar dit het naar de les gaan en het studeren zal doen verminderen (Heller 2002).

2.5.2 Studeergedrag

Opletten in de les, goede notities nemen en voorbereid naar de les gaan, zullen de slaagkans van een student verhogen (Borg en Gall. (1989)) (Okpala, Okpala, en Ellis (2000)).

2.5.3. Tijd investeren

Een belangrijke factor die de slaagkans bepaalt, is de hoeveelheid tijd en moeite die de student steekt in zijn studies (Alexander en Murphy (1994); Astin (1993); Bailey, Jenkins en Leinbach (2005); Masui et al. (2012)) .

2.5.4. Contact tussen faculteit en student

Informeel contact tussen de student en faculteitsleden, zoals te gast zijn bij een professor thuis of werken aan een project met een faculteitslid heeft een positieve correlatie met ontwikkeling en engagement van een student en dus ook met zijn slaagkans (Astin, 1993; Kuh, 2003; Kuh and Hu 2001; Pascarella and Terenzini 1977, 1991, 2005).

2.5.5. Peer interactie

Interactie met medestudenten kan gaan van werken en studeren in groepsverband tot het lid zijn van studentenverenigingen. Dit kan de academische ontwikkeling, het verkrijgen van kennis en het zelfbeeld van een student positief beïnvloeden. Dit alles zal dus ook gaan leiden tot een hogere slaagkans (Kuh 1993, 1995).

2.5.6. Extra-curriculaire activiteiten

Deelnemen aan extra-curriculaire activiteiten zorgt ervoor dat studenten psychologisch en sociaal verbonden zijn met een groep en zo een hoger verlangen hebben om te willen slagen (Hanks and Eckland (1976)). Ook zullen deze studenten zich inzetten voor activiteiten die hen zullen helpen de juiste vaardigheden en competenties te verwerven om te kunnen slagen in hun studies (Pascarella and Terenzini (2005)).

2.5.7. Studententevredenheid over de instelling

Als een student een bepaalde betrokkenheid met en loyaliteit voelt voor de instelling, zal deze betere prestaties behalen (Bean, 1980; Bean and Bradley, 1986; Bean and Vesper, 1994) (Pike 1991,1993).

2.5.8. Leven op de campus

Doordat studenten op de campus leven, zullen ze meer interacties hebben met de faculteit en hun medestudenten en dus ook een hogere beleving hebben van hun studies (Pascarella and Terenzini 1991; 2005). Dit alles zal zorgen voor een positief effect op de slaagkans van de student (Blimling 1993).

2.5.9. Eerstejaarsseminarie

Een eerstejaarsseminarie is een cursus special ontworpen om de academische vaardigheden en sociale ontwikkeling van studenten te verbeteren. Het volgen van zulk seminarie zorgt ervoor dat een student meer interactie zal hebben met de faculteit, de faculteit positiever zal gaan bekijken en

meer gebruik zal maken van services op de campus (NSSE 2005). Dit alles zal er toe leiden dat studenten die een seminarie volgden hogere punten gaan behalen en minder snel stoppen met hun studies (Pascarella en Terenzini (2005)) (Carstens (2000)).

2.5.10 Gebruik academische support services

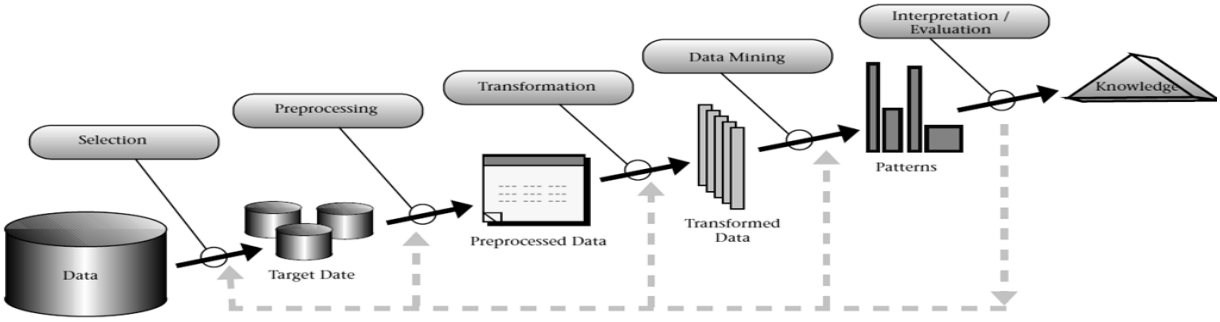
Het gebruik van hulpmiddelen door de faculteit aangeboden zoals schrijfvaardigheidlessen, leren plannen, financieel advies of bijlessen heeft een positief effect op de slaagkans van de student (Hossler, Kuh en Olsen 2001) (Hoyt 1999) (The Pell Institute (2004)).

2.5.11 Factoren buiten de student om

Tijdens de studie kunnen er ook onverwachte situaties optreden zoals ziekte of sterfgevallen in de familie. Deze zullen de slaagkans ook negatief beïnvloeden.

Hoofdstuk 3: Het knowledge discovery process

Om ons model te maken, gebruiken we de stappen in het knowledge discovery process (KDD) zoals hieronder beschreven in Figuur 4. Het doel van het knowledge discovery process is om nieuwe, bruikbare, geldige en begrijpbare patronen te vinden in data (Fayyad,Piatetsky-Shapiro, and Smyth 1996). Belangrijk is om het verschil duidelijk te maken tussen KDD en datamining. KDD verwijst naar het algemene proces van het vinden van bruikbare informatie in data, terwijl datamining slechts een stap is in dat proces. Meer specifiek is datamining de toepassing van algoritmes om de patronen in de data te vinden.



Figuur 4 : Het knowledge discovery process - Fayyad 1996.

3.1. Stappen KDD process

We beginnen vanuit de data. We moeten het domein van de data begrijpen en het doel van de studie duidelijk vastleggen. Dan volgt de eerste stap, de 'selection'. Daarin wordt de benodigde data, waarop de effectieve datamining gedaan gaat worden, geselecteerd uit de hele dataset. Zo komen we tot de target data. Volgens het principe "garbage in, garbage out" moet de data grondig gecontroleerd worden op outliers, missing data en het behandelen van bepaalde datatypen. Er moet alles aan gedaan worden om zoveel mogelijk ruis uit de data te verwijderen, dit heet 'preprocessing'. Uit deze stap volgt dan de preprocessed data. Nu moeten de data 'getransformeerd' worden. Dat wil zeggen dat we de data in de juiste vorm zetten, zodat deze kan worden gelezen door het datamining programma. Eventueel kan er bij deze stap ook nog aan datareductie worden gedaan om tot een lager aantal variabelen te komen. Nu volgt de belangrijkste stap in het proces, namelijk de datamining zelf. Zo proberen we te komen tot de onderliggende patronen die de data bevatten. Hier gaan we in het volgende deel dieper op in. Als we de gevonden patronen dan kritische gaan interpreteren en evalueren, komen we tot nieuwe kennis.

3.2. Datamining algoritmes

Datamining algoritmes kunnen we onderverdelen in 2 grote categorieën, namelijk supervised en unsupervised learning algoritmes. Het onderscheid tussen deze 2 methodes bestaat in de manier waarop het algoritme de data classificeert. Bij supervised learning is de doelvariabele, die we willen gaan voorspellen, op voorhand vastgelegd. Dit terwijl bij unsupervised learning algoritmes er niks op voorhand vastligt. Daar is het namelijk de bedoeling verbanden te zoeken tussen alle variabelen onderling terwijl we bij supervised learning de doelvariabele willen gaan voorspellen aan de hand van de andere variabelen. We bekijken in deze masterproef voornamelijk de supervised learning algoritmes omdat het ons uiteindelijke doel is om een vaststaande variabele, namelijk de slaagkans van een student, te gaan voorspellen. Allereerst bekijken we de OneR, het meest eenvoudige algoritme dat er is. Dit algoritme baseert zijn classificatie op één regel. Verder kiezen we voor drie van de meest bekende en gebruikte soorten algoritmes, namelijk naïeve bayes, beslissingsbomen en het ingewikkeldere neurale netwerk. Bij deze 3 modellen zal het algoritme een model leren en bouwen aan de hand van de voorbeelddata die we als input geven. Deze voorbeelddata bevat zowel de inputvariabelen als de doelvariabele die we willen gaan voorspellen. Aan de hand van deze voorbeelddata kan het algoritme dan een model leren en maken zodat, als we een nieuwe case hebben, het model de doelvariabele zal kunnen gaan voorspellen. We bekijken tevens nog een iets afwijkend algoritme, namelijk het k-Nearest neighbor algoritme. Dit algoritme bouwt geen echt model, maar zal nieuwe cases gaan vergelijken met de voorbeelddata om zo de doelvariabele te gaan proberen te voorspellen. We kunnen nu nog niet voorspellen welk algoritme het best zal presteren, maar met deze vijf algoritmes hebben we een weide verscheidenheid aan supervised learning algoritmes om onze voorspellingen te gaan doen. Elke algoritme heeft wel zijn eigen voor-

en nadelen. Deze zullen we hieronder in het kort bespreken. Voor meer uitleg verwijzen we echter naar het boek "Data Mining: Practical Machine Learning Tools and Techniques" (Frank, Witten (2005)).

Als laatste bekijken we nog een unsupervised learning algoritme, namelijk clustering. Dit algoritme zal geen doelvariabele voorspellen, maar zal de cases gaat groeperen. Clustering kan ons ook eventueel nuttige verbanden aantonen. Het kan bijvoorbeeld groepen van vakken gaan groeperen op basis van de resultaten behaald op deze vakken. Zo zou het kunnen dat bijvoorbeeld een slecht resultaat op wiskunde meestal gepaard gaan met een slecht resultaat op statistiek.

3.2.1 OneR classifier

OneR, dat staat voor one rule is het meest eenvoudige algoritme. Dit algoritme gaat een regel generen voor elke variabele in de data. Daarna pakt het de regel met de laagste voorspellingsfout als classificatieregel. Hoewel het zeer ongecompliceerd is, zal oneR toch vaak in staat zijn om een goede classificatie te doen. De zeer eenvoudige interpreteerbaarheid zien we als een groot voordeel van deze methode. OneR is ontwikkeld met als doel het aantonen dat eenvoudige classifiers vaak ook een goede accuraatheid konden behalen en dat een eenvoudige regel meestal niet veel slechter presteert dan de ingewikkelde classifiers. Het percentage juist geclassificeerde cases van OneR kan dus gebruikt worden om de prestaties van de andere algoritmes te benchmarken (Holte (1993)).

We geven hieronder een voorbeeld.

Which one is the best predictor ?

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Op basis van bovenstaande tabel gaan we een eenvoudige telling doen. Voor elke variabele wordt apart gekeken welke waarde de doelvariabele het best voorspelt. Dan bekijken we welke variabele globaal de kleinste foutenmarge geeft en die wordt gekozen om de regel op te stellen, in dit geval outlook.

Frequency Tables

★		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3

Figuur 5 : Hoeveelheidstellingen per variabelen

★		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

IF Outlook = Sunny THEN PlayGolf = Yes
IF Outlook = Overcast THEN PlayGolf = Yes
IF Outlook = Rainy THEN PlayGolf = No

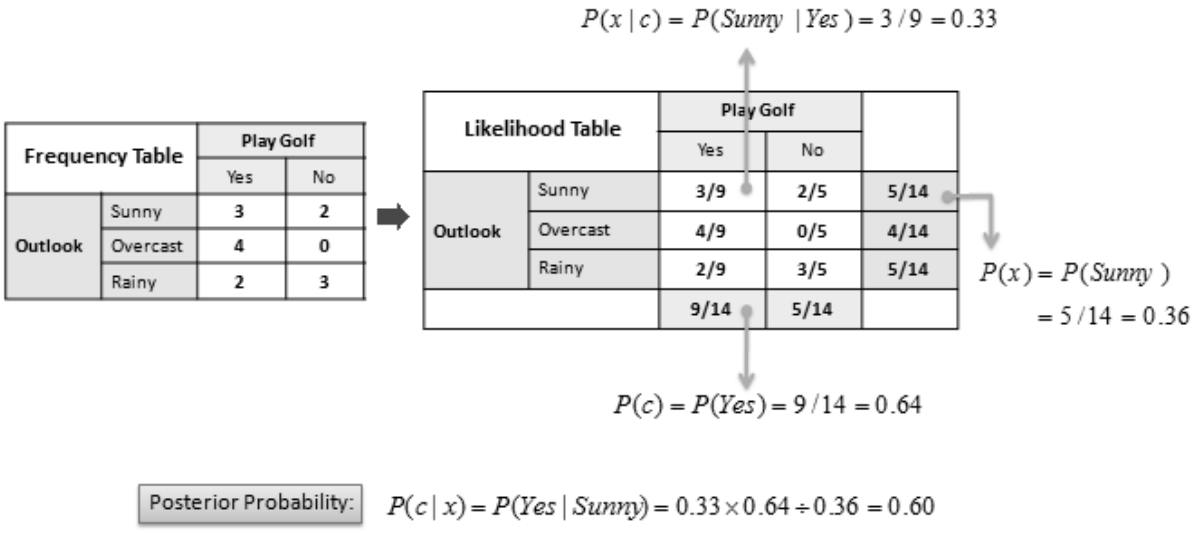
Figuur 6: OneR beslissingsregel

3.2.2. Naive bayes classifier

De naive bayes classifier is gebaseerd op Bayes' theorema met 'naïeve' onafhankelijkheids-assumpties tussen de variabelen. De waarde van een variabele is dus niet gerelateerd aan de waarde van een andere variabele binnen dezelfde case. Alle variabelen worden bijgevolg behandeld alsof ze compleet onafhankelijk zijn. Daarnaast hebben alle variabelen een even belangrijke contributie.

Ondanks het naïeve design en de eenvoudige basisassumpties kan een naive bayes classifier toch goede classificatie resultaten behalen in vele complexe echte wereld situaties. Het grote voordeel van een naive Bayes classifier is dat er slechts een kleine hoeveelheid data nodig is om de classifier te trainen. Andere voordelen zijn de gemakkelijke interpreteerbaarheid en het feit dat het perfect kan omgaan met missing values.

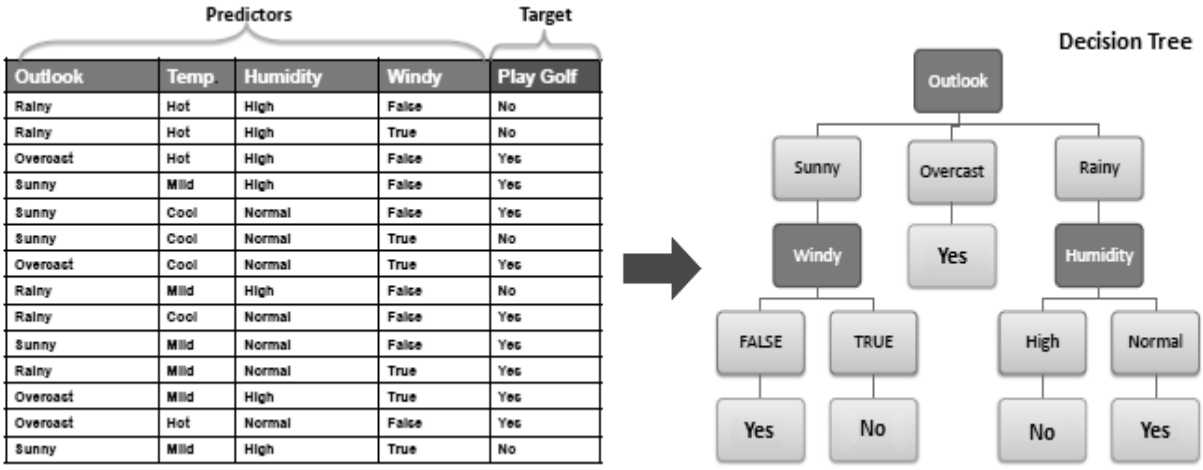
We illustreren dit met een voorbeeld. Eerst wordt er per variabele een telling uitgevoerd van elke mogelijke waarde ten opzichte van de doelvariabele. Deze hoeveelheidstelling wordt dan omgevormd naar fracties. Voor elke nieuwe case kan er dan, via een statische berekening, voor elke mogelijke uitkomst van de targetvariabele een kansvoorspelling worden gegeven.



Figuur 7: Berekening van de naive bayes classifier

3.2.3. Beslissingsbomen

Een beslissingsboom is een verzameling van knooppunten die verbonden zijn met keuzentakken. Ze gaan zo naar onder gaan tot ze uiteindelijk tot een classificatie komen. Op elk knooppunt wordt er een test uitgevoerd op een variabele. Het ontwikkelingsproces van beslissingsbomen bestaat uit twee delen. Allereerst wordt de boom ontwikkeld van boven naar beneden. Daarna wordt er terug 'gesnoeid' (pruning). Bij het snoeien worden overbodige takken verwijderd indien het de classificatie bevorderd. Een van de meest gebruikte beslissingsboom algoritmes is C4.5 (Quinlan 1993), dat gaat aan de hand van de information gain bij elke stap kijken welke variabele de dataset het beste opsplitst in subsets. De uitleg hiervan is redelijk technisch dus voor meer info hierover verwijzen we naar opnieuw naar het boek Data Mining: Practical Machine Learning Tools and Techniques (Frank, Witten (2005)).



Figuur 8 : Voorbeeld van omzetting van tabel naar beslissingsboom

Het grote voordeel van beslissingsbomen is dat deze gemakkelijk te interpreteren zijn. Andere voordelen zijn dat er weinig datapreparatie nodig is en een beslissingsboom zowel continue als categorische data aankan. Het grootste probleem van beslissingsbomen is overfitting. Het algoritme vormt de beslissingsboom op basis van de trainingsset. In sommige gevallen gaat de boom teveel gefit zijn op de trainingset en dus minder goed presteren op de nieuwe ongeclassificeerde cases. Het model gaat dan ook de error mee beschrijven en niet enkel de onderliggende relatie. Overfitting kan deels tegengegaan worden door het 'snoeien' van de boom.

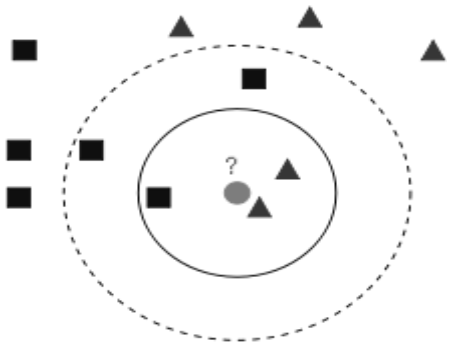
Na het vormen van de beslissingsboom, kan er aan de hand van de boom beslissingsregels worden ontwikkeld. Die regels leggen de classificatie van de boom vast in zinnen. Bijvoorbeeld: *Als* Outlook zonnig is & er is wind *Dan* gaan de kinderen buiten spelen.

3.2.4 K-Nearest neighbor algoritme

Het k-nearest neighbor algoritme is een instance-based algoritme. In dit soort algoritmes wordt een nieuw, niet-geclassificeerd geval vergeleken met alle reeds geclassificeerde cases. Via een afstandsfunctie, meestal de Euclidische afstand, wordt dan gekeken met welke cases de nieuwe case het meest overeen komt. Dan dient er nog bepaald te worden welke classificatie de nieuwe case krijgt. Dit gebeurt door alle cases die de meeste overeenkomsten hebben met de nieuwe case te gaan combineren om zo tot een classificatie te komen. We kunnen dit doen door gewogen of ongewogen stemming. Bij beide methodes dient eerst de waarde van de variabele k vastgelegd te worden. Deze waarde staat voor het aantal cases dat gebruikt gaat worden om de classificatie te doen. Bij een niet gewogen stemming gaan dan alle cases een even groot rol spelen, terwijl bij gewogen stemming de cases die het meeste gelijkenis hebben met de nieuwe niet-geclassificeerde case de grootste rol gaan spelen bij de classificatie.

Het k-nearest neighbor algoritme bezit ook enkele nadelen. Zo kan het bijvoorbeeld bestaan dat alle attributen van een case een even grote rol gaan spelen in het bepalen van de classificatie, terwijl dit zelden zo zal zijn in echte wereld situaties. Ook moeten alle mogelijk gevallen genoeg vertegenwoordigd zijn tussen de al geclassificeerde cases, zodat de resultaten niet vertekend gaan zijn. Verder is het algoritme ook zeer afhankelijk van de gekozen k waarde.

We geven hierbij een voorbeeld.

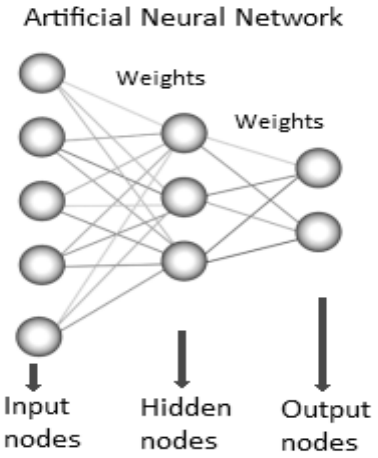


Het bolletje dient geclassificeerd te worden. Indien we k=3 nemen (volle cirkel), zal het geclassificeerd worden als een driehoek. Pakken we echter k = 5 (cirkel bestaande uit stippellijn), dan zal het bolletje geclassificeerd worden als vierkant.

3.2.5 Neurale netwerken

Een neuraal netwerk is een niet lineaire data modellering tool. Het wordt gebruikt om complexe relaties tussen input en output te modelleren of om verbanden in de data te vinden. Een neuraal netwerk bestaat uit vele knooppunten die met elkaar verbonden zijn. Het netwerk leert door individuele cases te bekijken en een voorspelling te maken van elke case. Als het een incorrecte voorspelling maakt, zal het de gewichten in zijn netwerk gaan aanpassen. Ook hiervan is de uitleg zeer technisch en ingewikkeld, voor meer info verwijzen we naar Warren (2004) .

Het grote voordeel van neurale netwerken is het feit dat ze zeer goed kunnen omgaan met missende waardes en noisy data. Het belangrijkste nadeel is dat ze moeilijk te interpreteren zijn voor de mens.



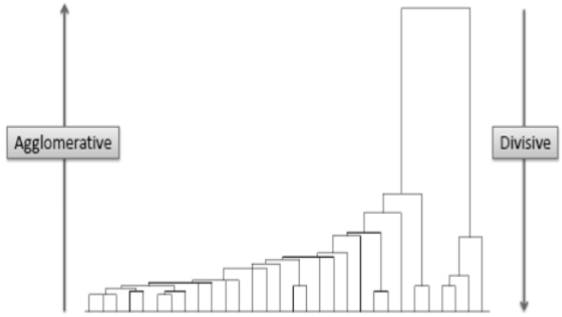
Figuur 9: Grafische voorstelling van een neuraal netwerk

3.2.6. Clustering

Bij clustering worden op elkaar gelijkende cases gegroepeerd. De bedoeling is om zo verschillende groepen van gelijkende cases te verkrijgen, waarbij binnen een groep de cases zo gelijk mogelijk zijn en de verschillen tussen de groepen cases, dus clusters, zo groot mogelijk zijn. Bij clustering is er dus geen predictie van een doelvariabele. De twee meest gebruikte types zijn hiërarchische en k-means clustering. Voor meer info verwijzen we naar Fung 2001. Bij hiërarchische clustering wordt stap per stap telkens een enkele case bij een andere gevoegd totdat alles uiteindelijk in een cluster zit. Naderhand wordt er dan gekeken aan de hand van enkele statistische kenmerken welk aantal clusters het beste is. Bij k-means clustering wordt het aantal clusters op voorhand bepaald tesamen met de clustergemiddeldes. Vervolgens wordt elke case toegevoegd aan de cluster waarbij hij zich het dichtst bevindt. Daarna worden de clustergemiddeldes opnieuw berekend en het proces opnieuw uitgevoerd. Dit wordt dan herhaald tot er een stabiele oplossing is.

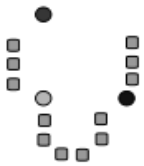
Clustering kan ook als een voorbereidende stap in het datamining proces worden gebruikt. Dan kunnen de resultaten van de clustering dienen als input voor andere algoritmes zoals bijvoorbeeld neurale netwerken. Op deze manier zal de datagrootte eerst verkleind worden door het groeperen en zullen de volgende algoritmes minder tijd nodig hebben voor hun model te bouwen.

Hiërarchische clustering

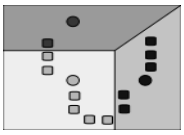


Figuur 10: Grafische voorstelling van hiërarchische clusterring

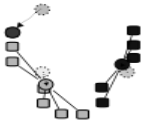
K-means clustering



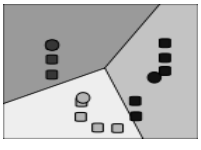
1) Drie initiële gemiddeldes worden random gegenereerd (k=3), voorgesteld door de bolletjes op de tekening.



2) Drie clusters worden gevormd door elke observatie te associëren met zijn dichtste gemiddelde.



3) Het midden van elke cluster wordt het nieuwe gemiddelde



4) Stappen twee en drie worden herhaald tot convergentie is bereikt

Hoofdstuk 4: Resultaten van andere onderzoeken

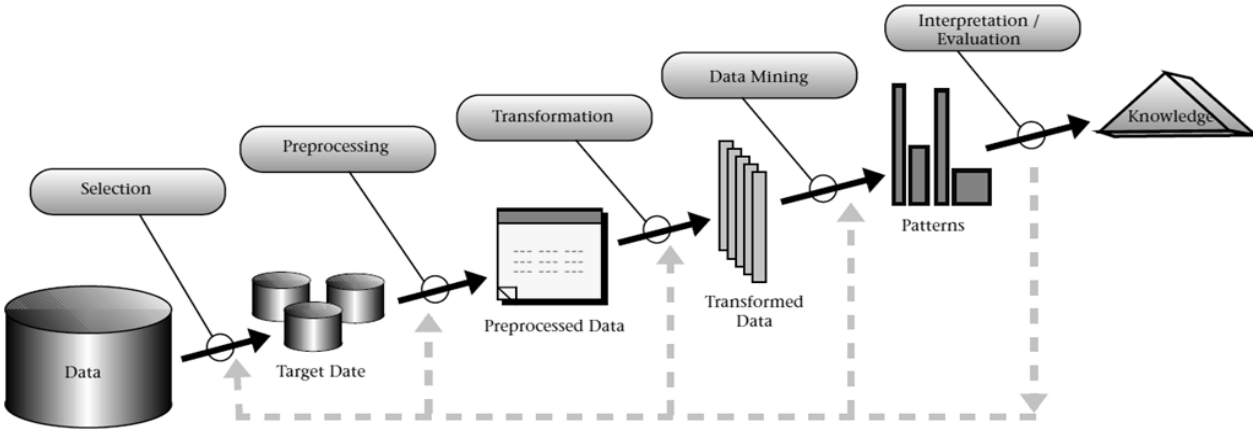
We bekeken voor deze masterproef ook enkele eerder uitgevoerde onderzoeken naar de slaagkans van een student, de gebruik maken van datamining. Welke variabelen deze onderzoeken als inputvariabelen gebruikten, is terug te vinden in Bijlage 1. Het percentage correct geclassificeerde cases van de algoritmes in deze onderzoeken is samengevat in volgende tabel.

	OneR	Naive Bayes	Decision tree	k-nearest neighbour	Neuraal netwerk
Kabakchieva (2012)	67%		73%	70%	74%
Dekker (2009)	68%	71%	70%		
Osmanbegović et al, (2012)		77%	74%		71%
Dragičević et al, (2012)			83%		
Oladokun et al, (2008)					74%
Bijayananda, Srinivasan (2004)					89%
Schumacher et al. (2010)			78%		86%
Zlatko en Kovačić (2010)			60%		

Er is duidelijk niet één bepaald algoritme dat er bovenuit steekt. Afhankelijk van het onderzoek presteert een ander algoritme beter dan de anderen. Als we dan toch een algoritme als beste moeten bestempelen, zien we dat een neuraal netwerk meestal het best presteert en dat dit in sommige gevallen zelfs een zeer hoog percentage correct geclassificeerde cases bevat. Dit is dan ook het meest ingewikkeld te interpreteren algoritme. We zien ook dat het percentage correct geclassificeerde telkens ongeveer rond de 70% schommelt. We kunnen dit dus aannemen als een goed richtpercentage voor ons eigen onderzoek.

Hoofdstuk 5: Toepassing KDD process

In dit hoofdstuk gaan we over tot de toepassing van de theorie die we in de vorige hoofdstukken hebben besproken. We zullen nu proberen om een model te bouwen dat de slaagkans van een student kan voorspellen aan de hand van zijn of haar eigenschappen. Hiervoor volgen we de stappen in het knowledge discovery process (KDD), zoals beschreven in onderstaande figuur.



Figuur 11 : Het knowledge discovery process - Fayyad 1996.

5.1. Kennismaking met de data

Via de UHasselt krijgen we toegang tot de studentendatabase BEW van het jaar 2004. Deze database bevat info over 192 studenten en is opgesplitst in vier delen. Deze delen zijn aan elkaar te linken via een fictief studenten ID, de enige variabele die ze alle 4 gemeenschappelijk hebben. We beschrijven de 4 delen van de database hieronder met de variabelen die ze bevatten. In Bijlage 2 stellen we een voorbeeld beschikbaar van hoe de data er uitzien met telkens 2 voorbeeldcases.

Student

Dit deel bevat info over de achtergrond van de student, zijn eerdere studies, en het behaalde eindresultaat aan de instelling.

- Student ID
- Geslacht (binair)
- Nationaliteit (categorie)
- Gemeente van afkomst (categorie)
- Land van afkomst (categorie)
- Op kot of niet (binair)
- Beursstudent of niet (binair)
- Naam secundaire school (categorie)
- Gemeente secundaire school (categorie)
- Richting secundaire school (categorie)
- Aantal uren van vakken gevolgd in secundaire school (categorie)
    - Wiskunde
    - Natuurkunde
    - Scheikunde
    - Frans
    - Engels
    - Duits
- Studievertraging in secundair (binair)
- Generatiestudent of niet (binair) (in 1986 geboren)
- Studiepunten behaald in eerst jaar (categorie)
- Studiepunten behaald na deliberatie (categorie)
- Vakken niet afgelegd (aantal)

- Percentage op het einde van heet jaar (percentage)
- Graad op het einde van het jaar (categorie)
- 1ste of 2de zit geslaagd (binair)
- Welk bachelordiploma behaald
- Het academiejaar in welk men het bachelordiploma heeft behaald
- Welk Master diploma behaald
- Het academiejaar in welk men het Master diploma heeft behaald

Resultaten 2004

Het tweede deel bevat informatie over de resultaten die de student behaalde in zijn eerste bachelorjaar.

- Student ID
- Academiejaar (altijd 2004)
- Naam van de opleiding (altijd bachelor Tew-Hi-Bi)
  (In 2004 was het eerste jaar Tew/Hi/Bi nog volledig gemeenschappelijk. Een keuze moest pas gemaakt worden na het 1ste jaar)
- Datum uitschrijving (altijd: niet gebeurd)
- Intern vaknummer
- Vaknummer volgens studiegids
- Naam van het vak
- Type vak (hoofdvak/deelvak)
- Code voor het vak gedurende de drie trimesters (afwezig, uitgesteld, ..)
- Punten voor het vak gedurende de drie trimesters
- Code voor de totale eerste kans per vak
- Punt voor de totale eerste kans per vak
- Code voor de tweede kans per vak
- Punt voor de tweede kans per vak
- Code voor het totale eerste jaar per vak
- Punt voor het totale eerste jaar per vak

Studietijdmeting

Gedurende het eerste trimester moesten een deel studenten uit de database ook verplicht deelnemen aan de studietijdmeting. Zij moesten elke week hun studieactiviteit neerschrijven in een daarvoor ontwikkelde applicatie. Op deze manier beschikken we ook over de studiegegevens van de deelnemende studenten, namelijk 59 van de 192 studenten, gedurende het eerste trimester.

- Vaknummer volgens de studiegids
- Weeknummer
- Type activiteit (HC, ZS, WZ , …)
- Naam van het vak
- Aantal gespendeerd minuten
- Datum registratie
- Student ID
- Voorziene studiebelasting

Denk- en leestest

Aan het begin van het eerste trimester wordt er bij de studenten ook een denk- en leestest afgenomen. Dit werd gedaan om te peilen naar de intellectuele capaciteit van de studenten.

- Student ID
- Deelgenomen aan leestest of niet
- Totaal resultaat leestest
- Totaal resultaat op denktest
- Deelresultaat denktest verbaal
- Deelresultaat denktest numeriek
- Deelresultaat denktest grafisch

De data in de gehele database zijn volledig geanonimiseerd. Dat wil zeggen dat we niet werken met echte namen en waarden, maar met vervangende waardes. Geslacht, secundaire school, steden, richting secundair, op kot of niet, beursstudent of niet, studievertraging of niet en nationaliteiten zijn omgezet naar nummers. Deze anonimiteit bewaart de privacy van de studenten.

De studentendatabase bevat ook slechts een beperkt aantal factoren die een relatie hebben met de slaagkans van een student. Als we vergelijken met de factoren uit de literatuurstudie zien we dat een heel deel factoren niet zijn inbegrepen in de studenten database. Dit kan verschillende redenen hebben, een deel factoren zijn moeilijk meetbaar, andere factoren zijn dan weer gewoon weg niet opgenomen in de database. We overlopen de factoren uit de literatuurstudie even systematisch.

Als we kijken naar de persoonskenmerken zien we dat er een heel deel factoren, rechtstreeks of onrechtstreeks, zijn opgenomen in de database (leeftijd, geslacht, etniciteit, taal, cognitieve intelligentie). Toch ontbreken er nog een heel deel factoren (persoonlijkheid, zelfvertrouwen, sociale responslabiliteit, flexibiliteit). Dit komt waarschijnlijk omdat deze factoren moeilijk meetbaar zijn, toch valt het eventueel te overwegen deze factoren in de toekomst toch op te nemen in de database. Deze kunnen dan bijvoorbeeld gemeten worden aan de hand van persoonlijkheidstesten.

Als we dan verder gaan kijken naar de factoren uit het persoonlijke verleden van de student, zien we dat ook hier weer een deel factoren zijn opgenomen (eerdere studies, eerder opgelopen studievertraging) maar een andere deel dan weer niet (werkervaring, andere factoren zoals parttime studeren, zorg dragen voor een kind, …). Er is niet direct een voor de hand liggende reden waarom deze factoren niet zijn opgenomen in de database, dus het is eventueel ook te overwegen om deze factoren in de toekomst ook op te nemen in de database.

Dan komen we bij de factoren die betrekking hebben op de familiale achtergrond van de student. Buiten of de student financiële hulp ontvangt onder de vorm van een studiebeurs, zijn hiervan geen factoren opgenomen in de database. Dit komt waarschijnlijk grotendeels omdat deze factoren een hoge privacywaarde hebben (diploma ouders,ondersteuning van het gezin, gescheiden ouders, socio-economische status gezin). Het zal altijd gevoelig liggen om deze factoren op te nemen in de database, maar omdat ook deze factoren volgens de literatuurstudie een invloed hebben, loont het de moeite om hieraan in de toekomst ook aandacht te besteden.

Als laatste groep factoren die een relatie hebben met de slaagkans van een student, hebben we de factoren tijdens de studie. Hiervan zijn er slecht een klein deel opgenomen in de database (tijd investeren in de studies, leven op de campus). Van een deel andere factoren is het raar dat deze niet zijn opgenomen in de database (werken tijdens de studies, extra- curriculaire activiteiten, eerste jaar seminaries, gebruik academische support services). Andere zijn dan weer zeer moeilijk meetbaar (studeergedrag, faculteit – student contact, peer interactie, studententevredenheid over de instelling). Wederom kan het hier de moeite lonen om deze factoren in de toekomst proberen op te nemen in de database. Alles omvattend hebben we in deze database toch een heel nuttige factoren waarmee we aan de slag kunnen gaan om ons voorspelend model te gaan bouwen.

5.2. Selectie, voorverwerking & transformatie

De volgende stap is het selecteren van de effectieve data die we gaan gebruiken en tevens het verwijderen van de niet-gebruikte variabelen. In deze stap realiseren we ook al het preprocessen van de data. We zetten de data in de juiste vorm, dus het specificeren van de nominale data en numerieke data. Daarnaast definiëren we de missende waardes. Verder moeten we de data ook transformeren, dat wil zeggen dat we moeten zorgen dat de data in de juist vorm staan voor in te laten in het datamining programma.

Om onze analyses te kunnen uitvoeren, maken we van vier delen één groot bestand, met telkens één rij per student. We combineren de tabellen via student ID, verwijderen de voor onze analyses niet nuttige variabelen en creëren daarnaast zelf ook enkele samenvattende variabelen.

Verwijderd:

- Student ID (ID's hebben geen betekenis)
- Aantal vakken niet afgelegd (weinig nuttige info)
- Info over master diploma (we concentreren ons alleen op de bachelor)

Zelf gecreëerd:

- Diploma behaald en in welk jaar omgezet naar eigen variabele die bepaalt of het gehaald is of niet en zo ja, of dit volgens het modeltraject ging of niet.
- Van studiemeting hebben we een totaal gemaakt van alle vakken per student.
- Aantal herexamens
- Aantal onvoldoendes op het einde van het jaar

Zo krijgen we als input variabelen:

- Geslacht
- Nationaliteit
- Gemeente
- Land
- Op kot of niet
- Secundaire school
- Gemeente secundaire school
- Studierichting secundaire school
- Aantal uren wiskunde secundaire school
- Aantal uren natuurkunde secundaire school
- Aantal uren scheikunde secundaire school
- Aantal uren Frans secundaire school
- Aantal uren Engels secundaire school
- Aantal uren Duits secundaire school
- Beursstudent of niet
- Studievertraging opgelopen
- Generatiestudent of niet
- Totaal resultaat leestest
- Totaal resultaat op denktest
- Deelresultaat denktest verbaal
- Deelresultaat denktest numeriek
- Deelresultaat denktest grafisch
- Resultaat financieel boekhouden
- Resultaat geo-economie en economische geschiedenis
- Resultaat macro-economie
- Resultaat sociologie en demografie
- Resultaat wiskunde
- Studietijdmeting (totaal)

En als output variabelen:

- Resultaten op het einde van het jaar voor alle vakken
- Aantal herexamens
- Aantal onvoldoendes op het einde van het jaar
- Aantal studiepunten behaald
- Aantal studiepunten behaald na deliberatie
- Resultaat eerste jaar (%)
- Eindbeoordeling eerste jaar
- Eerste zit of tweede zit
- Bachelordiploma behaald

Om nu de verschillende datamining algoritmes te kunnen uitvoeren moeten we onze data nog inladen in het datamining programma dat we gaan gebruiken, namelijk Weka. Weka[1] is een populair en gratis datamining programma geschreven in Java en ontwikkeld door de universiteit van Waikato in Nieuw–Zeeland. We dienen onze data in te laden als .csv file en alle ontbrekende waarden te benoemen met een ? omdat Weka deze gebruikt voor het definiëren van ontbrekende waarden.

5.3. Datamining

Nu de data zijn ingeladen in Weka kunnen we aan de eigenlijke datamining analyses beginnen. Voor elk voorspellend model moeten we allereerst een variabele selecteren die we willen gaan voorspellen aan de hand van de input variabelen. Deze variabele noemen we de doelvariabele en is steeds een van outputvariabelen.

We gaan vijf verschillende analyses uitvoeren:

- Voorspellen of een student gaat slagen in zijn 1ste bachelorjaar

Hiervoor gebruiken we de output variabele: eindbeoordeling 1ste jaar
Deze is als volgt verdeeld 97 studenten zijn geslaagd/ 95 studenten zijn niet geslaagd.

- Voorspellen van het aantal herexamens in het 1ste bachelorjaar

Hiervoor gebruiken we de output variabele: resultaat eerste jaar (%)
Dit in een continue variabele met als minimum 0 en maximum 11, het gemiddelde is 4.3 en de standaarddeviatie 3.67.

- Voorspellen van het behaalde percentage in het 1ste bachelorjaar

Hiervoor gebruiken we de output variabele: aantal herexamens
Dit in een continue variabele met als minimum 18% en maximum 77%, het gemiddelde is 53 % en de standaarddeviatie 12.6

- Voorspellen van het wel of niet slagen op bepaalde vakken

Hiervoor gebruiken we de outputvariabele : resultaten op het einde van het jaar, we doen dit voor de vakken: wiskunde, financieel boekhouden en statistiek. Dit zijn continue variabele met volgende verdeling.

- Wiskunde : minimum 0 / maximum 18 / gemiddelde 9,7 / standaarddeviatie: 4,78
- Boekhouden: minimum 1 / maximum 17 / gemiddelde 9,7 / standaarddeviatie: 3,37
- Statistiek : minimum 0 / maximum 20 / gemiddelde 11,4 / standaarddeviatie: 5,2

- Voorspellen van het wel of niet behalen van het bachelordiploma

Hiervoor gebruiken we de output variabele: bachelordiploma behaald
Deze is als volgt verdeeld 124 studenten hebben hun diploma behaald / 68 niet.

Bij elke analyse doen we ook voorspellingen op drie momenten in de tijd. Voor het begin van de studies, na de resultaten van het eerste trimester en op het einde van het eerste jaar. Op deze manier bevat elke analyse telkens meer informatie en kunnen we ook zien of die informatie bijdraagt tot een beter model, zoals we verwachten dat het gaat zijn. Bij de analyses over het eerste bachelorjaar valt het laatste moment in de tijd natuurlijk weg, want het heeft weinig zin dingen te gaan voorspellen die, op het moment dat we de inputgegevens hebben, al gebeurd zijn.

[1] http://www.cs.waikato.ac.nz/ml/weka/

Hieronder een lijst van welke variabelen telkens inbegrepen zijn op het tijdsmoment van de analyse.

- Voor het begin van de studies:

    - Geslacht
    - Nationaliteit
    - Gemeente
    - Land
    - Op kot of niet
    - Secundaire school
    - Gemeente secundaire school
    - Studierichting secundaire school
    - Aantal uren wiskunde secundaire school
    - Aantal uren natuurkunde secundaire school
    - Aantal uren scheikunde secundaire school
    - Aantal uren Frans secundaire school
    - Aantal uren Engels secundaire school
    - Aantal uren Duits secundaire school
    - Beursstudent of niet
    - Studievertraging opgelopen
    - Generatiestudent of niet
    - Totaal resultaat leestest
    - Totaal resultaat op denktest
    - Deelresultaat denktest verbaal
    - Deelresultaat denktest numeriek
    - Deelresultaat denktest grafisch

De resultaten van de denk- en leestest nemen we al op bij deze tijdsperiode omdat ze reeds in de eerste week van het jaar worden afgenomen. Moest blijken dat ze echt een duidelijk verband hebben met de slaagkans, zouden deze testen eventueel al kunnen worden afgenomen bij inschrijving aan de instelling.

- Na eerste trimester, dus na Kerstvakantie, worden volgende variabelen nog bijgevoegd:

    - Resultaat financieel boekhouden
    - Resultaat geo-economie en economische geschiedenis
    - Resultaat macro-economie
    - Resultaat sociologie en demografie
    - Resultaat wiskunde
    - Studietijdmeting (totaal)

De resultaten van de studietijdmeting handelen over het eerste trimester.

- Op het einde van het jaar worden volgende variabelen toegevoegd:

    - Resultaten op het einde van het jaar voor alle vakken
    - Aantal herexamens
    - Aantal onvoldoendes op het einde van het jaar
    - Aantal studiepunten behaald
    - Aantal studiepunten behaald na deliberatie
    - Resultaat eerste jaar (%)
    - Eindbeoordeling eerste jaar
    - Eerste zit of tweede zit

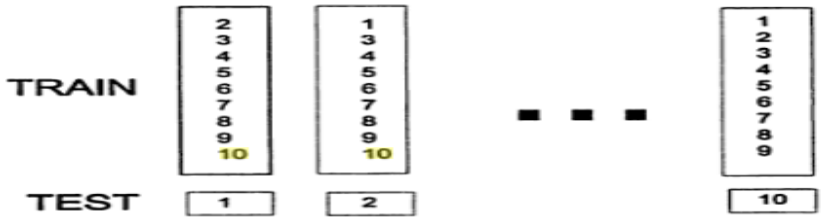
Voor het uitvoeren van deze analyses gebruiken we voor elk deel en elk moment in de tijd, de datamining technieken zoals beschreven in Hoofdstuk 3. Voor elke analyse zullen we dus telkens de volgende vijf datamining algoritmes toelichten en evalueren:

- OneR classifier
- Naive Bayes classifier
- Beslissingsboom
- k-Nearest neighbor algoritme
- Neurale netwerken

Als algoritme voor het creëren van beslissingsbomen kiezen we voor C4.5, veruit het meest gebruikte algoritme voor beslissingsbomen. Het k-Nearest neighbor algoritme is in Weka geïmplementeerd onder de naam Ibk. Voor het bouwen van neurale netwerken gebruiken we het MultilayerPerceptron zoals ook terug te vinden in Weka.

Oorspronkelijk was het de bedoeling om de vakken te clusteren zodat we groepen van vakken konden bekomen waarop studenten hetzelfde presteren. Zo zouden we dan kunnen voorspellen op welke groepen van vakken studenten het moeilijk zouden kunnen hebben. Helaas bleken onze clusterresultaten veel te instabiel en gaven ze afhankelijk van het gekozen cluster algoritme een heel verschillend resultaat, zie Bijlage 3. Daarom hebben we ervoor geopteerd om de vakken niet te groeperen maar ze gewoon individueel te behouden en voorspellingen te doen voor individuele vakken in de plaats van voor clusters van vakken.

De resultaten van de datamining algoritmes zullen telkens gemeten worden via een 10 fold cross validation. In deze techniek wordt de dataset random opgesplitst in tien gelijke delen. Het algoritme word dan geleerd en gebouwd op negen van deze tien delen en getest op het resterende deel. Dit proces word dan tien maal uitgevoerd zodat elk deel een keer het testdeel is geweest. Dit is algemeen de meest gebruikte techniek om datamining algoritmes te evalueren en geeft de minst vertekende resultaten.



Figuur 12: 10 fold cross validation

Dan leggen we nog enkele richtcijfers vast voor het percentage correct geclassificeerde cases. Aan de hand van deze percentages kunnen we de modellen gaan evalueren. Aan de hand van de eerder gevoerde onderzoeken, zie Hoofdstuk 4, kiezen we voor 70 % als breekpunt om te bepalen of het model goed presteert of niet. Indien het percentage hoger ligt dan 85% spreken we over een goed model.

Percentages correct geclassificeerde cases:

- Kleiner dan 70%: Slecht
- Groter of gelijk aan 70% en kleiner dan 85%: Redelijk
- Groter of gelijk aan 85%: Goed

5.4. Interpretatie en evaluatie

Voor elke deelanalyse, namelijk per moment in de tijd en per algoritme, hebben we telkens het percentage correct geclassificeerde cases berekend, samen met de confusion matrix. Deze matrix geeft in een tabel de echte waarde van een case weer ten opzichte van de waarde die het model gaat voorspellen voor deze case. Voornoemde gegevens zijn terug te vinden in Bijlage 4. Voor onze bespreking hebben we telkens per analyse alle percentages correct geclassificeerde cases samengevat in een tabel.

Voor het OneR algoritme vermelden we ook telkens de beslissingsvariabele waarop het algoritme zijn classificatie maakt. Deze variabele voorspelt het best de doelvariabele en kan dus in sommige gevallen zeer nuttige informatie geven. Dit geeft geen statische zekerheid, daar het datamining blijft. Maar het kan ons toch al een mooie aanzet geven voor het begrijpen van de factoren die een relatie hebben met de slaagkans van een student.

Voor het beslissingsboom algoritme geven we ook telkens de verkregen beslissingsboom mee. Men mag geen causale verbanden afleiden uit deze boom, daar dit niet de bedoeling is van datamining. Maar we kunnen wel telkens nagaan of de boom logisch is opgebouwd. Dit kan mogelijke pistes geven voor verder onderzoek indien het percentage correct geclassificeerde cases hoog genoeg is. We kunnen ook bekijken of hij niet teveel gefit is op de voorbeelddata, de aantallen tussen haakjes achter de eindepunten van de boom geven weer hoeveel cases er in dat eindepunt zitten en achter de backslash hoeveel van deze er incorrect zijn voorspeld. De kans op een te hevige fitting op de data is zeer groot doordat we maar een klein aantal voorbeeldcases hebben waarop het model gebouwd wordt, namelijk slechts 192. Dit kunnen we illustreren met volgend voorbeeld. Van de 192 voorbeeldcases waarop het model gebouwd is, zijn er 100 met twee uur Engels in het secundair, 87 met één uur Engels en 5 met drie uur Engels. Moesten van deze vijf studenten met drie uur Engels er nu toevallig vier niet slagen in hun eerste bachelorjaar dan zal het model drie uur Engels aanduiden als een factor voor het niet slagen. Terwijl er helemaal geen goede, logische basis is voor te besluiten dat het meer uren hebben van Engels een factor is voor het niet slagen. Dit vloeit voor uit het feit dat vijf studenten een te kleine groep is om een significant verband aan te tonen.

5.4.1. Voorspellen of een student gaat slagen in zijn eerst jaar bachelorjaar

Dit model zal proberen te voorspellen of een student slaagt (G) of niet slaagt (NG) voor zijn eerste bachelorjaar. We spreken van slagen als een student na de deliberatie alle 60 studiepunten heeft behaald.

Resultaten:

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
geslaagd of niet	vooraanvang	49	71	64	60	64
	naresult1	76	84	81	64	88

De modellen die alleen gebruik maken van de gegevens die bekend zijn aan het begin van het jaar presteren allemaal maar matig tot slecht. Enkel de naive bayes classifier valt net binnen het aanvaardbare, met een percentage correct geclassificeerde cases van 71% . Dit model zouden we dus kunnen gebruiken om een voorspelling te maken van het wel of niet slagen in het eerste bachelorjaar met als inputgegevens enkel de variabelen die bekend zijn aan het begin van de studies.

Als we de resultaten van het eerste trimester toevoegen aan de inputvariabelen zien we, zoals verwacht, in de meeste gevallen een zeer hoge stijging in het percentage correct geclassificeerde cases. Dit geeft in bijna elke geval een zeer acceptabel percentage met als uitschieter het neuraal netwerk met 88%. Op Ibk na, zou dus elk algoritme een goede voorspelling kunnen maken van het wel of niet slagen in het eerste bachelorjaar als het beschikt over de gegevens na het eerste trimester.

OneR beslissingsvariabele :

Model voor het begin van de studies → secundaire school

Secundaire school geeft ons geen extra nuttige informatie, deze variabele wordt door OneR alleen gekozen omdat deze veel mogelijke waardes heeft en OneR via deze variabele het best een zo specifiek mogelijke opdeling kan maken tussen de studenten.

Model na resultaten eerste trimester → macro economie

Dit is al een veel interessantere variabele. Dit wil zeggen dat de resultaten van macro economie van het eerste trimester de beste voorspeller zijn voor het slagen in het eerste bachelorjaar.

J48 beslissingsboom :

Boom voor het begin van de studies:

```
J48 pruned tree
------------------

SO_STUDIEVERTRAGING = nee
|   SOENGELS = 2tot2uur: G (127.29/43.76)
|   SOENGELS = 3ofmeeruur: NG (39.71/10.48)
SO_STUDIEVERTRAGING = ja
|   DENKTEST.NUMERIEK <= 15: NG (22.06/0.88)
|   DENKTEST.NUMERIEK > 15: G (2.94/0.82)
```

Deze boom zit eigenlijk niet logisch in elkaar, enkel de opdeling op basis van de denktest numeriek heeft een logische grond.

Boom na de resultaten van het 1ste trimester:

```
J48 pruned tree
------------------

Macro-economie <= 11
|   Financieel boekhouden <= 9
|   |   Wiskunde <= 9: NG (68.52/1.0)
|   |   Wiskunde > 9
|   |   |   GESLACHT = GESLACHT1: NG (5.13/1.0)
|   |   |   GESLACHT = GESLACHT2: G (2.13/0.13)
|   Financieel boekhouden > 9
|   |   LEESTEST.TOTAAL <= 79: G (5.77/0.06)
|   |   LEESTEST.TOTAAL > 79: NG (2.31/0.29)
Macro-economie > 11
|   Wiskunde <= 4
|   |   Financieel boekhouden <= 10: NG (8.18)
|   |   Financieel boekhouden > 10: G (2.02/0.02)
|   Wiskunde > 4: G (97.93/12.93)
```

Ook in deze boom zitten niet zoveel logische elementen. De opsplitsingen op resultaat op de vakken houden meestal nog wel steek, maar deze op leestest is onlogisch en duidelijk teveel gefit op de data.

5.4.2. Voorspellen van het aantal herexamens in het eerste bachelorjaar

Dit model zal voorspellen hoeveel herexamens een student zal hebben in zijn eerste jaar. Het aantal herexamens wordt in dit geval gedefinieerd als elk vak waarop de student een resultaat behaald minder dan een 10, maar er is ook nog deliberatie mogelijk. Dat wil dus zeggen dat sommige onvoldoendes niet hernomen moeten worden. Dit model voorspelt dus het aantal herexamens dat een student effectief moet maken. Studenten die gedelibereerd worden voor sommige resultaten zullen dus voorspeld worden als hebbende nul herexamens.

Omdat de meeste van onze datamining algoritmes niet overweg kunnen met continue numerieke data, hebben we zelf een opdeling gemaakt in categorieën voor het aantal herexamens. Deze opdeling ziet eruit als volgt:

- Geen herexamens
- Een Herexamen
- 2of3 herexamens
- 4 of 5 herexamens
- Meer als 5 herexamens

Resultaten:

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
#herex	vooraanvang	34	48	40	36	39
	naresult1	59	71	60	41	63

De modellen om het aantal herexamens van een student te voorspellen, presteren allemaal niet erg goed. Dit heeft deels te maken met het feit dat de doelvariabele die we gaan voorspellen 5 keuze mogelijkheden heeft, dit zorgt voor een daling in het percentage correct geclassificeerde cases. Van de modellen is er slechts een enkel dat een aanvaardbaar percentage correct geclassificeerde cases zal behalen en dat is hetgene van de naive bayes classifier met de resultaten van het eerste trimester. Dat model haalt een percentage van 71, wat nog altijd relatief laag is.

OneR beslissingsvariabele :

Model voor het begin van de studies → gemeente van afkomst

Gemeente van afkomst geeft ons geen extra nuttige informatie, deze variabele word door OneR alleen gekozen omdat deze veel mogelijke waardes heeft en OneR via deze variabele het best een zo specifiek mogelijke opdeling kan maken tussen de studenten.

Model na resultaten eerste trimester → wiskunde

Dit is al een veel interessantere variabele, wat wil zeggen dat de resultaten op wiskunde van het eerste trimester de beste voorspeller zijn voor het aantal herexamens in het eerste bachelorjaar

J48 beslissingsboom :

Boom voor het begin van de studies:

```
SO_STUDIEVERTRAGING = nee: 0.0 (167.0/104.0)
SO_STUDIEVERTRAGING = ja: meeralsvijf (25.0/5.0)
```

Dit is een zeer korte boom, maar de opsplitsing op eerder op gelopen studievertraging lijkt wel steek te houden.

Boom na de resultaten van het 1ste trimester:

```
Macro-economie <= 11
|   Financieel boekhouden <= 9
|   |   Wiskunde <= 9
|   |   |   BEURSSTUDENT = BEURSSTUDENT2
|   |   |   |   Geo-economie en economische geschiedenis <= 12: meeralsvijf (48.44/2.0)
|   |   |   |   Geo-economie en economische geschiedenis > 12
|   |   |   |   |   DENKTEST.GRAFISCH <= 13: meeralsvijf (2.43)
|   |   |   |   |   DENKTEST.GRAFISCH > 13: meeralsdrie (3.65/0.65)
|   |   |   BEURSSTUDENT = BEURSSTUDENT1
|   |   |   |   Macro-economie <= 9: meeralsvijf (8.0/1.0)
|   |   |   |   Macro-economie > 9: meeralsdrie (6.0/1.0)
|   |   Wiskunde > 9
|   |   |   GESLACHT = GESLACHT1: meeralsdrie (5.13/0.13)
|   |   |   GESLACHT = GESLACHT2: tweedrie (2.13/1.13)
|   Financieel boekhouden > 9
|   |   Macro-economie <= 10: meeralsdrie (4.04/1.04)
|   |   Macro-economie > 10
|   |   |   DENKTEST.TOTAAL <= 38: een (2.02/0.02)
|   |   |   DENKTEST.TOTAAL > 38: 0.0 (2.02/0.02)
Macro-economie > 11
|   Wiskunde <= 10: tweedrie (47.94/30.94)
|   Wiskunde > 10
|   |   Geo-economie en economische geschiedenis <= 9
|   |   |   LEESTEST.TOTAAL <= 71: 0.0 (2.25/0.25)
|   |   |   LEESTEST.TOTAAL > 71: tweedrie (3.37/1.37)
|   |   Geo-economie en economische geschiedenis > 9: 0.0 (54.57/4.57)
```

Deze boom maakt over het algemeen redelijk logische opsplitsingen en is nuttig om te bekijken voor verdere analyses.

Omdat de percentages correct geclassificeerde cases niet goed zijn, opteren we ervoor het nog eens te proberen met een doelvariabele met minder opties. Zo krijgen we hopelijk een model met een beter percentage correct geclassificeerde cases.

We kiezen voor een opdeling in de volgende categorieën:

- 0 Herexamens
- 1 tot en met 5 herexamens
- Meer als 5 herexamens

De achterliggende reden voor deze opsplitsing is dat we één tot en met vijf herexamens nog haalbaar achten om op te halen voor een student. Vanaf meer dan vijf herexamens moet de student toch wel in vraag stellen of de richting wel geschikt is voor hem.

<u>Resultaten:</u>

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
#herex2	vooraanvang	43	51	41	46	46
	naresult1	71	82	71	51	71

Zoals verwacht geven deze modellen al veel betere resultaten. Op het moment van de aanvang van de studies is er echter nog geen enkel model dat goede resultaten behaalt. Met de toevoeging

van de resultaten van het eerste trimester krijgen we wel enkele betere modellen, met als uitschieter de naive bayes met een percentage correctly classified instances van 82.

OneR beslissingsvariabele:

Model voor het begin van de studies → gemeente van afkomst
Model na resultaten eerste trimester → wiskunde

Dit zijn dezelfde resultaten als we reeds bekwamen bij de vorige categorieopdeling voor het voorspellen van het aantal herexamens.

J48 beslissingsboom:

Boom voor het begin van de studies:

```
SO_STUDIEVERTRAGING = nee
|   GEMEENTE = GEMEENTE172
|   |   GESLACHT = GESLACHT1: 1totmet5 (2.0)
|   |   GESLACHT = GESLACHT2: meeralsvijf (4.0/1.0)
|   GEMEENTE = GEMEENTE37
|   |   SO_RICHTING = SO_RICHTING37: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING28: 1totmet5 (1.0)
|   |   SO_RICHTING = SO_RICHTING34: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING23: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING24: meeralsvijf (3.0/1.0)
|   |   SO_RICHTING = SO_RICHTING9: 1totmet5 (5.0/2.0)
|   |   SO_RICHTING = SO_RICHTING7: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING29: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING30: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING22: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING12: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING32: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING4: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING1: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING16: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING3: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING27: 1totmet5 (0.0)
|   GEMEENTE = GEMEENTE99: 1totmet5 (3.0/2.0)
|   GEMEENTE = GEMEENTE34: meeralsvijf (1.0)
|   GEMEENTE = GEMEENTE30: meeralsvijf (1.0)
|   GEMEENTE = GEMEENTE159: 0.0 (1.0)
|   GEMEENTE = GEMEENTE58
|   |   SOSCHEI = 2ofmeeruur: 0.0 (2.0)
```

Deze boom loopt nog een heel eindje door maar de resultaten hiervan zijn niet echt bruikbaar aangezien hij opsplitisingen maakt op gemeente. Sommige gemeentes bevatten immers veel te weinig studenten om te kunnen spreken van duidelijke verbanden.

Boom na de resultaten van het eerste trimester:

```
Financieel boekhouden <= 9
|   Sociologie en demografie <= 9: meeralsvijf (29.31)
|   Sociologie en demografie > 9
|   |   Macro-economie <= 11
|   |   |   Wiskunde <= 9
|   |   |   |   BEURSSTUDENT = BEURSSTUDENT2: meeralsvijf (30.43/5.0)
|   |   |   |   BEURSSTUDENT = BEURSSTUDENT1
|   |   |   |   |   Macro-economie <= 9: meeralsvijf (5.0/1.0)
|   |   |   |   |   Macro-economie > 9: 1totmet5 (6.0)
|   |   |   Wiskunde > 9: 1totmet5 (7.07/0.07)
|   |   Macro-economie > 11: 1totmet5 (44.46/21.46)
Financieel boekhouden > 9
|   Wiskunde <= 10
|   |   DENKTEST.NUMERIEK <= 13: 1totmet5 (16.41/2.17)
|   |   DENKTEST.NUMERIEK > 13: 0.0 (6.15/1.15)
|   Wiskunde > 10: 0.0 (47.17/4.17)
```

Dit is een boom met logische opsplitsingen die zeker gebruikt kan worden voor verder onderzoek.

5.4.3. Voorspellen van behaalde percentage in het eerste bachelorjaar

Dit model zal voorspellen welk gewogen percentage een student zal halen in zijn eerst jaar.

Net zoals bij het aantal herexamens zetten we het continue numeriek percentage om naar elkaar uitsluitende categorieën. Zo krijgen we volgende categorieën in de doelvariabele die we willen voorspellen:

- Percentage kleiner dan 35%
- Percentage groter dan of gelijk aan 35% & kleiner dan 50%
- Percentage groter dan of gelijk aan 50% & kleiner dan 65%
- Percentage groter dan of gelijk aan 65%

We kiezen specifiek voor deze opdeling om volgende redenen. Indien het gewogen percentage lager ligt dan 35%, moet de student zicht zwaar de vraag stellen of de richting wel iets voor hem is. Als het percentage tussen 35% en 50% ligt is het nog aanvaardbaar om verder mee te gaan. Een percentage boven de 50% duidt erop dat de student kan slagen in de richting en een percentage hoger dan 65% duidt op het zeer goed presteren.

Resultaten :

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
Percentage	vooraanvang	42	43	47	36	46
	naresult1	51	71	58	41	55

Bijna geen enkel model slaagt erin om een goede voorspelling te maken voor het percentage. Enkel de naive bayes met toevoeging van de resultaten van het eerste trimester haalt nipt een acceptabel percentage van 71%.

OneR beslissingsregelvariabele:

Model voor het begin van de studies → gemeente van afkomst

Wederom geef de OneR beslissingsvariabele bij het model voor het begin van de studies ons geen nuttige informatie.

Model na resultaten eerste trimester → macro economie

Dit geeft ons weer wel nuttige informatie, dit wil zeggen dat de resultaten op macro-economie van het eerste trimester de beste voorspeller zijn voor het percentage in het eerste bachelorjaar. Weliswaar is het percentage correct geclassificeerde cases van OneR zeer laag, dus mogen we besluiten dat het geen goede voorspeller is.

J48 beslissingsboom :

Boom voor het begin van de studies:

```
: kleiner65 (192.0/100.0)
```

Deze boom steekt alles in één klasse en is dus geen basis voor verder onderzoek.

Boom na de resultaten van het 1ste trimester:

```
Wiskunde <= 3: kleiner50 (47.48/22.24)
Wiskunde > 3
|   Macro-economie <= 12
|   |   SO_STUDIEVERTRAGING = nee: kleiner65 (51.56/14.81)
|   |   SO_STUDIEVERTRAGING = ja: kleiner50 (6.05/2.05)
|   Macro-economie > 12
|   |   Sociologie en demografie <= 14: kleiner65 (67.71/20.71)
|   |   Sociologie en demografie > 14: groter65 (19.2/4.0)
```

Dit is een korte boom met logische opsplitsingen. Maar toch is het behaalde percentage correct geclassificeerde cases veel te laag om deze als basis te nemen voor verder onderzoek.

5.4.4. Voorspellen van slagen op bepaalde vakken in het eerste bachelorjaar

In dit deel gaan we proberen de slaagkans van enkele vakken uit het eerste bachelorjaar te voorspellen. We proberen dit niet voor alle vakken, maar we selecteren de 3 vakken waarvan de resultaten het best het wel of niet slagen in het totale eerste jaar bepalen. Dit zijn in volgorde: wiskunde, financieel boekhouden en statistiek. Deze 3 vakken leiden we af door het OneR algoritme toe te passen om de slaagkans in het 1ste bachelorjaar te bepalen aan de hand van alle inputvariabelen.

5.4.4.1. Wiskunde

Resultaten :

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
Wiskunde	vooraanvang	57	62	70	55	59
	naresult1	82	79	81	61	77

Bij de modellen met als inputvariabelen de gegevens aan het begin van het jaar, hebben we niet echt bruikbare modellen. Enkel de beslissingsboom haalt een iets of wat aanvaardbaar percentage correct geclassificeerde cases van 70%. De modellen met de resultaten van het eerste trimester bijgevoegd, presteren op IBk na allemaal redelijk goed. Omdat One R zijn regel baseert op de resultaten op wiskunde van het eerste trimester kunnen we stellen dat deze resultaten een heel goede voorspeller zijn voor het algemene resultaat van wiskunde, wat ook redelijk logisch is.

OneR beslissingsvariabele :

Model voor het begin van de studies → secundaire school

Ook hier geef de OneR beslissingsvariabele bij het model voor het begin van de studies ons geen nuttige informatie.

Model na resultaten eerste trimester → wiskunde

Dit wil dus zeggen dat de resultaten van het eerste trimester wiskunde de beste voorspeller zijn van het eindresultaat op wiskunde. Dit is een logische regel. Het percentage van 82% correct geclassificeerde cases is daarom ook redelijk hoog.

J48 beslissingsboom :

Boom voor het begin van de studies:

```
SO_STUDIEVERTRAGING = nee
|   SOENGELS = 2tot2uur: geslaagd (122.57/29.0)
|   SOENGELS = 3ofmeeruur: kleinder8 (33.43/16.43)
SO_STUDIEVERTRAGING = ja: kleinder8 (19.0/5.0)
```

In deze boom is de opsplitsing op basis van eerder opgelopen studievertraging logisch, maar wederom is de opsplitsing op Engels te zeer gefit op de data en heeft deze een zeer grote foutmarge.

Boom na de resultaten van het 1ste trimester:

```
Wiskunde <= 7
|   Wiskunde <= 3
|   |   Sociologie en demografie <= 12: kleinder8 (32.0/1.0)
|   |   Sociologie en demografie > 12
|   |   |   Macro-economie <= 10: kleinder10 (2.21/0.21)
|   |   |   Macro-economie > 10: kleinder8 (2.0)
|   Wiskunde > 3: geslaagd (48.28/29.0)
Wiskunde > 7: geslaagd (90.52/3.0)
```

De basisopsplitsing op basis van de resultaten op wiskunde van het eerste trimester is zeer logisch, maar verderop is er een onlogische opsplitsing op de resultaten van macro-economie. Deze laatste opsplitsing is duidelijk teveel gefit op de data.

5.4.4.2 Financieel boekhouden

Resultaten :

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
Boekhouden	vooraanvang	46	61	65	54	65
	naresult1	82	78	81	55	73

Tussen de modellen met als inputvariabelen de gegevens bij aanvang van de studies, zit geen enkel voortreffelijk en dus zeker bruikbaar model. Voegen we echter de resultaten van het eerste trimester toe aan de inputvariabelen krijgen we enkele goede modellen. Naive Bayes, J48 en zelfs OneR halen hier een goed percentage correctly classified instances.

OneR beslissingsvariabele :

Model voor het begin van de studies → gemeente van afkomst
Ook hier geef de OneR beslissingsvariabele bij het model voor het begin van de studies ons geen nuttige informatie.

Model na resultaten eerste trimester → boekhouden

Dit wil dus zeggen dat de resultaten van het eerste trimester boekhouden de beste voorspeller zijn van het eindresultaat op boekhouden. Dit is een logische regel. Het percentage van 82% correct geclassificeerde cases is ook redelijk hoog.

J48 beslissingsboom:

Boom voor het begin van de studies:

```
NATIONALITEIT = NATIONALITEIT3
|   SOFRANS = 4ofmeeruur
|   |   SO_STUDIEVERTRAGING = nee: geslaagd (36.58/17.29)
|   |   SO_STUDIEVERTRAGING = ja: kleinder8 (9.0/4.0)
|   SOFRANS = 0tot3uur: geslaagd (112.42/21.71)
NATIONALITEIT = NATIONALITEIT16: kleinder8 (1.0)
NATIONALITEIT = NATIONALITEIT14: kleinder8 (4.0/1.0)
NATIONALITEIT = NATIONALITEIT4: geslaagd (0.0)
NATIONALITEIT = NATIONALITEIT7: kleinder8 (1.0)
NATIONALITEIT = NATIONALITEIT9: kleinder8 (1.0)
```

Deze boom bied weinig nuttigs voor verder onderzoek. Een opsplitsing op nationaliteit heeft weinig nut daar de meeste nationaliteiten te weinig gegevens bevatten.

Boom na de resultaten van het eerste trimester:

```
Financieel boekhouden <= 7
|   Wiskunde <= 5
|   |   BEURSSTUDENT = BEURSSTUDENT2: kleinder8 (24.57)
|   |   BEURSSTUDENT = BEURSSTUDENT1
|   |   |   Sociologie en demografie <= 10: kleinder8 (4.0)
|   |   |   Sociologie en demografie > 10: geslaagd (2.0)
|   Wiskunde > 5
|   |   Macro-economie <= 12: kleinder8 (14.43/9.0)
|   |   Macro-economie > 12: kleinder10 (9.0/4.0)
Financieel boekhouden > 7
|   Wiskunde <= 4
|   |   OP_KOT = OP_KOT1: geslaagd (4.0)
|   |   OP_KOT = OP_KOT2
|   |   |   SO_RICHTING = SO_RICHTING37: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING28: geslaagd (0.14)
|   |   |   SO_RICHTING = SO_RICHTING34: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING23: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING24: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING9: kleinder8 (5.0/1.0)
|   |   |   SO_RICHTING = SO_RICHTING7: geslaagd (5.0/1.0)
|   |   |   SO_RICHTING = SO_RICHTING29: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING30: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING22: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING12: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING32: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING4: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING1: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING16: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING3: kleinder8 (1.0)
|   |   |   SO_RICHTING = SO_RICHTING27: geslaagd (0.0)
|   Wiskunde > 4: geslaagd (95.86/3.0)
```

Dit is een logische boom die zeker verder onderzocht kan worden om er conclusies uit te trekken. De opsplitsing op richting secundair onderwijs zou zeker logisch kunnen zijn, maar in dit onderzoek bevat elke richting te weinig studenten om hier echt verbanden te kunnen zien.

5.4.4.3 Statistiek

Resultaten :

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
Statistiek	vooraanvang	71	72	77	69	71
	naresult1	82	80	82	71	81

We zien dat de meeste modellen aan het begin van de studies nog aanvaardbaar presteren, J48 haalt zelfs een mooi percentage correctly classified instances. Zoals te verwachten zijn de modellen met de resultaten van het eerste trimester nog beter met, op Ibk na, allemaal een percentage van in de 80.

OneR beslissingsvariabele :

Model voor het begin van de studies → gemeente van afkomst

Ook hier geef de OneR beslissingsvariabele bij het model voor het begin van de studies ons geen nuttige informatie.

Model na resultaten eerste trimester → wiskunde

Dit wil dus zeggen dat de resultaten van het eerste trimester wiskunde de beste voorspeller zijn van het eindresultaat op statistiek. In tegenstelling tot de twee vorige gevallen is het niet statistiek zelf die de beste voorspeller is. Dit komt omdat statistiek geen resultaten heeft in het eerste trimester. Met 82% correct geclassificeerde cases is dit wederom een zeer goede voorspeller. We kunnen dus zeggen dat de resultaten op wiskunde een goede voorspeller zijn van de resultaten op statistiek.

J48 beslissingsboom :

Boom voor het begin van de studies:

```
SO_STUDIEVERTRAGING = nee: geslaagd (152.0/29.0)
SO_STUDIEVERTRAGING = ja: kleinder8 (16.0/7.0)
```

Een zeer kleine boom met slechts een enkele regel. Doch is deze regel de moeite waard om verder te onderzoeken, daar deze logisch is en ook een goed percentage correct geclassificeerde cases bevat.

Boom na de resultaten van het 1ste trimester:

```
Macro-economie <= 8: kleinder8 (25.0/7.0)
Macro-economie > 8
|   Wiskunde <= 3
|   |   BEURSSTUDENT = BEURSSTUDENT2: kleinder8 (9.2/2.1)
|   |   BEURSSTUDENT = BEURSSTUDENT1: geslaagd (5.0/1.0)
|   Wiskunde > 3
|   |   Financieel boekhouden <= 7
|   |   |   SOFRANS = 4ofmeeruur
|   |   |   |   Geo-economie en economische geschiedenis <= 11: kleinder8 (4.9/1.0)
|   |   |   |   Geo-economie en economische geschiedenis > 11: kleinder10 (3.33/1.33)
|   |   |   SOFRANS = 0tot3uur: geslaagd (16.67/3.0)
|   |   Financieel boekhouden > 7: geslaagd (103.9/1.0)
```

Een boom met een deel begrijpbare logische opsplitsingen. Zeker het opsplitsen op aantal uren Frans in het secundair is de moeite waard om verder te onderzoeken.

5.4.5. Voorspellen of een student zijn bachelordiploma zal behalen.

Dit model zal voorspellen of een student zal slagen op zijn volledige bachelor.

In een eerste model bekijken we enkel het slagen of niet slagen. In een tweede model gaan we dan ook nog bekijken hoeveel jaren een student erover doet om zijn bachelordiploma te behalen.

Resultaten :

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
Bachelor 1	vooraanvang	66	73	70	63	71
	naresult1	75	83	80	65	80
	eindejaar	89	88	89	85	88

We zien dat na de resultaten van het eerste jaar elk model zeer goed presteert. We kunnen dus stellen dat na het eerste jaar men zeer goed kan gaan voorspellen welke studenten uiteindelijk hun bachelordiploma zullen behalen. Ook de modellen met enkel de gegevens na het eerste trimester presteren goed. Het is dus mogelijk om na het 1ste trimester al goed te voorspellen of een student uiteindelijk zijn bachelordiploma zal behalen. De modellen, die opgebouwd zijn aan de hand van alleen de gegevens bekend voor aanvang van de studies, presteren maar matig. Maar ze geven toch al een goede indicatie voor het wel of niet behalen van het bacherlordiploma.

OneR beslissingsvariabele :

Model voor het begin van de studies → secundaire school

Wederom geef de OneR beslissingsvariabele bij het model voor het begin van de studies ons geen nuttige informatie.

Model na resultaten eerste trimester → financieel boekhouden

Dit wil dus zeggen dat de resultaten van het eerste trimester financieel boekhouden de beste voorspeller zijn voor het wel of niet behalen van het bachelordiploma.

Model op het einde van het eerste bachelorjaar→: # onvoldoendes eerste bachelorjaar.

Het aantal onvoldoendes in het eerste bachelorjaar is de beste voorspeller voor het wel of niet slagen op de gehele bacheloropleiding.

J48 beslissingsboom :

Boom voor het begin van de studies:

```
J48 pruned tree
------------------

SO_STUDIEVERTRAGING = nee
|   BEURSSTUDENT = BEURSSTUDENT2
|   |   SOWISK = 8uur: wel (18.0/3.0)
|   |   SOWISK = 6uur: wel (85.0/25.0)
|   |   SOWISK = 4/5uur
|   |   |   OP_KOT = OP_KOT1: niet behaald (4.0)
|   |   |   OP_KOT = OP_KOT2: wel (6.0/2.0)
|   |   SOWISK = 0tot3uur: niet behaald (15.0/2.0)
|   BEURSSTUDENT = BEURSSTUDENT1: wel (39.0/2.0)
SO_STUDIEVERTRAGING = ja: niet behaald (25.0/6.0)
```

Deze logisch opgebouwde boom bied veel stof voor verder onderzoek. Eerder opgelopen studievertraging speelt ook hier weer een rol, samen met het krijgen van een beurs, het aantal uren wiskunde en het op kot zitten of niet. Deze factoren zijn in overeenstemming met de besluiten uit de literatuurstudie en zijn het dus zeker waard om verder onderzoek op uit te voeren.

Boom na de resultaten van het 1ste trimester:

```
J48 pruned tree
------------------

Financieel boekhouden <= 7
|   Macro-economie <= 11
|   |   BEURSSTUDENT = BEURSSTUDENT2
|   |   |   Wiskunde <= 8: niet behaald (45.37/1.0)
|   |   |   Wiskunde > 8: wel (5.28/1.28)
|   |   BEURSSTUDENT = BEURSSTUDENT1
|   |   |   SO_STUDIEVERTRAGING = nee: wel (7.0/2.0)
|   |   |   SO_STUDIEVERTRAGING = ja: niet behaald (4.0)
|   Macro-economie > 11
|   |   GESLACHT = GESLACHT1: wel (13.08/2.08)
|   |   GESLACHT = GESLACHT2: niet behaald (3.08/1.0)
Financieel boekhouden > 7
|   GENERATIESTUDENT = nee
|   |   BEURSSTUDENT = BEURSSTUDENT2
|   |   |   SONAT = 3ofmeeruur: niet behaald (0.0)
|   |   |   SONAT = 2uur: wel (4.0/1.0)
|   |   |   SONAT = 0/1uur: niet behaald (4.19)
|   |   BEURSSTUDENT = BEURSSTUDENT1: wel (4.0)
|   GENERATIESTUDENT = ja: wel (102.0/7.0)
```

Deze boom haalt een goed percentage correct geclassificeerde cases, ook de keuzeknooppunten lijken logisch. Dus het is zeker de moeite om hier verder onderzoek naar te doen. De opsplitsing op aantal uren natuurkunde (sonat) is wel duidelijk teveel gefit op de data, want hier zit geen logisch verband in.

Boom na het volledige eerste bachelorjaar:

```
J48 pruned tree
------------------

onvoldoendes <= 4
|   Financieel boekhouden  EIND <= 7
|   |   Talen  EIND <= 10: wel (3.0)
|   |   Talen  EIND > 10
|   |   |   STPTN_BEHAALD = 54<= X < 60: niet behaald (0.0)
|   |   |   STPTN_BEHAALD = 60.0: niet behaald (0.0)
|   |   |   STPTN_BEHAALD = <30: niet behaald (0.0)
|   |   |   STPTN_BEHAALD = 30<= X < 48: niet behaald (5.25/0.17)
|   |   |   STPTN_BEHAALD = 48<= X < 54: wel (3.0/1.0)
|   Financieel boekhouden  EIND > 7: wel (123.75/6.92)
onvoldoendes > 4
|   BEURSSTUDENT = BEURSSTUDENT2: niet behaald (51.0)
|   BEURSSTUDENT = BEURSSTUDENT1
|   |   Financieel boekhouden <= 5: niet behaald (4.0)
|   |   Financieel boekhouden > 5: wel (2.0)
```

Deze boom voorspelt zeer goed het slagen op de volledige bacheloropleiding. Toch maakt hij enkele zeer onlogische keuzes, meer bepaald in de opsplitsing bij studiepunten behaald. Deze keuzes zijn duidelijk teveel gefit op de data.

Nu splitsen we de geslaagde studenten nog eens op in twee categorieën, namelijk zij die hun bachelordiploma behalen volgens het modeltraject en zij die er langer dan drie jaar over doen.

Resultaten :

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
Bachelor 2	vooraanvang	42	55	55	44	54
	naresult1	72	66	70	45	63
	eindejaar	82	80	83	74	76

Logischerwijs zijn de percentages correct geclassificeerde cases lager dan bij de modellen met maar twee te voorspellen targetvariabelen. Maar toch zien we hier ook enkele goede resultaten, zeker de modellen met alle gegevens over het eerste jaar presteren nog goed. En ook de OneR en beslissingsboom met als input de gegevens van het eerste trimester zijn zeker ook nog bruikbaar om te gaan voorspellen hoe een student zijn bachelor zal volbrengen.

OneR beslissingsvariabele :

Deze regels blijven hetzelfde als bij de vorige opdeling in categorieën.

J48 beslissingsboom :

Boom voor het begin van de studies:

```
J48 pruned tree
------------------

SO_STUDIEVERTRAGING = nee
|   BEURSSTUDENT = BEURSSTUDENT2
|   |   SODUITS = 0uur: normaal (42.32/22.99)
|   |   SODUITS = 1uur: normaal (38.19/13.9)
|   |   SODUITS = 2uur: normaal (23.74/9.56)
|   |   SODUITS = 3ofmeeruur: niet behaald (23.74/8.37)
|   BEURSSTUDENT = BEURSSTUDENT1: normaal (39.0/18.0)
SO_STUDIEVERTRAGING = ja
|   BEURSSTUDENT = BEURSSTUDENT2: niet behaald (17.0/2.0)
|   BEURSSTUDENT = BEURSSTUDENT1
|   |   SOWISK = 8uur: meer (0.0)
|   |   SOWISK = 6uur: niet behaald (5.0/1.0)
|   |   SOWISK = 4/5uur: meer (1.0)
|   |   SOWISK = 0tot3uur: meer (2.0)
```

Wederom maakt deze boom een opsplitsing op eerder opgelopen studievertraging, wat al meermaals is voorgekomen. De opsplitsing op uren Duits in het secundair onderwijs lijkt dan weer minder logisch te zijn, net zoals deze op uren wiskunde. De opsplitsing op basis van beursstudent of niet, is ook zeker de moeite waard om verder te onderzoeken.

Boom na de resultaten van het eerste trimester:

```
J48 pruned tree
------------------

Financieel boekhouden <= 7
|   Wiskunde <= 7
|   |   BEURSSTUDENT = BEURSSTUDENT2: niet behaald (49.94/3.0)
|   |   BEURSSTUDENT = BEURSSTUDENT1
|   |   |   SO_STUDIEVERTRAGING = nee: meer (7.0/1.0)
|   |   |   SO_STUDIEVERTRAGING = ja: niet behaald (3.0)
|   Wiskunde > 7: meer (17.88/9.88)
Financieel boekhouden > 7
|   GENERATIESTUDENT = nee
|   |   BEURSSTUDENT = BEURSSTUDENT2
|   |   |   SONAT = 3ofmeeruur: niet behaald (0.0)
|   |   |   SONAT = 2uur: normaal (4.0/1.0)
|   |   |   SONAT = 0/1uur: niet behaald (4.19)
|   |   BEURSSTUDENT = BEURSSTUDENT1: meer (4.0/1.0)
|   GENERATIESTUDENT = ja: normaal (102.0/25.0)
```

Deze boom houdt weinig steek qua logische opbouw, de opsplitsing op basis van de resultaten van het eerst trimester is onlogisch en ook deze op basis van aantal uren natuurkunde in het secundair lijkt teveel gefit op de data.

Boom na het volledige eerste bachelorjaar:

```
onvoldoendes <= 2
|   Beginselen van het recht. Privaat en publiek recht. Handelsrecht EIND <= 7: meer (6.0/1.0)
|   Beginselen van het recht. Privaat en publiek recht. Handelsrecht EIND > 7
|   |   Wiskunde  EIND <= 6: meer (4.0)
|   |   Wiskunde  EIND > 6
|   |   |   Talen  EIND <= 10: meer (6.0/1.0)
|   |   |   Talen  EIND > 10: normaal (98.0/13.0)
onvoldoendes > 2
|   onvoldoendes <= 4
|   |   Geo-economie en economische geschiedenis  EIND <= 11: meer (7.0)
|   |   Geo-economie en economische geschiedenis  EIND > 11
|   |   |   Macro-economie  EIND <= 11
|   |   |   |   LEESTEST.TOTAAL <= 84: niet behaald (7.0)
|   |   |   |   LEESTEST.TOTAAL > 84: meer (3.0/1.0)
|   |   |   Macro-economie  EIND > 11: meer (4.0)
|   onvoldoendes > 4
|   |   BEURSSTUDENT = BEURSSTUDENT2: niet behaald (51.0)
|   |   BEURSSTUDENT = BEURSSTUDENT1
|   |   |   Financieel boekhouden <= 5: niet behaald (4.0)
|   |   |   Financieel boekhouden > 5: meer (2.0)
```

Deze boom maakt logische opsplitsingen, maar zit wel redelijk ingewikkeld in elkaar. Hij maakt daarnaast op zijn eindpunten vaak beslissingen gebaseerd op weinig cases. Dit duidt op een hevige fitting op de data.

Algemene opmerking:

We bekijken de OneR beslissingsvariabelen nader. We kunnen dan zien dat het algoritme hier bij de modellen voor aanvang van de studies telkens een variabele neemt die gewoon het beste resultaat behaalt omdat deze de beste opsplitsingmogelijkheden heeft. Daarom gaan we eens eenmalig proberen telkens de onnuttige variabelen te verwijderen tot we een logischere resultaat krijgen. We doen dit bij het model 5.4.5., dat voorspelt of een student zijn bachelordiploma zal behalen. Na het verwijderen van de variabelen secundaire school en gemeente maakt de OneR een opsplitsing op denktest numeriek wat een opsplitsing is waarop wel voortgebouwd kan worden in verder onderzoek.

Hoofdstuk 6 : Besluit

6.1. Algemeen besluit

We beginnen ons besluit met een terugblik op de centrale onderzoeksvraag die we ons op het begin stelden.

"Kunnen we, met behulp van het knowledge discovery process, voorspellingen gaan doen over de slaagkans van een student aan de UHasselt?"

Om deze vraag te kunnen beantwoorden, bekijken we onderstaande tabel met de percentages correct geclassificeerde cases (%<70 zijn doorstreept en aangeduid en het hoogste percentage per rij is telkens groter weergegeven).

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
geslaagd of niet	vooraanvang	~~49~~	71	~~64~~	~~60~~	~~64~~
	naresult1	76	84	81	~~64~~	88
#herex	~~vooraanvang~~	~~34~~	~~48~~	~~40~~	~~36~~	~~39~~
	naresult1	~~59~~	71	~~60~~	~~41~~	~~63~~
#herex2	~~vooraanvang~~	~~43~~	~~51~~	~~41~~	~~46~~	~~46~~
	naresult1	71	82	71	~~51~~	71
Percentage	~~vooraanvang~~	~~42~~	~~43~~	~~47~~	~~36~~	~~46~~
	naresult1	~~51~~	71	~~58~~	~~41~~	~~55~~
Wiskunde	vooraanvang	~~57~~	~~62~~	70	~~55~~	~~59~~
	naresult1	82	79	81	~~61~~	77
Boekhouden	~~vooraanvang~~	~~46~~	~~61~~	~~65~~	~~54~~	~~65~~
	naresult1	82	78	81	~~55~~	73
Statistiek	vooraanvang	71	72	77	~~69~~	71
	naresult1	82	80	82	71	81
Bachelor 1	vooraanvang	~~66~~	73	70	~~63~~	71
	naresult1	75	83	80	~~65~~	80
	eindejaar	89	88	89	85	88
Bachelor 2	~~vooraanvang~~	~~42~~	~~55~~	~~55~~	~~44~~	~~54~~
	naresult1	72	~~66~~	70	~~45~~	~~63~~
	eindejaar	82	80	83	74	76

Wie zien dat we met enkele basisgegevens van een student en zijn resultaten op de denk- en leestest al met een correctheid van minimum 70% kunnen voorspellen of een student:

- Slaagt in zijn eerste jaar
- Zijn bachelordiploma zal behalen
- Welk resultaat hij zal behalen op statistiek

Maar we kunnen niet met een voldoende groot percentage voorspellen hoeveel herexamens een student zal behalen of in welke categorie van percentage hij zal vallen. Ook de resultaten voor wiskunde en boekhouden zijn niet te voorspellen met een zekerheid groter dan 70% .

Verder stellen we vast dat hoe meer inputgegevens, zoals aanvang studies/na resultaten 1ste trimester/einde van het jaar, we toevoegen, hoe beter de resultaten zullen zijn. Dit is een logisch gegeven, maar het toont aan hoe belangrijk een evaluatie van een student na het eerste trimester is. Na de toevoeging van de resultaten van het eerste trimester zien we in elk van de analyses een grote toename in het percentage correct geclassificeerde cases.

Als we dan een blik werpen op de analyses die plaatsvinden na de resultaten van het eerste trimester, zien we dat we voor elke doelvariabele een model hebben dat beter presteert dan 70%. We kunnen dus na het eerste trimester voorspellen, met een zekerheid groter als 70%, of een student:

- Slaagt in zijn eerste jaar
- Hoeveel herexamens hij zal behalen
- In welke categorie van totaal percentage hij zal vallen
- Welke resultaat hij zal gaan behalen op wiskunde / statistiek / boekhouden
- Zijn bachelordiploma zal behalen en of hij dit zal doen volgens het modeltraject of niet.

Met toevoeging van de resultaten en gegevens op het einde van het jaar kunnen we zelfs nog beter voorspellingen gaan doen in verband met behalen van het bachelordiploma.

Op basis van de resultaten van onze analyses kunnen we stellen dat onze modellen zeer nuttig zouden kunnen zijn als adviesfunctie voor een student. De modellen bij aanvang van de studies presteren nog niet goed genoeg, maar deze na het eerste trimester zijn zeker goed genoeg om te gebruiken als advies naar een student toe. Zo kan een student eventueel geheroriënteerd worden na het eerste trimester waardoor hij geen heel jaar zal verliezen.

Als we naar de OneR beslissingsvariabelen kijken, zien we dat het algoritme hier bij de modellen voor aanvang van de studies telkens een variabele neemt die gewoon het beste resultaat behaalt omdat deze de beste opsplitsingmogelijkheden heeft. Dit haalden we eerder al aan in de opmerking bij Hoofdstuk 5. Deze beslissingsvariabelen zijn van weinig nut omdat ze geen logische grond hebben. Bij de modellen na de resultaten van het eerste trimester was de OneR beslissingsvariabele altijd wel een variabele die nuttige info gaf.

Voor de J48 beslissingsbomen hebben we vaak afwisselende resultaten. Soms is de boom logisch en begrijpbaar terwijl somsde boom veel te veel gefit op de data is en deze onlogische beslissingen maakt. Zoals eerder vermeld, mogen we uit deze bomen geen causale verbanden afleiden, maar toch reiken deze bomen ons vaak verbanden aan die verder onderzocht kunnen worden.

Als we in de tabel puur kijken naar de dataminingalgoritmes zien we dat het k-Nearest neighbor algoritme (Ibk) zeer slecht presteert op bijna elk model, deze afwijkend methode (zie uitleg algoritmes) is dus niet geschikt voor deze dataset. Het neurale netwerk presteert zeker niet slecht maar wordt in bijna alle analyses geoutperformd door de andere algoritmes. De andere algoritmes presteren alle drie gelijkaardig met naïeve Bayes die toch lichtjes naar voor komt als sterkste algoritme. Ook bij de eerdere gedane onderzoeken, vermeld in Hoofdstuk 4, zien we dat waar de naive bayes classifier gebruikt werd, deze beter presteerde dan de andere classifiers. Voor verder onderzoek in de toekomst raden we dus allereerst deze classifier aan, maar ook J48 en het MultilayerPerceptron lonen zeker de moeite om te gebruiken.

6.2. Opmerkingen, werkpunten en punten voor verder onderzoek

Omdat er slechts een kleine database te onzer beschikking was, zijn deze analyses gebaseerd op één jaar en een klein aantal studenten. Dit betekent dat, om te kijken of ze veralgemeenbaar zijn, we deze nog zouden moeten testen op meer studenten en andere jaren. Door dit te kleine aantal studenten zouden de modellen ook teveel gefit kunnen zijn op de data. Dit geeft vooral problemen bij de variabelen met veel mogelijkheden, zoals onder meer secundaire school en nationaliteit, die nu vaak mogelijkheden hebben met slechts een paar cases in. Dit feit zorgt ervoor dat de resultaten niet veralgemeend kunnen worden.

Ook de basisgegevens over een student zijn niet volledig. Deze zouden nog veel uitgebreider kunnen. Hierbij verwijzen we graag naar de literatuurstudie die nog een groot deel factoren aanhaalt die een belangrijke rol spelen in de slaagkans van een student. Als deze factoren opgenomen zouden worden in de database zou dit het voorspellend vermogen van de modellen zeker omhoog halen.

Lijst van geraadpleegde werken

Adelman, C. (1999). Answers in the Toolbox: Academic Intensity, Attendance Patterns, and Bachelor's Degree Attainment. U.S. Department of Education. Washington, DC: Office of Educational Research and Improvement.

Adelman, C. (2006). The Toolbox Revisited: Paths to Degree Completion From High School Through College. U.S. Department of Education. Washington, DC: Office of Vocational and Adult Education.

Ahmadi, M., Raiszadeh, F., & Helms, M. (1997) An examination of the admission criteria for MBA programs: A case study. Education, 117, 540–546.

Alexander, P. A., and Murphy, P. K. (1994). The Research Base for APA's Learner-Centered Psychological Principles. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Anderson, G., & Benjamin, D. (1994). The determinants of success in university introductory economics courses. Journal of Economic Education, 25, 99–118.

Astin, A. W. (1993). What Matters in College? Four Critical Years Revisited (1st Ed.). San Francisco: Jossey-Bass.

Attinasi, L. C., Jr. (1989). Getting in: Mexican Americans' Perceptions of University Attendance and the Implications for Freshman Year Persistence. Journal of Higher Education, 60(3):247-277.

Bailey, T., Jenkins, D., and Leinbach, T. (2005). Graduation Rates, Student Goals, and Measuring Community College Effectiveness (CCRC Brief Number 28). New York: Columbia University, Community College Research Center.

Bean, J. P. (1980). Dropouts and Turnover: The Synthesis and Test of a Causal Model of Student Attrition. Research in Higher Education, 12(2): 155-187.

Bean, J. P., and Bradley, R. (1986). Untangling the Satisfaction-Performance Relationship for College Students. Journal of Higher Education, 57(4): 393-412.

Bean, J. P., and Vesper, N. (1994). Gender Differences in College Student Satisfaction. Paper presented at the annual meeting of the Association for the Study of Higher Education, Tucson, AZ.

Belga "Rectoren Vlaamse universiteiten pleiten voor oriënteringsproef " Knack, 26 augustus, 2011.

Belga "Vlaamse rectoren willen oriënteringsproef invoeren" De morgen, 26 augustus, 2011.

Bijayananda, N., Srinivasan, R. (2004) "Using neural networks to predict MBA student success" College Student Journal 1: 143-150

Blimling, G. S. (1993). The Influence of College Residence Halls on Students. In Higher Education:Handbook of Theory and Research, Vol. 9, edited by J. C. Smart, 248-307. New York: Agathon.

Blimling, G. S., and Whitt, E. J. (1999). Identifying the Principles That Guide Student Affairs Practice. In Good Practice in Student Affairs: Principles To Foster Student Learning, edited by G. S.

Blimling and E. J. Whitt,. San Francisco: Jossey-Bass : 1-20

Borg, W. R. and Gall, M. D. (1989). Educational Research: An Introduction, 5th ed. New York: Longman.

Burton, L., Dowling, D. (2005) "In search of the key factors that influence student success at university" Higher education in a changing world, Proceedings of the 28th HERDSA Annual Conference

Carstens, J. B. (2000). Effects of a Freshman Orientation Course on Academic Outcomes, Quality of Effort and Estimated Intellectual Gain. PhD diss., University of Iowa, 2000. Abstract in Dissertation Abstracts International, 61, 2206.

Cheung, L. W., & Kan, C. N. (2002). Evaluation of factors related to student performance in a distance-learning business communication course. Journal of Education for Business, 77 : 257–259.

Choy, S. P. (1999). College Access and Affordability. Education Statistics Quarterly, 1(2):74-90.

Clayton, G. E., & Cate, T. (2004). Predicting MBA no-shows and graduation success with discriminate analysis. International Advances in Economic Research, 10, 235–243.

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. American Journal of Sociology, 94: 95-120.

Community College Survey of Student Engagement (CCSSE). (2005). Engaging Students, Challenging the Odds: 2005 Findings. Austin, TX: Author.

Deckro, R., & Woudenberg, H. (1997). MBA admission criteria and academic success. Decision Science, 8, 765-769.

Declercq, K., Verboven, F. " Slaagkansen aan Vlaamse universiteiten: tijd om het beleid bij te sturen? " Vives Briefings (2010)

Dekker, G. "Predicting students drop out: a case study" (2009)

Deredactie.be "Slaagcijfers universiteiten zijn laag", 27 September, 2010

Deredactie.be "Smet wil geen examen, maar wel een proef", 27 September, 2010

Didia, D., & Hasnat, B. (1998). The determinantsof performance in the university introductory finance course. Financial Practice and Education, 8(1), 102–107.

Dragičević, M., Bach M., Šimičević V. (2012) "Predicting student performance using decision trees" Proceedings of the IBC 2012 Internet & Business: 186-191.

Dreher, G. F., & Ryan, K. C. (2000). Prior work experience and academic performance among first-year MBA students. Research in Higher Education, 41, 505–525.

Dugan, M. K., Grady,W. R., Payn, B., Baydar, N., & Johnson, T. R. (1996). The importance ofprior work experience for full time MBA students: Evidence from GMAT registrant survey.

Edin Osmanbegović, E., Suljić, M. "(2012) Datamining approach for prediciting student performance." Journal of Economics and Business 10: 3-12

Ekpenyong, B. (2000). Empirical analysis of the relationship between students' attributes and performance: A case study of the University of Ibadan (Nigeria) MBA programme. Journal of Financial Management & Analysis, 13(2), 54–63.

Ely, D. P., & Hittle, L. (1990). The impact of performance in managerial economics and basic finance course. Journal of Financial Education, 8(2), 59-61.

Fisher, J. B., & Resrick, D. A. (1990). Standardized testing and graduate business school admission: A review of issues and analysis of a Baruch college MBA cohort. College and University, 65, 137–148.

Florida Department of Education. (2005). Postsecondary Success Begins With High School Preparation. Data Trend #33. Tallahassee, FL: Florida Department of Education

Frank, E., Witten, I. (2005) "Data Mining: Practical Machine Learning Tools and Techniques (second Edition)" Morgan Kaufmann

Friedman, J. "Data mining and statistics: What's the connection ?"

Fung, G., (2001) "A comprehensive overview of basis clustering algorithms"

Gladieux, L. E., and Swail, W. S. (1998). Financial Aid is Not Enough: Improving the Odds of College Success. College Board Review, (185): 16-21, 30-32.

Gleason, P. (1993). College Student Employment, Academic Progress, and Post-College Labor Market Success. Journal of Student Financial Aid, 23(2): 5-14.

Goethals, Maarten."Oriëntatieproef voor studenten tegen 2013" De standaard, 26 augustus, 2011.

Graham, L. D. (1991). Predicting academic successof students in a master of business administrationprogram. Educational and Psychological Measurement, 5, 721–727

Gutierrez, R. (2000). Advancing African-American, Urban Youth in Mathematics: Unpacking the Success of One Math Department. American Journal of Education, 109(1): 63-111.

Hand, D., Steinbach ,M. et al. (2008) "Top 10 algorithms in data mining" Knowl Inf Syst 14 : 1–37

Hand, David J. (1998) "Data Mining: Statistics and More?" The American Statistician 52: 112 -118

Hand, David J. (1999) "Statistics and Data Mining: Intersecting Disciplines" Acm sigkdd 1: 16-19

Hanks, M. P., and Eckland, B. K. (1976). Athletics and Social Participation in the Education Attainment Process. Sociology of Education, 49(4): 271-294.

Heller, D. E. (Ed.). (2002). Conditions of Access: Higher Education for Lower-Income Students. Westport, CT: American Council on Education/Praeger Series on Higher Education

Horn, L. J., and Kojaku, L. K. (2001). High School Academic Curriculum and the Persistence Path Through College: Persistence and Transfer Behavior of Undergraduates 3 Years After Entering 4-Year Institutions. Education Statistics Quarterly, 3(3): 65-72.

Hossler, D., Kuh, G. D., and Olsen, D. (2001). Finding Fruit on the Vines: Using Higher Educational Research and Institutional Research to Guide Institutional Policies and Strategies. Part II. Research in Higher Education, 42(2): 223-235.

Hoyt, J. E. (1999). Remedial Education and Student Attrition. Community College Review, 27(2): 51-72.

Institute for Higher Education Policy (IHEP). (2001). Getting Through College: Voices of Low-Income and Minority Students in New England. Washington, DC

Johnson, M. (2005). Academic performance of transfer versus native students, College Student Journal, 39(3), 570-580.

Kabakchieva, D. (2012) "Student Performance Prediction by Using Data Mining Classification Algorithms" International Journal of Computer Science and Management Research 1

Kruzicevic, S., Barisic, K., Banozic, A., Esteban, C. Sapunar, D., Puljak, L. (2012) "Predictors of Attrition and Academic Success of Medical Students: A 30-Year Retrospective Study" Plos One 7

Kuh, G. D. (1993). In Their Own Words: What Students Learn Outside the Classroom. American Educational Research Journal, 30(2): 277-304.

Kuh, G. D. (1995). The Other Curriculum: Out-Of-Class Experiences Associated With Student Learning and Personal Development. Journal of Higher Education, 66(2): 123-155.

Kuh, G. D. (2003). What We're Learning About Student Engagement From NSSE: Benchmarks for Effective Educational Practices. Change, 35(2): 24-32.

Kuh, G. D. (2005). Student Engagement in the First Year of College. In Challenging and Supporting the First-Year Student: A Handbook for Improving the First Year of College, edited by M. L. Upcraft, J. N. Gardner, and B. O. Barefoot, 86-107. San Francisco: Jossey-Bass.

Kuh, G. D., and Hu, S. (2001). The Effects of Student-Faculty Interaction in the 1990s. Review of Higher Education, 24(3): 309-332.

Kuh, G., Kinzie, J., Buckley, J., Bridges, B., Hayek, J. (2006) "What Matters to Student Success:A Review of the Literature" Report for the National Symposium on Postsecondary Student Success

Kuonen, Diego. (2004) "Data Mining and Statistics: What is the Connection?" TDAN.com

Launius, M. H. (1997). College student attendance: Attitudes and academic performance. College Student Journal, 31(1), 86–92.

London, H. B. (1989). Breaking Away: A Study of First-Generation College Students and Their Families. American Journal of Education, 97(1): 144-170.

Martinez, M., and Klopott, S. (2003). Improving College Access for Minority, Low-Income, and First- Generation Students. Boston: Pathways to College Network.

Masui, C. (2002). Leervaardigheid bevorderen in het hoger onderwijs: een ontwerponderzoek bij eerstejaarsstudenten. Leuven : universitaire pers Leuven.

Masui, C., Broeckmans, J., Doumen, S., Groenen, A. & Molenberghs, G. (2012) " Do diligent students perform better? Complex relations between student and course characteristics, study time, and academic performance in higher education" Studies in Higher Education

McClure, R. H.,Wells, C. E., & Bowerman, B. L. (1986). A model of MBA student performance. Research in Higher Education, 25, 182–193.

Murdock, T. A. (1990). Financial Aid and Persistence: An Integrative Review of the Literature. NASPA Journal: 27(3): 213-221.

N. Havermans , S. Vanassche. (2013) "scheiding in Vlaanderen"

Nuñez, A. M., and Cuccaro-Alamin, S. (1998). First-Generation Students: Undergraduates Whose Parents Never Enrolled in Postsecondary Education (NCES 98-082). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Okpala, A.O., Okpala, C.O., and R. Ellis. (2000). "Academic efforts and study habits among students in a principles of macroeconomics course." Journal of Education for Business 75:219-224.

Oladokun, V., Adebanjo, A., Charles-Owaba, O. (2008) "Predicting Students' Academic Performance using Artificial Neural Network" The Pacific Journal of Science and Technology 1: 72-79

Paolillo, J. G. P. (1982). The predictive validity of selected admission variables relative to grade point average earned in a master of business administration program. Educational and Psychological Measurement, 42, 1163–1167.

Pascarella, E .T., and Terenzini, P. T. (2005). How College Affects Students: A Third Decade of Research. San Francisco: Jossey-Bass.

Pascarella, E. T. (2001). Cognitive Growth in College: Surprising and Reassuring Findings. Change, 33(6): 20-27.

Pascarella, E. T., and Terenzini, P. T. (1977). Patterns of Student-Faculty Informal Interaction Beyond the Classroom and Voluntary Freshman Attrition. Journal of Higher Education, 48(5): 540-552.

Pascarella, E. T., and Terenzini, P. T. (1991). How College Affects Students: Findings and Insights From Twenty-Years of Research (1st ed.). San Francisco: Jossey-Bass Publishers.

Pascarella, E. T., Bohr, L., Nora, A., and Terenzini, P. T. (1995). Cognitive Effects of 2-Year and 4-Year Colleges: New Evidence. Educational Evaluation and Policy Analysis, 17(1): 83-96.

Pascarella, E. T., Edison, M. I., Nora, A. Hagedorn, L. S., and Terenzini, P. T. (1996). Influences on Students' Openness to Diversity and Challenge in the First Year of College. Journal of Higher Education, 67(2): 174-195.

Pascarella, E. T., Edison, M. I., Nora, A., Hagedorn, L. S., and Terenzini, P. T. (1998). Does Work Inhibit Cognitive Development During College? Educational Evaluation and Policy Analysis, 20(2): 75- 93.

Pathways to College Network. (2004). A Shared Agenda: A Leadership Challenge to Improve College Access and Success. Boston: The Education Resources Institute

Peiperl, M., & Trevelyan, R. (1997). Predictors of performance at business school and beyond:Demographic factors and the contrast between individual and group outcomes. Journal of Management Development, 16(5), 354-367.

Pike, G. R. (1991). Dimensions of Academic Growth and Development During College: Using Alumni Reports to Evaluate Education Programs. Paper presented at the annual meeting of the Association for the Student of Higher Education, Boston, MA.

Pike, G. R. (1993). The Relationship Between Perceived Learning and Satisfaction With College: An Alternative View. Research in Higher Education, 34(1): 23-40.

Pike, G. R., and Saupe, J. L. (2002). Does High School Matter? An Analysis of Three Methods of Predicting First-Year Grades. Research in Higher Education, 43(2): 187-207.

Quinlan, J. (1993) "C4.5: programs for machine learning". Morgan Kaufmann

Schumacher, P., Olinsky, A., Quinn, J., Smith, R., "A Comparison of Logistic Regression, NeuralNetworks, and Classification Trees Predicting

Singh, Y., Chauchan A. (2009) "Neural networks in data mining" Journal of Theoretical and Applied Information Technology

Sparkman, L., Maulding, W., Roberts, J. (2012" Non-cognitive predictors of student success in college." College student Journal 46): 642 - 652

Sternberg, R. (2006) "The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills ". Intelligence 34: 321–350

Success of Actuarial Students(2010)" Journal of Education for business 85: 258 - 263

Sulaiman, A., Mohezar, S. (2006) "Student Success Factors:Identifying Key Predictors" Journal of Education for Business: 328 - 333

The Pell Institute. (2004). Indicators of Opportunity in Higher Education: Fall 2004 Status Report. Washington, DC: Author.

Tierney, W. G., Corwin, Z. B., and Colyar, J. E. (Eds.) (2004). Preparing for College: Nine Elements for Effective Outreach. Albany, NY: State University of New York Press.

Warburton, E. C., Bugarin, R., and Nuñez, A. M. (2001). Bridging the Gap: Academic Preparation and Postsecondary Success of First-Generation Students. Education Statistics Quarterly, 3(3): 73-77.

Warren, S., (2004) "Neural networks and statistical models" Proceedings of the Nineteenth Annual SAS Users Group International Conference

Yang, B., & Lu, D. R. (2001). Predicting academic performance in management education: An empirical investigation of MBA success. Journal of Education for Business, 77, 15–21.
York-Anderson, D., and Bowman, S. (1991). Assessing the College Knowledge of First-Generation and Second-Generation Students. Journal of College Student Development, 32: 116-122.

Zhang, A., Aasheim, C. (2011) "Academic Success Factors: An IT Student Perspective " Journal of Information Technology Education 10

Zlatko, J., Kovačić (2010) "Early Prediction of Student Success: Mining Students Enrolment Data" Proceedings of Informing Science & IT Education Conference

Bijlagen

Bijlage 1 : Inputvariabelen van de eerdere onderzoeken

Kabakchieva (2012)

TABLE I
FINAL DATASET USED FOR THE DATA MINING ANALYSIS

Type of Data	Attribute Name	Attribute Type	Values
Personal Data	Gender	Nom	Male (49%), Female (51%)
	Age	Nom	29 distinct values (17-43,46,48,53) (18-21–95%, 19–78%)
	BirthYear	Nom	29 distinct values
Pre-University Data	PlacePrevEdu	Nom	7 distinct values (Sofia, NE, NC, NW, SE, SC, SW)
	ProfilePrevEdu	Nom	9 distinct values (Language, Natural_Math, Humanitarian, Economics, Technological, Business_Management, Arts, Sports, General)
	ScorePrevEdu	Num	3.40-6.00
	AdmissionYear	Nom	2007, 2008, 2009
	AdmissionExam	Nom	5 distinct values (BG,Math,Geography, History,Economics)
	AdmissionExamScore	Num	0.00-6.00
	AdmissionScore	Num	0.00-35.98
University Data	UnivSpecialtyName	Nom	10 distinct values
	CurrentSemester	Nom	1-10
	NumFailures	Nom	0-12 (0 – 86,4%)
	StudentClass	**Nom**	**Weak (5340), Strong (4727)**

Dekker (2009)

Attributes	Type	Typical value	Remarks
IDNR	numeric	0553453	Is removed after importing, used to check data integrity
VWO_Year	nominal	2001-2003	Five categories including 'n/a', corresponding to major changes in the Dutch education system
VWO_Profile	nominal	VWO(N+T)	The curriculum taken at pre-university education. Six categories including 'n/a'
VWO_nCourses	nominal	6	The number of courses taken. Integer values.
VWO_mean	nominal	'good'	{ n/a, poor, average, 'above average', good, excellent }
VWO_Science-_nCourses	nominal	>3	{ n/a, <3, 3, >3 }
VWO_Science-_mean	nominal	'good'	As VWO_mean
VWO_Math-_nCourses	nominal	1	{n/a, 0,1,2}
VWO_Math-_mean	nominal	'good'	As VWO_mean
HO_Education	nominal	'Electrical'	{n/a, Electrical, Technical, Other}
HO_Year	nominal	2001-2003	Same categories as VWO_Year
HO_Grade	nominal	'good'	As VWO_mean
Gapyear	nominal	-1	{n/a, <-1, -1, 0, 1, >1 }
Classification	nominal	1	{-1, 1}

Osmanbegović en Suljić (2012)

Table 2.1. Student related variables

Br.	Variable	Coding	Br.	Variable	Coding
1.	Gender (S)	A – male B – female	2.	Family (BCD)	Numeric value
3.	Distance (UAS)	Numeric value	4.	High School (VSS)	A – Grammar School B – High school for economics C – Rest
5.	GPA (PO)	Numeric value	6.	Entrance exam (URK)	Numeric value
7.	Scholarships (SS)	A – Not B – Sometimes C – Yes	8.	Time (VRI)	A – less then 1 hour B – from 1 to 2 hours C – from 2 to 3 hours D – from 3 to 4 hours E – from 4 to 5 hours
9.	Materials (MAT)	A – book, B – the notes of other students, C – notebook from the lectures, D – notes edited or made by student E – all that is available to student	10.	the Internet (INT)	A – Yes B – No
11.	Grade importance (VO)	A – Not important at all, B – not important C – Somewhat important, D – Important, E – Very important	12.	Earnings (MPD)	A – less then 500 KM B – from 500 to 1000 KM C – from 1000 to 1500 KM D – from 1500 to 2000 KM E – over 2000 KM

Dragičević, Bach en Šimičević (2012)

Variable group	Variables used for modelling	Dependent
		Grade Point Average
Demographic	Gender	not used
	Age	not used
Characteristics of study	Model of tuition fee	not used
	Length of study	not used
	Parallel study on another college	not used
	Did the student finished already some college program	not used
Situational criteria	Working while studying	✓ (yes)
	Location of permanent residence	not used
	Location of temporary residence	not used
	Dormitory or parents	
	Who is financing expenses and accommodation for the time of studying	✓ (student)
Parent background	Parents employment	not used
	Parents income	not used
	Parents level of degree	✓ (mother)

	Decision tree GPA criterion
Correctly classified instances	82,5%
Incorrectly classified instances	17,5%

Oladokun, Adebanjo en Charles-Owaba (2008)

S/N	Input variable	Domain		
1	UME score*	Score	Normalized score	
2	O/level results	Math	A1–A2 A3–C4 C5–C6	1 2 3
		English	A1–A2 A3–C4 C5–C6	1 2 3
		Physics	A1–A2 A3–C4 C5–C6	1 2 3
		Chemistry	A1–A2 A3–C4 C5–C6	1 2 3
3	Further math	Present and passed Present, not passed Not present		1 2 3
4	Age at entry	Below 23 years 23 years – above		1 2
5	Time before admission	1 year 2 years 3 years – above		1 2 3
6	Educated parent(s)	Yes No		1 2
7	Zone of secondary school attended	South-west South-south East North Lagos		1 2 3 4 5
8	Type of Secondary school	Private State Federal		1 2 3
9	Location of school	Located in home state Outside home state		1 2
10	Gender	Male Female		1 2

Table 2: Output Data Transformation.

S/N	Output Variable	Domain	
		Class	CGPA
1	GOOD	1st Class 2nd Class Upper	6.0 – 7.0 4.6 – 5.9
2	AVERAGE	2nd Class Lower	2.4 – 4.5
3	POOR	3rd Class Pass	1.8 – 2.3 1.0 – 1.7

Bijayananda, Srinivasan (2004)

In dit onderzoek werd gebruik gemaakt van volgende variabelen om de slaagkans te voorspellen : Campus locatie, burgerlijke status, geslacht, etniciteit,leeftijd, vroegere school, eerdere studierichting, 4-year undergraduate grade point average, junior-senior year grade point average, graduate management admission test score.

Schumacher et al. (2010)

In dit onderzoek werd een decision tree en een neural network gebouwd aan de hand van volgende significante factoren : Mathematics SAT scores, verbal SAT scores, percentile rank in high school graduating class en percentage score mathematics placement exam.

Zlatko en Kovačić (2010)

Table 1: Description of variables and their domains

Variable	Description (Domain)
	Student demographics
Gender	Student gender (binary: female or male)
Age	Student's age (numeric: 1 – under 30, 2 – 30 to 40 or 3 – over 40)
Ethnicity	Student's ethnic group (nominal: Pakeha, Maori & Pacific Islanders or Others)
Disability	Student has a disability (binary: yes or no)
Secondary school	Student's highest level of achievement from a secondary school on the New Zealand National Qualifications Framework (nominal: No secondary qualification, NCEA1, NCEA2, University entrance, NCEA3, Overseas or Other)
Work status	Student is working (binary: yes or no)
Early enrolment	Student enrolled for the first time in the course before start of the course (binary: yes or no)
	Study environment
Course programme	Programme (nominal: Bachelor of Business or Bachelor of Applied Science)
Course block	Semester in which a course is offered (Semester 1, Semester 2 or Semester 3)
	Dependent variable
Study outcome	Study outcome (nominal: Pass – successful completion, Fail – unsuccessful completion includes also withdrawals, academic withdrawals and transfers)

Table 3: Best predictors for dependent variable

Variable	Chi-square	*P*-value
Ethnicity	19.35	0.00006
Course programme	7.80	0.00523
Course block	5.51	0.06354
Secondary school	2.06	0.35748
Early enrolment	1.16	0.28131
Disability	0.86	0.35363
Gender	0.58	0.44774
Age	0.28	0.86750
Work status	0.07	0.78940

Bijlage 2 : voorbeeldcases dataset

Student

A	B	C	D	E	F	G	H	I	J	K	L	M
STUDENT_ID	GESLACHT	NATIONALITEIT	GEMEENTE	LAND	OP_KOT	SCHOOL_SO	SO_GEMEENTE	SO_RICHTING	SOWISK	SONAT	SOSCHEI	SOFRANS
61	GESLACHT1	NATIONALITEIT3	GEMEENTE172	LAND1	OP_KOT1	SCHOOL_SO113	SO_GEMEENTE73	SO_RICHTING37	8	3	2	NA
64	GESLACHT1	NATIONALITEIT3	GEMEENTE37	LAND1	OP_KOT2	SCHOOL_SO111	SO_GEMEENTE23	SO_RICHTING28	6	2	1	NA

N	O	P	Q	R	S	T	U	V
SOENGELS	SODUITS	BEURSSTUDENT	SO_STUDIEVERTRAGING	GENERATIESTUDENT	STPTN_BEHAALD	STPTN_NADELIBERATIE	VAKKEN_NIET_AFGELEGD	PERCENTAGE
NA	NA	BEURSSTUDENT2	0	0	54<= X < 60	60	0	58
NA	NA	BEURSSTUDENT2	0	0	60	60	0	59

W	X	Y	Z	AA	AB
EINDBEOORDELING	DATUM_VASTLEGGING_EINDBEOORDELING	BA_DIPLOMA	BA_DIPL_AJ	MA_DIPLOMA	MA_DIPL_AJ
V	tweedezit	bachelor BI	2009	NA	NA
V	tweedezit	bachelor TEW	2006	master TEW-BM	2007

Resultaten2004

A	B	C	D	E	F	G	H	I	J	K	L	M	N
STUDENT_	ACJAAR	GRNAAM	DATUIT	MODELTRAJECT	VAKNR	VKSTUGID	VKNAAM	TYPE	PTKODE1	PTP1	VKGEW1	PTKODE2	PTP2
61	2004	bachelor TEW-HI-HIB	NA	1 bachelor TEW-HI-HIB	1	1169	Macro-ec	hoofdvak		+17.0	3		+08.0
61	2004	bachelor TEW-HI-HIB	NA	1 bachelor TEW-HI-HIB	2	1237	Geo-econ	hoofdvak		+13.0	3		

O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
VKGEW2	PTKODE3	PTP3	VKGEW3	PTKODE4	PTP4	VKGEW4	PTKODE5	PTP5	VKGEW5	PTKODE6	PTP6	VKGEW6	PTKODEK1	PTPK1	PTKODEK2	PTPK2	PTKODE	PTP
3		NA	0			0		NA	0		NA	0		+13.0				+13.0
0		NA	0			0		NA	0		NA	0		+13.0				+13.0

Studietijdmeting

A	B	C	D	E	F	G	H
VKSTUGIDS	WEEKNR	KLASSE	OANM	MINUTEN	DATREG	STUDENT_ID	SBU
1169	0	Zelfstudieopdracht	Macro-economie	495	20041206	309	360
1169	0	Zelfstudieopdracht	Macro-economie	660	20041206	109	360

Denk en leestest

A	B	C	D	E	F	G	H
STUDENT_ID	STUDENTID	LEESTEST.DEELNAME	LEESTEST.TOTAAL	DENKTEST.TOTAAL	DENKTEST.VERBAAL	DENKTEST.NUMERIEK	DENKTEST.GRAFISCH
61	104	deelgenomen	77	46	15	15	16
64	105	deelgenomen	79	23	9	5	9

Bijlage 3 : resultaten clustering

Hierachisch 4cluster

Beginselen van het recht. Privaat en publiek recht. Handelsrecht	cluster0
Financieel boekhouden	cluster0
Geo-economie en economische geschiedenis	cluster1
Informatica	cluster1
Macro-economie	cluster1
Management en Management accounting	cluster1
Onderzoeksmethoden en psychologie	cluster1
Sociologie en demografie	cluster1
Talen	cluster1
Statistiek	cluster2
Wiskunde	cluster3

Simple K means 4cluster

Management en Management accounting	cluster0
Beginselen van het recht. Privaat en publiek recht. Handelsrecht	cluster1
Financieel boekhouden	cluster1
Geo-economie en economische geschiedenis	cluster2
Macro-economie	cluster2
Onderzoeksmethoden en psychologie	cluster2
Sociologie en demografie	cluster2
Talen	cluster2
Wiskunde	cluster2
Informatica	cluster3
Statistiek	cluster3

Simple K means 3cluster

Management en Management accounting	cluster0
Beginselen van het recht. Privaat en publiek recht. Handelsrecht	cluster1
Financieel boekhouden	cluster1
Geo-economie en economische geschiedenis	cluster2
Informatica	cluster2
Macro-economie	cluster2
Onderzoeksmethoden en psychologie	cluster2
Sociologie en demografie	cluster2
Statistiek	cluster2
Talen	cluster2
Wiskunde	cluster2

Simple K means 5cluster

Management en Management accounting	cluster0
Beginselen van het recht. Privaat en publiek recht. Handelsrecht	cluster1
Financieel boekhouden	cluster1
Geo-economie en economische geschiedenis	cluster2
Macro-economie	cluster2
Onderzoeksmethoden en psychologie	cluster2
Sociologie en demografie	cluster2
Wiskunde	cluster2
Informatica	cluster3
Statistiek	cluster3
Talen	cluster4

FarthestFirst algorithm 4cluster

Geo-economie en economische geschiedenis	cluster0
Informatica	cluster0
Macro-economie	cluster0
Onderzoeksmethoden en psychologie	cluster0
Sociologie en demografie	cluster0
Talen	cluster0
Beginselen van het recht. Privaat en publiek recht. Handelsrecht	cluster1
Financieel boekhouden	cluster1
Management en Management accounting	cluster1
Statistiek	cluster2
Wiskunde	cluster3

Bijlage 4 : Weka output van de analyses

1) Voorspellen of een student gaat slagen in zijn eerst jaar.

- **OneR Classifier**
  - ***Voor het begin van de studies***

```
Correctly Classified Instances          95               49.4792 %
Incorrectly Classified Instances        97               50.5208 %
Kappa statistic                         -0.0139
Mean absolute error                      0.5052
Root mean squared error                  0.7108
Relative absolute error                101.0296 %
Root relative squared error            142.1298 %
Total Number of Instances              192


=== Confusion Matrix ===

  a  b   <-- classified as
 64 33 |  a = G
 64 31 |  b = NG
```

OneR beslissingsvariabele : secundaire school

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         146               76.0417 %
Incorrectly Classified Instances        46               23.9583 %
Kappa statistic                          0.5203
Mean absolute error                      0.2396
Root mean squared error                  0.4895
Relative absolute error                 47.911  %
Root relative squared error             97.8764 %
Total Number of Instances              192

 a  b   <-- classified as
79 18 |  a = G
28 67 |  b = NG
```

OneR beslissingsvariabele : macro economie

- **Naive Bayes**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances         136               70.8333 %
Incorrectly Classified Instances        56               29.1667 %
Kappa statistic                          0.4152
Mean absolute error                      0.3249
Root mean squared error                  0.4989
Relative absolute error                 64.9632 %
Root relative squared error             99.7628 %
Total Number of Instances              192


=== Confusion Matrix ===

  a  b   <-- classified as
 80 17 |  a = G
 39 56 |  b = NG
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         161               83.8542 %
Incorrectly Classified Instances        31               16.1458 %
Kappa statistic                          0.6768
Mean absolute error                      0.1589
Root mean squared error                  0.3653
Relative absolute error                 31.7751 %
Root relative squared error             73.0375 %
Total Number of Instances              192


  a  b   <-- classified as
 85 12 |  a = G
 19 76 |  b = NG
```

.

- **J48**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances         123               64.0625 %
Incorrectly Classified Instances        69               35.9375 %
Kappa statistic                          0.2781
Mean absolute error                      0.4217
Root mean squared error                  0.4831
Relative absolute error                 84.3365 %
Root relative squared error             96.6029 %
Total Number of Instances              192
```

```
=== Confusion Matrix ===

  a  b   <-- classified as
 82 15 |  a = G
 54 41 |  b = NG


J48 pruned tree
------------------

SO_STUDIEVERTRAGING = nee
|   SOENGELS = 2tot2uur: G (127.29/43.76)
|   SOENGELS = 3ofmeeruur: NG (39.71/10.48)
SO_STUDIEVERTRAGING = ja
|   DENKTEST.NUMERIEK <= 15: NG (22.06/0.88)
|   DENKTEST.NUMERIEK > 15: G (2.94/0.82)
```

- ***<u>Na resultaten 1ste trimester</u>***

```
Correctly Classified Instances         156               81.25   %
Incorrectly Classified Instances        36               18.75   %
Kappa statistic                          0.625
Mean absolute error                      0.228
Root mean squared error                  0.3897
Relative absolute error                 45.5967 %
Root relative squared error             77.9181 %
Total Number of Instances              192


=== Confusion Matrix ===

  a  b   <-- classified as
 79 18 |  a = G
 18 77 |  b = NG

J48 pruned tree
------------------

Macro-economie <= 11
|   Financieel boekhouden <= 9
|   |   Wiskunde <= 9: NG (68.52/1.0)
|   |   Wiskunde > 9
|   |   |   GESLACHT = GESLACHT1: NG (5.13/1.0)
|   |   |   GESLACHT = GESLACHT2: G (2.13/0.13)
|   Financieel boekhouden > 9
|   |   LEESTEST.TOTAAL <= 79: G (5.77/0.06)
|   |   LEESTEST.TOTAAL > 79: NG (2.31/0.29)
Macro-economie > 11
|   Wiskunde <= 4
|   |   Financieel boekhouden <= 10: NG (8.18)
|   |   Financieel boekhouden > 10: G (2.02/0.02)
|   Wiskunde > 4: G (97.93/12.93)
```

- **IBk**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances         115               59.8958 %
Incorrectly Classified Instances        77               40.1042 %
Kappa statistic                          0.197
Mean absolute error                      0.4022
Root mean squared error                  0.6297
Relative absolute error                 80.425  %
Root relative squared error            125.9108 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b   <-- classified as
 63 34 |  a = G
 43 52 |  b = NG
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         123               64.0625 %
Incorrectly Classified Instances        69               35.9375 %
Kappa statistic                          0.2813
Mean absolute error                      0.361
Root mean squared error                  0.5961
Relative absolute error                 72.1882 %
Root relative squared error            119.1913 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b   <-- classified as
 62 35 |  a = G
 34 61 |  b = NG
```

- **MultilayerPerceptron**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances         122               63.5417 %
Incorrectly Classified Instances        70               36.4583 %
Kappa statistic                          0.2703
Mean absolute error                      0.3658
Root mean squared error                  0.5588
Relative absolute error                 73.1465 %
Root relative squared error            111.7297 %
Total Number of Instances              192
```

```
 a  b   <-- classified as
65 32 |  a = G
38 57 |  b = NG
```

- ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         168               87.5    %
Incorrectly Classified Instances        24               12.5    %
Kappa statistic                          0.7499
Mean absolute error                      0.1331
Root mean squared error                  0.3342
Relative absolute error                 26.6148 %
Root relative squared error             66.8261 %
Total Number of Instances              192
```

```
  a  b   <-- classified as
 87 10 |  a = G
 14 81 |  b = NG
```

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
geslaagd of niet	vooraanvang	49	71	64	60	64
	naresult1	76	84	81	64	88

2) Voorspellen van aantal herexamens in het eerste jaar

- **OneR Classifier**
  - ***Voor het begin van de studies***

```
Correctly Classified Instances          65               33.8542 %
Incorrectly Classified Instances       127               66.1458 %
Kappa statistic                          0.1144
Mean absolute error                      0.2646
Root mean squared error                  0.5144
Relative absolute error                 91.1269 %
Root relative squared error            135.1651 %
Total Number of Instances              192
```

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  9  0  7  6  3 |  a = tweedrie
  1  0  1  1  2 |  b = een
 16  0 34 14  3 |  c = meeralsvijf
 24  0 17 18  5 |  d = 0.0
 12  3  7  5  4 |  e = meeralsdrie
```

OneR beslissingsvariabele : gemeente van afkomst

- ***<u>Na resultaten 1ste trimester</u>***

```
Correctly Classified Instances         113               58.8542 %
Incorrectly Classified Instances        79               41.1458 %
Kappa statistic                          0.4204
Mean absolute error                      0.1646
Root mean squared error                  0.4057
Relative absolute error                 56.6853 %
Root relative squared error            106.6046 %
Total Number of Instances              192
```

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  2  0  7  8  8 |  a = tweedrie
  0  0  1  1  3 |  b = een
  6  0 52  3  6 |  c = meeralsvijf
  7  0  1 51  5 |  d = 0.0
  6  0 11  6  8 |  e = meeralsdrie
```

OneR beslissingsvariabele : wiskunde

- **<u>Naive Bayes</u>**
  - ***<u>Voor het begin van de studies</u>***

```
Correctly Classified Instances          93               48.4375 %
Incorrectly Classified Instances        99               51.5625 %
Kappa statistic                          0.2627
Mean absolute error                      0.2251
Root mean squared error                  0.3987
Relative absolute error                 77.5161 %
Root relative squared error            104.7791 %
Total Number of Instances              192
```

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  3  0 10 10  2 |  a = tweedrie
  1  0  2  1  1 |  b = een
  5  0 37 19  6 |  c = meeralsvijf
  1  2  4 48  9 |  d = 0.0
  3  0  7 16  5 |  e = meeralsdrie
```

- ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         136               70.8333 %
Incorrectly Classified Instances        56               29.1667 %
Kappa statistic                          0.5926
Mean absolute error                      0.124
Root mean squared error                  0.2961
Relative absolute error                 42.6928 %
Root relative squared error             77.8128 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  7  0  6  7  5 |  a = tweedrie
  1  1  0  2  1 |  b = een
  5  0 57  1  4 |  c = meeralsvijf
  2  0  0 54  8 |  d = 0.0
  4  1  3  6 17 |  e = meeralsdrie
```

➢ **J48**

- ***Voor het begin van de studies***

```
Correctly Classified Instances          76               39.5833 %
Incorrectly Classified Instances       116               60.4167 %
Kappa statistic                          0.0912
Mean absolute error                      0.2827
Root mean squared error                  0.3815
Relative absolute error                 97.3587 %
Root relative squared error            100.25   %
Total Number of Instances              192
```

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  0  0 10 14  1 |  a = tweedrie
  0  0  3  2  0 |  b = een
  0  0 33 33  1 |  c = meeralsvijf
  0  0 18 43  3 |  d = 0.0
  0  0 14 17  0 |  e = meeralsdrie

SO_STUDIEVERTRAGING = nee: 0.0 (167.0/104.0)
SO_STUDIEVERTRAGING = ja: meeralsvijf (25.0/5.0)
```

- ***<u>Na resultaten 1ste trimester</u>***

```
Correctly Classified Instances         115               59.8958 %
Incorrectly Classified Instances        77               40.1042 %
Kappa statistic                          0.4407
Mean absolute error                      0.1851
Root mean squared error                  0.3383
Relative absolute error                 63.7611 %
Root relative squared error             88.8926 %
Total Number of Instances              192
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  9  0  6  7  3 |  a = tweedrie
  0  0  0  2  3 |  b = een
  5  0 52  2  8 |  c = meeralsvijf
  6  0  2 47  9 |  d = 0.0
  6  1  8  9  7 |  e = meeralsdrie


Macro-economie <= 11
|   Financieel boekhouden <= 9
|   |   Wiskunde <= 9
|   |   |   BEURSSTUDENT = BEURSSTUDENT2
|   |   |   |   Geo-economie en economische geschiedenis <= 12: meeralsvijf (48.44/2.0)
|   |   |   |   Geo-economie en economische geschiedenis > 12
|   |   |   |   |   DENKTEST.GRAFISCH <= 13: meeralsvijf (2.43)
|   |   |   |   |   DENKTEST.GRAFISCH > 13: meeralsdrie (3.65/0.65)
|   |   |   BEURSSTUDENT = BEURSSTUDENT1
|   |   |   |   Macro-economie <= 9: meeralsvijf (8.0/1.0)
|   |   |   |   Macro-economie > 9: meeralsdrie (6.0/1.0)
|   |   Wiskunde > 9
|   |   |   GESLACHT = GESLACHT1: meeralsdrie (5.13/0.13)
|   |   |   GESLACHT = GESLACHT2: tweedrie (2.13/1.13)
|   Financieel boekhouden > 9
|   |   Macro-economie <= 10: meeralsdrie (4.04/1.04)
|   |   Macro-economie > 10
|   |   |   DENKTEST.TOTAAL <= 38: een (2.02/0.02)
|   |   |   DENKTEST.TOTAAL > 38: 0.0 (2.02/0.02)
Macro-economie > 11
|   Wiskunde <= 10: tweedrie (47.94/30.94)
|   Wiskunde > 10
|   |   Geo-economie en economische geschiedenis <= 9
|   |   |   LEESTEST.TOTAAL <= 71: 0.0 (2.25/0.25)
|   |   |   LEESTEST.TOTAAL > 71: tweedrie (3.37/1.37)
|   |   Geo-economie en economische geschiedenis > 9: 0.0 (54.57/4.57)
```

- **IBk**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances          70               36.4583 %
Incorrectly Classified Instances       122               63.5417 %
Kappa statistic                          0.1308
Mean absolute error                      0.256
Root mean squared error                  0.4971
Relative absolute error                 88.1767 %
Root relative squared error            130.6348 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  2  2  7  8  6 |  a = tweedrie
  1  0  1  2  1 |  b = een
  9  3 33 15  7 |  c = meeralsvijf
  8  2 13 28 13 |  d = 0.0
  3  2  5 14  7 |  e = meeralsdrie
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances          79               41.1458 %
Incorrectly Classified Instances       113               58.8542 %
Kappa statistic                          0.2027
Mean absolute error                      0.2378
Root mean squared error                  0.4785
Relative absolute error                 81.9007 %
Root relative squared error            125.7267 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  2  2  7  6  8 |  a = tweedrie
  2  0  0  2  1 |  b = een
 12  3 37  4 11 |  c = meeralsvijf
  7  1 12 33 11 |  d = 0.0
  5  2  5 12  7 |  e = meeralsdrie
```

- **MultilayerPerceptron**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances          74               38.5417 %
Incorrectly Classified Instances       118               61.4583 %
Kappa statistic                          0.152
Mean absolute error                      0.2498
Root mean squared error                  0.4595
Relative absolute error                 86.0514 %
Root relative squared error            120.7488 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  5  0  9  6  5 |  a = tweedrie
  0  0  4  0  1 |  b = een
  3  1 38 13 12 |  c = meeralsvijf
  5  1 14 25 19 |  d = 0.0
  5  2  7 11  6 |  e = meeralsdrie
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         120               62.5    %
Incorrectly Classified Instances        72               37.5    %
Kappa statistic                          0.4812
Mean absolute error                      0.1489
Root mean squared error                  0.3463
Relative absolute error                 51.2902 %
Root relative squared error             91.0118 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  5  1  5  4 10 |  a = tweedrie
  2  0  0  2  1 |  b = een
  6  0 55  2  4 |  c = meeralsvijf
  4  3  4 47  6 |  d = 0.0
  5  1  4  8 13 |  e = meeralsdrie
```

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
#herex	vooraanvang	34	48	40	36	39
	naresult1	59	71	60	41	63

- **OneR Classifier**
  - ***Voor het begin van de studies***

```
Correctly Classified Instances          83               43.2292 %
Incorrectly Classified Instances       109               56.7708 %
Kappa statistic                          0.1529
Mean absolute error                      0.3785
Root mean squared error                  0.6152
Relative absolute error                 85.1991 %
Root relative squared error            130.5218 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
 39 12 10 |  a = 1totmet5
 26 28 13 |  b = meeralsvijf
 33 15 16 |  c = 0.0
```

OneR beslissingsvariabele : gemeente van afkomst

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         136               70.8333 %
Incorrectly Classified Instances        56               29.1667 %
Kappa statistic                          0.5631
Mean absolute error                      0.1944
Root mean squared error                  0.441
Relative absolute error                 43.772  %
Root relative squared error             93.5543 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
 38 12 11 |  a = 1totmet5
 20 46  1 |  b = meeralsvijf
 12  0 52 |  c = 0.0
```

OneR beslissingsvariabele : wiskunde

- **Naive Bayes**
  - ***Voor het begin van de studies***

```
Correctly Classified Instances          97               50.5208 %
Incorrectly Classified Instances        95               49.4792 %
Kappa statistic                          0.2574
Mean absolute error                      0.3361
Root mean squared error                  0.4881
Relative absolute error                 75.659  %
Root relative squared error            103.5512 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
 20 17 24 |  a = 1totmet5
 13 36 18 |  b = meeralsvijf
 19  4 41 |  c = 0.0
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         158               82.2917 %
Incorrectly Classified Instances        34               17.7083 %
Kappa statistic                          0.7344
Mean absolute error                      0.15
Root mean squared error                  0.3318
Relative absolute error                 33.7644 %
Root relative squared error             70.3897 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
 46  5 10 |  a = 1totmet5
  8 58  1 |  b = meeralsvijf
 10  0 54 |  c = 0.0
```

- **J48**
  - ***Voor het begin van de studies***

```
Correctly Classified Instances          79               41.1458 %
Incorrectly Classified Instances       113               58.8542 %
Kappa statistic                          0.1169
Mean absolute error                      0.4246
Root mean squared error                  0.5277
Relative absolute error                 95.5884 %
Root relative squared error            111.9562 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
 20 16 25 |  a = 1totmet5
 14 31 22 |  b = meeralsvijf
 23 13 28 |  c = 0.0


SO_STUDIEVERTRAGING = nee
|   GEMEENTE = GEMEENTE172
|   |   GESLACHT = GESLACHT1: 1totmet5 (2.0)
|   |   GESLACHT = GESLACHT2: meeralsvijf (4.0/1.0)
|   GEMEENTE = GEMEENTE37
|   |   SO_RICHTING = SO_RICHTING37: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING28: 1totmet5 (1.0)
|   |   SO_RICHTING = SO_RICHTING34: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING23: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING24: meeralsvijf (3.0/1.0)
|   |   SO_RICHTING = SO_RICHTING9: 1totmet5 (5.0/2.0)
|   |   SO_RICHTING = SO_RICHTING7: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING29: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING30: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING22: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING12: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING32: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING4: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING1: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING16: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING3: 1totmet5 (0.0)
|   |   SO_RICHTING = SO_RICHTING27: 1totmet5 (0.0)
|   GEMEENTE = GEMEENTE99: 1totmet5 (3.0/2.0)
|   GEMEENTE = GEMEENTE34: meeralsvijf (1.0)
|   GEMEENTE = GEMEENTE30: meeralsvijf (1.0)
|   GEMEENTE = GEMEENTE159: 0.0 (1.0)
|   GEMEENTE = GEMEENTE58
|   |   SOSCHEI = 2ofmeeruur: 0.0 (2.0)
```

```
|   |   SOSCHEI = 0/1uur
|   |   |   DENKTEST.GRAFISCH <= 12: meeralsvijf (2.0)
|   |   |   DENKTEST.GRAFISCH > 12: 1totmet5 (4.0)
|   GEMEENTE = GEMEENTE1: 1totmet5 (1.0)
|   GEMEENTE = GEMEENTE65: 1totmet5 (3.0/2.0)
|   GEMEENTE = GEMEENTE16
|   |   GESLACHT = GESLACHT1
|   |   |   SONAT = 3ofmeeruur: 0.0 (2.0)
|   |   |   SONAT = 2uur: 1totmet5 (2.0/1.0)
|   |   |   SONAT = 0/1uur: 0.0 (0.0)
|   |   GESLACHT = GESLACHT2: 1totmet5 (3.0)
|   GEMEENTE = GEMEENTE145: 0.0 (3.0/1.0)
|   GEMEENTE = GEMEENTE98: 1totmet5 (9.0/3.0)
|   GEMEENTE = GEMEENTE144: 1totmet5 (1.0)
|   GEMEENTE = GEMEENTE55: 1totmet5 (1.0)
|   GEMEENTE = GEMEENTE147
|   |   DENKTEST.NUMERIEK <= 13: meeralsvijf (2.0)
|   |   DENKTEST.NUMERIEK > 13: 0.0 (4.0/1.0)
|   GEMEENTE = GEMEENTE49
|   |   SONAT = 3ofmeeruur: 0.0 (2.0)
|   |   SONAT = 2uur: meeralsvijf (0.0)
|   |   SONAT = 0/1uur: meeralsvijf (9.0/2.0)
|   GEMEENTE = GEMEENTE46: 1totmet5 (1.0)
|   GEMEENTE = GEMEENTE18: 0.0 (1.0)
|   GEMEENTE = GEMEENTE174: 1totmet5 (1.0)
|   GEMEENTE = GEMEENTE79: 0.0 (7.0/2.0)
|   GEMEENTE = GEMEENTE9: meeralsvijf (2.0/1.0)
|   GEMEENTE = GEMEENTE13: 0.0 (1.0)
|   GEMEENTE = GEMEENTE84
|   |   DENKTEST.NUMERIEK <= 11: meeralsvijf (2.0/1.0)
|   |   DENKTEST.NUMERIEK > 11: 1totmet5 (4.0/1.0)
|   GEMEENTE = GEMEENTE94: meeralsvijf (3.0/1.0)
```

```
|   GEMEENTE = GEMEENTE113: meeralsvijf (2.0)
|   GEMEENTE = GEMEENTE101: meeralsvijf (1.0)
|   GEMEENTE = GEMEENTE29: 0.0 (4.0/1.0)
|   GEMEENTE = GEMEENTE75: 1totmet5 (1.0)
|   GEMEENTE = GEMEENTE15: 1totmet5 (3.0/1.0)
|   GEMEENTE = GEMEENTE161: 0.0 (1.0)
|   GEMEENTE = GEMEENTE142: 0.0 (0.0)
|   GEMEENTE = GEMEENTE50: meeralsvijf (2.0/1.0)
|   GEMEENTE = GEMEENTE60: 1totmet5 (1.0)
|   GEMEENTE = GEMEENTE118: 0.0 (1.0)
|   GEMEENTE = GEMEENTE122: 1totmet5 (2.0/1.0)
|   GEMEENTE = GEMEENTE130
|   |   DENKTEST.GRAFISCH <= 13: meeralsvijf (2.0/1.0)
|   |   DENKTEST.GRAFISCH > 13: 1totmet5 (2.0)
|   GEMEENTE = GEMEENTE128: meeralsvijf (1.0)
|   GEMEENTE = GEMEENTE28: meeralsvijf (3.0/1.0)
|   GEMEENTE = GEMEENTE22: 0.0 (4.0/1.0)
|   GEMEENTE = GEMEENTE2
|   |   GESLACHT = GESLACHT1: 0.0 (3.0)
|   |   GESLACHT = GESLACHT2: meeralsvijf (2.0)
|   GEMEENTE = GEMEENTE109: meeralsvijf (2.0/1.0)
|   GEMEENTE = GEMEENTE62
|   |   DENKTEST.GRAFISCH <= 14: 1totmet5 (2.0)
|   |   DENKTEST.GRAFISCH > 14: meeralsvijf (2.0/1.0)
|   GEMEENTE = GEMEENTE73: 1totmet5 (1.0)
|   GEMEENTE = GEMEENTE111: 0.0 (1.0)
|   GEMEENTE = GEMEENTE164: 0.0 (2.0)
|   GEMEENTE = GEMEENTE124: meeralsvijf (2.0/1.0)
|   GEMEENTE = GEMEENTE35
|   |   OP_KOT = OP_KOT1: 0.0 (2.0)
|   |   OP_KOT = OP_KOT2: 1totmet5 (2.0/1.0)
|   GEMEENTE = GEMEENTE131: meeralsvijf (1.0)
|   GEMEENTE = GEMEENTE74: 0.0 (2.0)
|   GEMEENTE = GEMEENTE59: 0.0 (0.0)
|   GEMEENTE = GEMEENTE133: 1totmet5 (1.0)
|   GEMEENTE = GEMEENTE82: 0.0 (3.0/1.0)
|   GEMEENTE = GEMEENTE141: meeralsvijf (1.0)
|   GEMEENTE = GEMEENTE140: meeralsvijf (3.0/1.0)
|   GEMEENTE = GEMEENTE135: 1totmet5 (2.0)
|   GEMEENTE = GEMEENTE149: 0.0 (1.0)
|   GEMEENTE = GEMEENTE56: 1totmet5 (2.0/1.0)
|   GEMEENTE = GEMEENTE54: 1totmet5 (2.0)
|   GEMEENTE = GEMEENTE87: 0.0 (0.0)
|   GEMEENTE = GEMEENTE6: 0.0 (1.0)
|   GEMEENTE = GEMEENTE90: 0.0 (1.0)
|   GEMEENTE = GEMEENTE47: 0.0 (1.0)
|   GEMEENTE = GEMEENTE123: meeralsvijf (1.0)
|   GEMEENTE = GEMEENTE138: 0.0 (1.0)
SO_STUDIEVERTRAGING = ja: meeralsvijf (25.0/5.0)
```

- ***Na resultaten 1ste trimester***

```
=== Summary ===

Correctly Classified Instances         137               71.3542 %
Incorrectly Classified Instances        55               28.6458 %
Kappa statistic                          0.5714
Mean absolute error                      0.2368
Root mean squared error                  0.3942
Relative absolute error                 53.3002 %
Root relative squared error             83.6346 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
 44  9  8 |  a = 1totmet5
 17 49  1 |  b = meeralsvijf
 19  1 44 |  c = 0.0
```

```
Financieel boekhouden <= 9
|   Sociologie en demografie <= 9: meeralsvijf (29.31)
|   Sociologie en demografie > 9
|   |   Macro-economie <= 11
|   |   |   Wiskunde <= 9
|   |   |   |   BEURSSTUDENT = BEURSSTUDENT2: meeralsvijf (30.43/5.0)
|   |   |   |   BEURSSTUDENT = BEURSSTUDENT1
|   |   |   |   |   Macro-economie <= 9: meeralsvijf (5.0/1.0)
|   |   |   |   |   Macro-economie > 9: 1totmet5 (6.0)
|   |   |   Wiskunde > 9: 1totmet5 (7.07/0.07)
|   |   Macro-economie > 11: 1totmet5 (44.46/21.46)
Financieel boekhouden > 9
|   Wiskunde <= 10
|   |   DENKTEST.NUMERIEK <= 13: 1totmet5 (16.41/2.17)
|   |   DENKTEST.NUMERIEK > 13: 0.0 (6.15/1.15)
|   Wiskunde > 10: 0.0 (47.17/4.17)
```

- **IBk**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances          89               46.3542 %
Incorrectly Classified Instances       103               53.6458 %
Kappa statistic                          0.1971
Mean absolute error                      0.3591
Root mean squared error                  0.593
Relative absolute error                 80.8426 %
Root relative squared error            125.8027 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
 29 10 22 |  a = 1totmet5
 17 34 16 |  b = meeralsvijf
 27 11 26 |  c = 0.0
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances          97               50.5208 %
Incorrectly Classified Instances        95               49.4792 %
Kappa statistic                          0.2597
Mean absolute error                      0.3318
Root mean squared error                  0.5695
Relative absolute error                 74.6962 %
Root relative squared error            120.8194 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
 30 11 20 |  a = 1totmet5
 25 37  5 |  b = meeralsvijf
 24 10 30 |  c = 0.0
```

- **MultilayerPerceptron**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances          88               45.8333 %
Incorrectly Classified Instances       104               54.1667 %
Kappa statistic                          0.1889
Mean absolute error                      0.3662
Root mean squared error                  0.5574
Relative absolute error                 82.4259 %
Root relative squared error            118.2626 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
 25 14 22 |  a = 1totmet5
 18 36 13 |  b = meeralsvijf
 29  8 27 |  c = 0.0
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         136               70.8333 %
Incorrectly Classified Instances        56               29.1667 %
Kappa statistic                          0.5626
Mean absolute error                      0.1987
Root mean squared error                  0.4024
Relative absolute error                 44.7281 %
Root relative squared error             85.3752 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
 37 10 14 |  a = 1totmet5
 12 52  3 |  b = meeralsvijf
 16  1 47 |  c = 0.0
```

#herex2

	OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
vooraanvang	43	51	41	46	46
naresult1	71	82	71	51	71

3) ***Voorspellen van behaalde percentage in het eerste jaar***

- **OneR Classifier**
  - ***Voor het begin van de studies***

```
Correctly Classified Instances          80               41.6667 %
Incorrectly Classified Instances       112               58.3333 %
Kappa statistic                          0.0362
Mean absolute error                      0.2917
Root mean squared error                  0.5401
Relative absolute error                 86.4656 %
Root relative squared error            131.6477 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 67 14  1 10 |  a = kleiner65
 27  3  1  4 |  b = groter65
 12  2  4  3 |  c = kleiner35
 30  5  3  6 |  d = kleiner50
```

OneR beslissingsvariabele gemeente van afkomst

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances          98               51.0417 %
Incorrectly Classified Instances        94               48.9583 %
Kappa statistic                          0.2139
Mean absolute error                      0.2448
Root mean squared error                  0.4948
Relative absolute error                 72.5693 %
Root relative squared error            120.6058 %
Total Number of Instances              192
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 70  8  1 13 |  a = kleiner65
 22 11  0  2 |  b = groter65
  6  1  4 10 |  c = kleiner35
 24  4  3 13 |  d = kleiner50
```

OneR beslissingsvariabele : macro economie

- **Naive Bayes**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances          83               43.2292 %
Incorrectly Classified Instances       109               56.7708 %
Kappa statistic                          0.1434
Mean absolute error                      0.29
Root mean squared error                  0.4562
Relative absolute error                 85.9648 %
Root relative squared error            111.1948 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 49 27  4 12 |  a = kleiner65
 23 10  1  1 |  b = groter65
  6  1 10  4 |  c = kleiner35
 22  4  4 14 |  d = kleiner50
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         137               71.3542 %
Incorrectly Classified Instances        55               28.6458 %
Kappa statistic                          0.5763
Mean absolute error                      0.1631
Root mean squared error                  0.3583
Relative absolute error                 48.3408 %
Root relative squared error             87.3417 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 73  9  0 10 |  a = kleiner65
  7 26  1  1 |  b = groter65
  2  0 12  7 |  c = kleiner35
  8  1  9 26 |  d = kleiner50
```

- **J48**
  - ***Voor het begin van de studies***

```
Correctly Classified Instances          91               47.3958 %
Incorrectly Classified Instances       101               52.6042 %
Kappa statistic                         -0.005
Mean absolute error                      0.337
Root mean squared error                  0.4122
Relative absolute error                 99.9127 %
Root relative squared error            100.4774 %
Total Number of Instances              192
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 91  0  0  1 |  a = kleiner65
 34  0  0  1 |  b = groter65
 21  0  0  0 |  c = kleiner35
 44  0  0  0 |  d = kleiner50

: kleiner65 (192.0/100.0)
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         112               58.3333 %
Incorrectly Classified Instances        80               41.6667 %
Kappa statistic                          0.3524
Mean absolute error                      0.2435
Root mean squared error                  0.3756
Relative absolute error                 72.1983 %
Root relative squared error             91.5619 %
Total Number of Instances              192


=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 70 14  0  8 |  a = kleiner65
 19 16  0  0 |  b = groter65
  4  0  2 15 |  c = kleiner35
 14  1  5 24 |  d = kleiner50


Wiskunde <= 3: kleiner50 (47.48/22.24)
Wiskunde > 3
|   Macro-economie <= 12
|   |   SO_STUDIEVERTRAGING = nee: kleiner65 (51.56/14.81)
|   |   SO_STUDIEVERTRAGING = ja: kleiner50 (6.05/2.05)
|   Macro-economie > 12
|   |   Sociologie en demografie <= 14: kleiner65 (67.71/20.71)
|   |   Sociologie en demografie > 14: groter65 (19.2/4.0)
```

- **IBk**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances          70               36.4583 %
Incorrectly Classified Instances       122               63.5417 %
Kappa statistic                          0.0406
Mean absolute error                      0.319
Root mean squared error                  0.5573
Relative absolute error                 94.5701 %
Root relative squared error            135.857  %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 48 26  6 12 |  a = kleiner65
 23  5  1  6 |  b = groter65
  9  1  5  6 |  c = kleiner35
 19  8  5 12 |  d = kleiner50
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances          78               40.625  %
Incorrectly Classified Instances       114               59.375  %
Kappa statistic                          0.1004
Mean absolute error                      0.2986
Root mean squared error                  0.5388
Relative absolute error                 88.5336 %
Root relative squared error            131.3285 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 49 22  5 16 |  a = kleiner65
 21  8  1  5 |  b = groter65
  8  0  4  9 |  c = kleiner35
 20  3  4 17 |  d = kleiner50
```

- **MultilayerPerceptron**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances          88               45.8333 %
Incorrectly Classified Instances       104               54.1667 %
Kappa statistic                          0.1675
Mean absolute error                      0.285
Root mean squared error                  0.4863
Relative absolute error                 84.5028 %
Root relative squared error            118.554  %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 59 14  2 17 |  a = kleiner65
 19 10  3  3 |  b = groter65
  7  2  5  7 |  c = kleiner35
 20  6  4 14 |  d = kleiner50
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         106               55.2083 %
Incorrectly Classified Instances        86               44.7917 %
Kappa statistic                          0.3247
Mean absolute error                      0.2225
Root mean squared error                  0.4373
Relative absolute error                 65.9611 %
Root relative squared error            106.5928 %
Total Number of Instances              192


=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 60 17  0 15 |  a = kleiner65
 17 18  0  0 |  b = groter65
  3  0  9  9 |  c = kleiner35
 17  1  7 19 |  d = kleiner50
```

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
Percentage	vooraanvang	42	43	47	36	46
	naresult1	51	71	58	41	55

4) Voorspellen van slagen op bepaalde vakken in het eerste jaar

1) Wiskunde

- **OneR Classifier**
  - ***Voor het begin van de studies***

```
Correctly Classified Instances          99               56.5714 %
Incorrectly Classified Instances        76               43.4286 %
Kappa statistic                          0.0243
Mean absolute error                      0.2895
Root mean squared error                  0.5381
Relative absolute error                 80.9452 %
Root relative squared error            127.5119 %
Total Number of Instances              175
Ignored Class Unknown Instances                 17

=== Confusion Matrix ===

  a  b  c   <-- classified as
 90  4 13 |  a = geslaagd
 15  2  1 |  b = kleinder10
 43  0  7 |  c = kleinder8
```

OneR beslissingsvariabele : secundaire school

- ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         144               82.2857 %
Incorrectly Classified Instances        31               17.7143 %
Kappa statistic                          0.6293
Mean absolute error                      0.1181
Root mean squared error                  0.3436
Relative absolute error                 33.0171 %
Root relative squared error             81.4376 %
Total Number of Instances              175
Ignored Class Unknown Instances                 17

=== Confusion Matrix ===

   a   b   c   <-- classified as
 105   0   2 |   a = geslaagd
  11   0   7 |   b = kleinder10
  11   0  39 |   c = kleinder8
```

OneR beslissingsvariabele :wiskunde

- **Naive Bayes**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances         109               62.2857 %
Incorrectly Classified Instances        66               37.7143 %
Kappa statistic                          0.2475
Mean absolute error                      0.2574
Root mean squared error                  0.4422
Relative absolute error                 71.9656 %
Root relative squared error            104.7873 %
Total Number of Instances              175
Ignored Class Unknown Instances                 17

=== Confusion Matrix ===

  a  b  c   <-- classified as
 85  2 20 |  a = geslaagd
 12  1  5 |  b = kleinder10
 22  5 23 |  c = kleinder8
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         139               79.4286 %
Incorrectly Classified Instances        36               20.5714 %
Kappa statistic                          0.6052
Mean absolute error                      0.1393
Root mean squared error                  0.3388
Relative absolute error                 38.9587 %
Root relative squared error             80.283  %
Total Number of Instances              175
Ignored Class Unknown Instances                 17

=== Confusion Matrix ===

  a  b  c   <-- classified as
 98  3  6 |  a = geslaagd
  7  2  9 |  b = kleinder10
  5  6 39 |  c = kleinder8
```

- **J48**
  - ***Voor het begin van de studies***

```
Correctly Classified Instances         123               70.2857 %
Incorrectly Classified Instances        52               29.7143 %
Kappa statistic                          0.3947
Mean absolute error                      0.2954
Root mean squared error                  0.3921
Relative absolute error                 82.5796 %
Root relative squared error             92.9305 %
Total Number of Instances              175
Ignored Class Unknown Instances                 17

=== Confusion Matrix ===

  a  b  c   <-- classified as
 91  0 16 |  a = geslaagd
 11  0  7 |  b = kleinder10
 18  0 32 |  c = kleinder8


SO_STUDIEVERTRAGING = nee
|   SOENGELS = 2tot2uur: geslaagd (122.57/29.0)
|   SOENGELS = 3ofmeeruur: kleinder8 (33.43/16.43)
SO_STUDIEVERTRAGING = ja: kleinder8 (19.0/5.0)
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         141               80.5714 %
Incorrectly Classified Instances        34               19.4286 %
Kappa statistic                          0.6006
Mean absolute error                      0.191
Root mean squared error                  0.3383
Relative absolute error                 53.4027 %
Root relative squared error             80.167  %
Total Number of Instances              175
Ignored Class Unknown Instances                 17

=== Confusion Matrix ===

   a   b   c   <-- classified as
 102   1   4 |   a = geslaagd
  10   0   8 |   b = kleinder10
  11   0  39 |   c = kleinder8
```

```
Wiskunde <= 7
|   Wiskunde <= 3
|   |   Sociologie en demografie <= 12: kleinder8 (32.0/1.0)
|   |   Sociologie en demografie > 12
|   |   |   Macro-economie <= 10: kleinder10 (2.21/0.21)
|   |   |   Macro-economie > 10: kleinder8 (2.0)
|   Wiskunde > 3: geslaagd (48.28/29.0)
Wiskunde > 7: geslaagd (90.52/3.0)
```

- **IBk**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances          96               54.8571 %
Incorrectly Classified Instances        79               45.1429 %
Kappa statistic                          0.1186
Mean absolute error                      0.3036
Root mean squared error                  0.5435
Relative absolute error                 84.8899 %
Root relative squared error            128.8    %
Total Number of Instances              175
Ignored Class Unknown Instances                 17

=== Confusion Matrix ===

  a  b  c   <-- classified as
 77 11 19 |  a = geslaagd
 11  1  6 |  b = kleinder10
 29  3 18 |  c = kleinder8
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         106               60.5714 %
Incorrectly Classified Instances        69               39.4286 %
Kappa statistic                          0.2366
Mean absolute error                      0.2663
Root mean squared error                  0.508
Relative absolute error                 74.4383 %
Root relative squared error            120.3746 %
Total Number of Instances              175
Ignored Class Unknown Instances                 17

=== Confusion Matrix ===

  a  b  c   <-- classified as
 80  8 19 |  a = geslaagd
 11  1  6 |  b = kleinder10
 22  3 25 |  c = kleinder8
```

- **MultilayerPerceptron**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances         103               58.8571 %
Incorrectly Classified Instances        72               41.1429 %
Kappa statistic                          0.1698
Mean absolute error                      0.2709
Root mean squared error                  0.4887
Relative absolute error                 75.733  %
Root relative squared error            115.8054 %
Total Number of Instances              175
Ignored Class Unknown Instances                 17

=== Confusion Matrix ===

  a  b  c   <-- classified as
 82  5 20 |  a = geslaagd
 14  0  4 |  b = kleinder10
 26  3 21 |  c = kleinder8
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         135               77.1429 %
Incorrectly Classified Instances        40               22.8571 %
Kappa statistic                          0.5575
Mean absolute error                      0.158
Root mean squared error                  0.3599
Relative absolute error                 44.1732 %
Root relative squared error             85.2913 %
Total Number of Instances              175
Ignored Class Unknown Instances                 17

=== Confusion Matrix ===

  a  b  c   <-- classified as
 98  4  5 |  a = geslaagd
  8  1  9 |  b = kleinder10
  7  7 36 |  c = kleinder8
```

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
Wiskunde	vooraanvang	57	62	70	55	59
	naresult1	82	79	81	61	77

2) Financieel boekhouden

- **OneR Classifier**
  - ***Voor het begin van de studies***

```
Correctly Classified Instances          76               46.0606 %
Incorrectly Classified Instances        89               53.9394 %
Kappa statistic                         -0.0504
Mean absolute error                      0.3596
Root mean squared error                  0.5997
Relative absolute error                113.2726 %
Root relative squared error            151.0852 %
Total Number of Instances              165
Ignored Class Unknown Instances                 27

=== Confusion Matrix ===

  a  b  c   <-- classified as
  0 10  2 |  a = kleinder10
 32 71  9 |  b = geslaagd
  9 27  5 |  c = kleinder8
```

OneR beslissingsvariabele : gemeente van afkomst

- ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         135               81.8182 %
Incorrectly Classified Instances        30               18.1818 %
Kappa statistic                          0.602
Mean absolute error                      0.1212
Root mean squared error                  0.3482
Relative absolute error                 38.1818 %
Root relative squared error             87.7177 %
Total Number of Instances              165
Ignored Class Unknown Instances                 27

=== Confusion Matrix ===

   a   b   c   <-- classified as
   0   3   9 |   a = kleinder10
   0 102  10 |   b = geslaagd
   0   8  33 |   c = kleinder8
```

OneR beslissingsvariabele: boekhouden

- **Naive Bayes**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances         100               60.6061 %
Incorrectly Classified Instances        65               39.3939 %
Kappa statistic                          0.1794
Mean absolute error                      0.2717
Root mean squared error                  0.463
Relative absolute error                 85.5782 %
Root relative squared error            116.6551 %
Total Number of Instances              165
Ignored Class Unknown Instances                 27

=== Confusion Matrix ===

  a  b  c   <-- classified as
  0  7  5 |  a = kleinder10
  9 82 21 |  b = geslaagd
  3 20 18 |  c = kleinder8
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         129               78.1818 %
Incorrectly Classified Instances        36               21.8182 %
Kappa statistic                          0.5241
Mean absolute error                      0.1524
Root mean squared error                  0.35
Relative absolute error                 48.0089 %
Root relative squared error             88.1927 %
Total Number of Instances              165
Ignored Class Unknown Instances                 27

=== Confusion Matrix ===

  a  b  c   <-- classified as
  1  8  3 |  a = kleinder10
  2 97 13 |  b = geslaagd
  1  9 31 |  c = kleinder8
```

- **J48**
  - ***Voor het begin van de studies***

```
Correctly Classified Instances         108               65.4545 %
Incorrectly Classified Instances        57               34.5455 %
Kappa statistic                          0.0938
Mean absolute error                      0.2961
Root mean squared error                  0.4087
Relative absolute error                 93.2813 %
Root relative squared error            102.9681 %
Total Number of Instances              165
Ignored Class Unknown Instances                 27

=== Confusion Matrix ===

   a   b   c   <-- classified as
   0   9   3 |   a = kleinder10
   0 100  12 |   b = geslaagd
   0  33   8 |   c = kleinder8
```

```
NATIONALITEIT = NATIONALITEIT3
|   SOFRANS = 4ofmeeruur
|   |   SO_STUDIEVERTRAGING = nee: geslaagd (36.58/17.29)
|   |   SO_STUDIEVERTRAGING = ja: kleinder8 (9.0/4.0)
|   SOFRANS = 0tot3uur: geslaagd (112.42/21.71)
NATIONALITEIT = NATIONALITEIT16: kleinder8 (1.0)
NATIONALITEIT = NATIONALITEIT14: kleinder8 (4.0/1.0)
NATIONALITEIT = NATIONALITEIT4: geslaagd (0.0)
NATIONALITEIT = NATIONALITEIT7: kleinder8 (1.0)
NATIONALITEIT = NATIONALITEIT9: kleinder8 (1.0)
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         134               81.2121 %
Incorrectly Classified Instances        31               18.7879 %
Kappa statistic                          0.595
Mean absolute error                      0.1514
Root mean squared error                  0.3115
Relative absolute error                 47.6958 %
Root relative squared error             78.4729 %
Total Number of Instances              165
Ignored Class Unknown Instances                 27
```

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
   5   6   1 |   a = kleinder10
   5 101   6 |   b = geslaagd
   4   9  28 |   c = kleinder8
```

```
Financieel boekhouden <= 7
|   Wiskunde <= 5
|   |   BEURSSTUDENT = BEURSSTUDENT2: kleinder8 (24.57)
|   |   BEURSSTUDENT = BEURSSTUDENT1
|   |   |   Sociologie en demografie <= 10: kleinder8 (4.0)
|   |   |   Sociologie en demografie > 10: geslaagd (2.0)
|   Wiskunde > 5
|   |   Macro-economie <= 12: kleinder8 (14.43/9.0)
|   |   Macro-economie > 12: kleinder10 (9.0/4.0)
Financieel boekhouden > 7
|   Wiskunde <= 4
|   |   OP_KOT = OP_KOT1: geslaagd (4.0)
|   |   OP_KOT = OP_KOT2
|   |   |   SO_RICHTING = SO_RICHTING37: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING28: geslaagd (0.14)
|   |   |   SO_RICHTING = SO_RICHTING34: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING23: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING24: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING9: kleinder8 (5.0/1.0)
|   |   |   SO_RICHTING = SO_RICHTING7: geslaagd (5.0/1.0)
|   |   |   SO_RICHTING = SO_RICHTING29: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING30: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING22: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING12: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING32: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING4: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING1: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING16: geslaagd (0.0)
|   |   |   SO_RICHTING = SO_RICHTING3: kleinder8 (1.0)
|   |   |   SO_RICHTING = SO_RICHTING27: geslaagd (0.0)
|   Wiskunde > 4: geslaagd (95.86/3.0)
```

- **IBk**
  - ***Voor het begin van de studies***

```
Correctly Classified Instances          89               53.9394 %
Incorrectly Classified Instances        76               46.0606 %
Kappa statistic                          0.0552
Mean absolute error                      0.3098
Root mean squared error                  0.5487
Relative absolute error                 97.5843 %
Root relative squared error            138.2462 %
Total Number of Instances              165
Ignored Class Unknown Instances                27
```

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
  2  6  4 |  a = kleinder10
 14 76 22 |  b = geslaagd
  3 27 11 |  c = kleinder8
```

- ***Na resultaten 1ste trimester***

```
Correctly Classified Instances          91               55.1515 %
Incorrectly Classified Instances        74               44.8485 %
Kappa statistic                          0.0648
Mean absolute error                      0.3019
Root mean squared error                  0.5414
Relative absolute error                 95.0888 %
Root relative squared error            136.4153 %
Total Number of Instances              165
Ignored Class Unknown Instances                27

=== Confusion Matrix ===

  a  b  c   <-- classified as
  1  7  4 |  a = kleinder10
 11 77 24 |  b = geslaagd
  2 26 13 |  c = kleinder8
```

➢ **MultilayerPerceptron**

- ***Voor het begin van de studies***

```
Correctly Classified Instances          92               55.7576 %
Incorrectly Classified Instances        73               44.2424 %
Kappa statistic                          0.0629
Mean absolute error                      0.2954
Root mean squared error                  0.5101
Relative absolute error                 93.0618 %
Root relative squared error            128.5245 %
Total Number of Instances              165
Ignored Class Unknown Instances                27

=== Confusion Matrix ===

  a  b  c   <-- classified as
  0  7  5 |  a = kleinder10
  5 78 29 |  b = geslaagd
  2 25 14 |  c = kleinder8
```

- ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         120               72.7273 %
Incorrectly Classified Instances        45               27.2727 %
Kappa statistic                          0.4362
Mean absolute error                      0.183
Root mean squared error                  0.3912
Relative absolute error                 57.6323 %
Root relative squared error             98.5594 %
Total Number of Instances              165
Ignored Class Unknown Instances                 27

=== Confusion Matrix ===

  a  b  c   <-- classified as
  1  7  4 |  a = kleinder10
 11 90 11 |  b = geslaagd
  1 11 29 |  c = kleinder8
```

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
Boekhouden	vooraanvang	46	61	65	54	65
	naresult1	82	78	81	55	73

3) Statistiek

➢ **OneR Classifier**

- ***Voor het begin van de studies***

```
Correctly Classified Instances         119               70.8333 %
Incorrectly Classified Instances        49               29.1667 %
Kappa statistic                         -0.0075
Mean absolute error                      0.1944
Root mean squared error                  0.441
Relative absolute error                 75.773  %
Root relative squared error            123.9397 %
Total Number of Instances              168
Ignored Class Unknown Instances                 24
```

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 117  11   0 |   a = geslaagd
  30   2   1 |   b = kleinder8
   6   1   0 |   c = kleinder10
```

OneR beslissingsvariabele gemeente van afkomst

- ***<u>Na resultaten 1ste trimester</u>***

```
Correctly Classified Instances         137               81.5476 %
Incorrectly Classified Instances        31               18.4524 %
Kappa statistic                          0.4213
Mean absolute error                      0.123
Root mean squared error                  0.3507
Relative absolute error                 47.938  %
Root relative squared error             98.581  %
Total Number of Instances              168
Ignored Class Unknown Instances                 24

=== Confusion Matrix ===

   a   b   c   <-- classified as
 122   6   0 |   a = geslaagd
  18  15   0 |   b = kleinder8
   4   3   0 |   c = kleinder10
```

OneR beslissingsvariabele : wiskunde

- **<u>Naive Bayes</u>**
  - ***<u>Voor het begin van de studies</u>***

```
Correctly Classified Instances         121               72.0238 %
Incorrectly Classified Instances        47               27.9762 %
Kappa statistic                          0.2751
Mean absolute error                      0.1979
Root mean squared error                  0.3925
Relative absolute error                 77.1023 %
Root relative squared error            110.3262 %
Total Number of Instances              168
Ignored Class Unknown Instances                 24
```

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 107  18   3 |   a = geslaagd
  15  14   4 |   b = kleinder8
   4   3   0 |   c = kleinder10
```

- ***<u>Na resultaten 1ste trimester</u>***

```
Correctly Classified Instances         134               79.7619 %
Incorrectly Classified Instances        34               20.2381 %
Kappa statistic                          0.4878
Mean absolute error                      0.1313
Root mean squared error                  0.3288
Relative absolute error                 51.1665 %
Root relative squared error             92.4156 %
Total Number of Instances              168
Ignored Class Unknown Instances                24
```

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 112  14   2 |   a = geslaagd
   9  21   3 |   b = kleinder8
   2   4   1 |   c = kleinder10
```

➢ **<u>J48</u>**

- ***<u>Voor het begin van de studies</u>***

```
Correctly Classified Instances         129               76.7857 %
Incorrectly Classified Instances        39               23.2143 %
Kappa statistic                          0.1764
Mean absolute error                      0.2346
Root mean squared error                  0.3503
Relative absolute error                 91.4266 %
Root relative squared error             98.4509 %
Total Number of Instances              168
Ignored Class Unknown Instances                24
```

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 123   5   0 |   a = geslaagd
  27   6   0 |   b = kleinder8
   5   2   0 |   c = kleinder10


SO_STUDIEVERTRAGING = nee: geslaagd (152.0/29.0)
SO_STUDIEVERTRAGING = ja: kleinder8 (16.0/7.0)
```

- ***<u>Na resultaten 1ste trimester</u>***

```
Correctly Classified Instances         137               81.5476 %
Incorrectly Classified Instances        31               18.4524 %
Kappa statistic                          0.4678
Mean absolute error                      0.1614
Root mean squared error                  0.3287
Relative absolute error                 62.8973 %
Root relative squared error             92.4005 %
Total Number of Instances              168
Ignored Class Unknown Instances                 24

=== Confusion Matrix ===

   a   b   c   <-- classified as
 119   8   1 |   a = geslaagd
  15  18   0 |   b = kleinder8
   2   5   0 |   c = kleinder10


Macro-economie <= 8: kleinder8 (25.0/7.0)
Macro-economie > 8
|   Wiskunde <= 3
|   |   BEURSSTUDENT = BEURSSTUDENT2: kleinder8 (9.2/2.1)
|   |   BEURSSTUDENT = BEURSSTUDENT1: geslaagd (5.0/1.0)
|   Wiskunde > 3
|   |   Financieel boekhouden <= 7
|   |   |   SOFRANS = 4ofmeeruur
|   |   |   |   Geo-economie en economische geschiedenis <= 11: kleinder8 (4.9/1.0)
|   |   |   |   Geo-economie en economische geschiedenis > 11: kleinder10 (3.33/1.33)
|   |   |   SOFRANS = 0tot3uur: geslaagd (16.67/3.0)
|   |   Financieel boekhouden > 7: geslaagd (103.9/1.0)
```

- **IBk**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances         116               69.0476 %
Incorrectly Classified Instances        52               30.9524 %
Kappa statistic                          0.1883
Mean absolute error                      0.211
Root mean squared error                  0.4499
Relative absolute error                 82.2184 %
Root relative squared error            126.4558 %
Total Number of Instances              168
Ignored Class Unknown Instances                 24

=== Confusion Matrix ===

   a   b   c   <-- classified as
 103  24   1 |   a = geslaagd
  19  13   1 |   b = kleinder8
   4   3   0 |   c = kleinder10
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         119               70.8333 %
Incorrectly Classified Instances        49               29.1667 %
Kappa statistic                          0.2283
Mean absolute error                      0.1993
Root mean squared error                  0.4367
Relative absolute error                 77.6696 %
Root relative squared error            122.7556 %
Total Number of Instances              168
Ignored Class Unknown Instances                 24
=== Confusion Matrix ===

   a   b   c   <-- classified as
 105  22   1 |   a = geslaagd
  18  14   1 |   b = kleinder8
   4   3   0 |   c = kleinder10
```

- **MultilayerPerceptron**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances         120               71.4286 %
Incorrectly Classified Instances        48               28.5714 %
Kappa statistic                          0.22
Mean absolute error                      0.1924
Root mean squared error                  0.4081
Relative absolute error                 74.9665 %
Root relative squared error            114.7146 %
Total Number of Instances              168
Ignored Class Unknown Instances                 24

=== Confusion Matrix ===

   a   b   c   <-- classified as
 108  18   2 |   a = geslaagd
  19  12   2 |   b = kleinder8
   4   3   0 |   c = kleinder10
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         136               80.9524 %
Incorrectly Classified Instances        32               19.0476 %
Kappa statistic                          0.4388
Mean absolute error                      0.1327
Root mean squared error                  0.3265
Relative absolute error                 51.6941 %
Root relative squared error             91.7613 %
Total Number of Instances              168
Ignored Class Unknown Instances                 24

=== Confusion Matrix ===

   a   b   c   <-- classified as
 121   6   1 |   a = geslaagd
  15  15   3 |   b = kleinder8
   3   4   0 |   c = kleinder10
```

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
Statistiek	vooraanvang	71	72	77	69	71
	naresult1	82	80	82	71	81

5) <u>Voorspellen of een student zijn bachelordiploma zal halen.</u>

- **<u>OneR Classifier</u>**
  - ***<u>Voor het begin van de studies</u>***

```
Correctly Classified Instances         126               65.625  %
Incorrectly Classified Instances        66               34.375  %
Kappa statistic                          0.1654
Mean absolute error                      0.3438
Root mean squared error                  0.5863
Relative absolute error                 75.0474 %
Root relative squared error            122.5644 %
Total Number of Instances              192

=== Confusion Matrix ===

   a   b   <-- classified as
 106  18 |   a = wel
  48  20 |   b = niet behaald
```

OneR beslissingsvariabele : secundiaire school

  - ***<u>Na resultaten 1ste trimester</u>***

```
Correctly Classified Instances         144               75      %
Incorrectly Classified Instances        48               25      %
Kappa statistic                          0.4642
Mean absolute error                      0.25
Root mean squared error                  0.5
Relative absolute error                 54.5799 %
Root relative squared error            104.5232 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b   <-- classified as
 97 27 |  a = wel
 21 47 |  b = niet behaald
```

OneR beslissingsvariabele : financieel boekhouden

- ***Na het eerste jaar***

```
Correctly Classified Instances         170               88.5417 %
Incorrectly Classified Instances        22               11.4583 %
Kappa statistic                          0.7444
Mean absolute error                      0.1146
Root mean squared error                  0.3385
Relative absolute error                 25.0158 %
Root relative squared error             70.7626 %
Total Number of Instances              192

=== Confusion Matrix ===

   a   b   <-- classified as
 116   8 |   a = wel
  14  54 |   b = niet behaald
```

OneR beslissingsvariabele : # onvoldoendes

- **Naive Bayes**
  - ***Voor het begin van de studies***

```
Correctly Classified Instances         140               72.9167 %
Incorrectly Classified Instances        52               27.0833 %
Kappa statistic                          0.3791
Mean absolute error                      0.2933
Root mean squared error                  0.4782
Relative absolute error                 64.0292 %
Root relative squared error             99.9618 %
Total Number of Instances              192

=== Confusion Matrix ===

   a   b   <-- classified as
 105  19 |   a = wel
  33  35 |   b = niet behaald
```

- ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         160               83.3333 %
Incorrectly Classified Instances        32               16.6667 %
Kappa statistic                          0.6381
Mean absolute error                      0.1703
Root mean squared error                  0.3742
Relative absolute error                 37.1842 %
Root relative squared error             78.215  %
Total Number of Instances              192
```

```
=== Confusion Matrix ===

   a   b   <-- classified as
 107  17 |   a = wel
  15  53 |   b = niet behaald
```

- ***Na het eerste jaar***

```
Correctly Classified Instances         169               88.0208 %
Incorrectly Classified Instances        23               11.9792 %
Kappa statistic                          0.7441
Mean absolute error                      0.1165
Root mean squared error                  0.3332
Relative absolute error                 25.4444 %
Root relative squared error             69.6509 %
Total Number of Instances              192

=== Confusion Matrix ===

   a   b   <-- classified as
 109  15 |   a = wel
   8  60 |   b = niet behaald
```

➢ **J48**

- ***Voor het begin van de studies***

```
Correctly Classified Instances         135               70.3125 %
Incorrectly Classified Instances        57               29.6875 %
Kappa statistic                          0.2712
Mean absolute error                      0.3923
Root mean squared error                  0.4647
Relative absolute error                 85.6388 %
Root relative squared error             97.1445 %
Total Number of Instances              192

   a   b   <-- classified as
 112  12 |   a = wel
  45  23 |   b = niet behaald
```

```
J48 pruned tree
------------------

SO_STUDIEVERTRAGING = nee
|   BEURSSTUDENT = BEURSSTUDENT2
|   |   SOWISK = 8uur: wel (18.0/3.0)
|   |   SOWISK = 6uur: wel (85.0/25.0)
|   |   SOWISK = 4/5uur
|   |   |   OP_KOT = OP_KOT1: niet behaald (4.0)
|   |   |   OP_KOT = OP_KOT2: wel (6.0/2.0)
|   |   SOWISK = 0tot3uur: niet behaald (15.0/2.0)
|   BEURSSTUDENT = BEURSSTUDENT1: wel (39.0/2.0)
SO_STUDIEVERTRAGING = ja: niet behaald (25.0/6.0)
```

- ***<u>Na resultaten 1ste trimester</u>***

```
Correctly Classified Instances         154               80.2083 %
Incorrectly Classified Instances        38               19.7917 %
Kappa statistic                          0.5494
Mean absolute error                      0.2472
Root mean squared error                  0.416
Relative absolute error                 53.9769 %
Root relative squared error             86.968  %
Total Number of Instances              192

=== Confusion Matrix ===

   a   b   <-- classified as
 111  13 |   a = wel
  25  43 |   b = niet behaald

J48 pruned tree
------------------

Financieel boekhouden <= 7
|   Macro-economie <= 11
|   |   BEURSSTUDENT = BEURSSTUDENT2
|   |   |   Wiskunde <= 8: niet behaald (45.37/1.0)
|   |   |   Wiskunde > 8: wel (5.28/1.28)
|   |   BEURSSTUDENT = BEURSSTUDENT1
|   |   |   SO_STUDIEVERTRAGING = nee: wel (7.0/2.0)
|   |   |   SO_STUDIEVERTRAGING = ja: niet behaald (4.0)
|   Macro-economie > 11
|   |   GESLACHT = GESLACHT1: wel (13.08/2.08)
|   |   GESLACHT = GESLACHT2: niet behaald (3.08/1.0)
Financieel boekhouden > 7
|   GENERATIESTUDENT = nee
|   |   BEURSSTUDENT = BEURSSTUDENT2
|   |   |   SONAT = 3ofmeeruur: niet behaald (0.0)
|   |   |   SONAT = 2uur: wel (4.0/1.0)
|   |   |   SONAT = 0/1uur: niet behaald (4.19)
|   |   BEURSSTUDENT = BEURSSTUDENT1: wel (4.0)
|   GENERATIESTUDENT = ja: wel (102.0/7.0)
```

- ***Na het eerste jaar***

```
Correctly Classified Instances         171               89.0625 %
Incorrectly Classified Instances        21               10.9375 %
Kappa statistic                          0.7552
Mean absolute error                      0.1311
Root mean squared error                  0.3252
Relative absolute error                 28.6254 %
Root relative squared error             67.9878 %
Total Number of Instances              192

=== Confusion Matrix ===

   a   b   <-- classified as
 117   7 |   a = wel
  14  54 |   b = niet behaald

J48 pruned tree
------------------

onvoldoendes <= 4
|   Financieel boekhouden  EIND <= 7
|   |   Talen  EIND <= 10: wel (3.0)
|   |   Talen  EIND > 10
|   |   |   STPTN_BEHAALD = 54<= X < 60: niet behaald (0.0)
|   |   |   STPTN_BEHAALD = 60.0: niet behaald (0.0)
|   |   |   STPTN_BEHAALD = <30: niet behaald (0.0)
|   |   |   STPTN_BEHAALD = 30<= X < 48: niet behaald (5.25/0.17)
|   |   |   STPTN_BEHAALD = 48<= X < 54: wel (3.0/1.0)
|   Financieel boekhouden  EIND > 7: wel (123.75/6.92)
onvoldoendes > 4
|   BEURSSTUDENT = BEURSSTUDENT2: niet behaald (51.0)
|   BEURSSTUDENT = BEURSSTUDENT1
|   |   Financieel boekhouden <= 5: niet behaald (4.0)
|   |   Financieel boekhouden > 5: wel (2.0)
```

- **IBk**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances         121               63.0208 %
Incorrectly Classified Instances        71               36.9792 %
Kappa statistic                          0.1667
Mean absolute error                      0.3713
Root mean squared error                  0.6046
Relative absolute error                 81.0579 %
Root relative squared error            126.3984 %
Total Number of Instances              192
```

```
=== Confusion Matrix ===

  a  b   <-- classified as
 93 31 |  a = wel
 40 28 |  b = niet behaald
```

- ***<u>Na resultaten 1ste trimester</u>***

```
Correctly Classified Instances         124               64.5833 %
Incorrectly Classified Instances        68               35.4167 %
Kappa statistic                          0.2101
Mean absolute error                      0.3558
Root mean squared error                  0.5917
Relative absolute error                 77.6857 %
Root relative squared error            123.6994 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b   <-- classified as
 93 31 |  a = wel
 37 31 |  b = niet behaald
```

- ***<u>Na het eerste jaar</u>***

```
Correctly Classified Instances         164               85.4167 %
Incorrectly Classified Instances        28               14.5833 %
Kappa statistic                          0.6833
Mean absolute error                      0.1499
Root mean squared error                  0.3797
Relative absolute error                 32.7232 %
Root relative squared error             79.3822 %
Total Number of Instances              192

=== Confusion Matrix ===

   a   b   <-- classified as
 109  15 |   a = wel
  13  55 |   b = niet behaald
```

- **MultilayerPerceptron**

  - ***Voor het begin van de studies***

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         136               70.8333 %
Incorrectly Classified Instances        56               29.1667 %
Kappa statistic                          0.3789
Mean absolute error                      0.3113
Root mean squared error                  0.5106
Relative absolute error                 67.9573 %
Root relative squared error            106.7325 %
Total Number of Instances              192
=== Confusion Matrix ===

  a  b   <-- classified as
 92 32 |  a = wel
 24 44 |  b = niet behaald
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         153               79.6875 %
Incorrectly Classified Instances        39               20.3125 %
Kappa statistic                          0.5485
Mean absolute error                      0.2015
Root mean squared error                  0.4304
Relative absolute error                 43.9922 %
Root relative squared error             89.9793 %
Total Number of Instances              192

=== Confusion Matrix ===

   a   b   <-- classified as
 107  17 |   a = wel
  22  46 |   b = niet behaald
```

- ***Na het eerste jaar***

```
Correctly Classified Instances         168               87.5    %
Incorrectly Classified Instances        24               12.5    %
Kappa statistic                          0.7249
Mean absolute error                      0.119
Root mean squared error                  0.3275
Relative absolute error                 25.9876 %
Root relative squared error             68.4559 %
Total Number of Instances              192

=== Confusion Matrix ===

   a   b   <-- classified as
 113  11 |   a = wel
  13  55 |   b = niet behaald
```

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
Bachelor 1	vooraanvang	66	73	70	63	71
	naresult1	75	83	80	65	80
	eindejaar	89	88	89	85	88

➢ **OneR Classifier**

- ***Voor het begin van de studies***

```
Correctly Classified Instances          81               42.1875 %
Incorrectly Classified Instances       111               57.8125 %
Kappa statistic                          0.1019
Mean absolute error                      0.3854
Root mean squared error                  0.6208
Relative absolute error                 90.9776 %
Root relative squared error            134.9301 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
 10 20  8 |  a = meer
 24 50 12 |  b = normaal
 21 26 21 |  c = niet behaald
```

OneR beslissingsvariabele: secundaire school

- ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         138               71.875  %
Incorrectly Classified Instances        54               28.125  %
Kappa statistic                          0.5238
Mean absolute error                      0.1875
Root mean squared error                  0.433
Relative absolute error                 44.2594 %
Root relative squared error             94.1118 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
  0 21 17 |  a = meer
  0 81  5 |  b = normaal
  0 11 57 |  c = niet behaald
```

OneR beslissingsvariabele : Financieel boekhouden

- ***Na het eerste jaar***

```
Correctly Classified Instances         157               81.7708 %
Incorrectly Classified Instances        35               18.2292 %
Kappa statistic                          0.704
Mean absolute error                      0.1215
Root mean squared error                  0.3486
Relative absolute error                 28.6866 %
Root relative squared error             75.7672 %
Total Number of Instances              192


=== Confusion Matrix ===

  a  b  c   <-- classified as
 16 15  7 |  a = meer
  2 84  0 |  b = normaal
  7  4 57 |  c = niet behaald
```

OneR beslissingsvariabele : # onvoldoendes

- **Naive Bayes**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances         105               54.6875 %
Incorrectly Classified Instances        87               45.3125 %
Kappa statistic                          0.2592
Mean absolute error                      0.316
Root mean squared error                  0.4766
Relative absolute error                 74.5873 %
Root relative squared error            103.5828 %
Total Number of Instances              192


=== Confusion Matrix ===

  a  b  c   <-- classified as
  6 21 11 |  a = meer
  8 65 13 |  b = normaal
 10 24 34 |  c = niet behaald
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         126               65.625  %
Incorrectly Classified Instances        66               34.375  %
Kappa statistic                          0.4568
Mean absolute error                      0.2331
Root mean squared error                  0.4168
Relative absolute error                 55.0271 %
Root relative squared error             90.5968 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
 12 16 10 |  a = meer
 18 64  4 |  b = normaal
  8 10 50 |  c = niet behaald
```

- ***Na het eerste jaar***

```
Correctly Classified Instances         154               80.2083 %
Incorrectly Classified Instances        38               19.7917 %
Kappa statistic                          0.6865
Mean absolute error                      0.1316
Root mean squared error                  0.3534
Relative absolute error                 31.06   %
Root relative squared error             76.8042 %
Total Number of Instances              192
```

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 21 11  6 |  a = meer
  8 78  0 |  b = normaal
  9  4 55 |  c = niet behaald
```

➢ **J48**

- ***Voor het begin van de studies***

```
Correctly Classified Instances         106               55.2083 %
Incorrectly Classified Instances        86               44.7917 %
Kappa statistic                          0.2222
Mean absolute error                      0.3755
Root mean squared error                  0.4414
Relative absolute error                 88.6306 %
Root relative squared error             95.9362 %
Total Number of Instances              192
```

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
  1 32  5 |  a = meer
  0 79  7 |  b = normaal
  3 39 26 |  c = niet behaald
```

```
J48 pruned tree
------------------

SO_STUDIEVERTRAGING = nee
|   BEURSSTUDENT = BEURSSTUDENT2
|   |   SODUITS = 0uur: normaal (42.32/22.99)
|   |   SODUITS = 1uur: normaal (38.19/13.9)
|   |   SODUITS = 2uur: normaal (23.74/9.56)
|   |   SODUITS = 3ofmeeruur: niet behaald (23.74/8.37)
|   BEURSSTUDENT = BEURSSTUDENT1: normaal (39.0/18.0)
SO_STUDIEVERTRAGING = ja
|   BEURSSTUDENT = BEURSSTUDENT2: niet behaald (17.0/2.0)
|   BEURSSTUDENT = BEURSSTUDENT1
|   |   SOWISK = 8uur: meer (0.0)
|   |   SOWISK = 6uur: niet behaald (5.0/1.0)
|   |   SOWISK = 4/5uur: meer (1.0)
|   |   SOWISK = 0tot3uur: meer (2.0)
```

- ***<u>Na resultaten 1ste trimester</u>***

```
Correctly Classified Instances         135               70.3125 %
Incorrectly Classified Instances        57               29.6875 %
Kappa statistic                          0.5135
Mean absolute error                      0.2606
Root mean squared error                  0.3875
Relative absolute error                 61.5233 %
Root relative squared error             84.2095 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
 10 23  5 |  a = meer
  5 78  3 |  b = normaal
  9 12 47 |  c = niet behaald
```

```
J48 pruned tree
------------------

Financieel boekhouden <= 7
|   Wiskunde <= 7
|   |   BEURSSTUDENT = BEURSSTUDENT2: niet behaald (49.94/3.0)
|   |   BEURSSTUDENT = BEURSSTUDENT1
|   |   |   SO_STUDIEVERTRAGING = nee: meer (7.0/1.0)
|   |   |   SO_STUDIEVERTRAGING = ja: niet behaald (3.0)
|   Wiskunde > 7: meer (17.88/9.88)
Financieel boekhouden > 7
|   GENERATIESTUDENT = nee
|   |   BEURSSTUDENT = BEURSSTUDENT2
|   |   |   SONAT = 3ofmeeruur: niet behaald (0.0)
|   |   |   SONAT = 2uur: normaal (4.0/1.0)
|   |   |   SONAT = 0/1uur: niet behaald (4.19)
|   |   BEURSSTUDENT = BEURSSTUDENT1: meer (4.0/1.0)
|   GENERATIESTUDENT = ja: normaal (102.0/25.0)
```

- ***<u>Na het eerste jaar</u>***

```
Correctly Classified Instances         159               82.8125 %
Incorrectly Classified Instances        33               17.1875 %
Kappa statistic                          0.724
Mean absolute error                      0.159
Root mean squared error                  0.3272
Relative absolute error                 37.5339 %
Root relative squared error             71.1234 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
 20 13  5 |  a = meer
  4 82  0 |  b = normaal
  7  4 57 |  c = niet behaald
```

```
onvoldoendes <= 2
|   Beginselen van het recht. Privaat en publiek recht. Handelsrecht EIND <= 7: meer (6.0/1.0)
|   Beginselen van het recht. Privaat en publiek recht. Handelsrecht EIND > 7
|   |   Wiskunde  EIND <= 6: meer (4.0)
|   |   Wiskunde  EIND > 6
|   |   |   Talen  EIND <= 10: meer (6.0/1.0)
|   |   |   Talen  EIND > 10: normaal (98.0/13.0)
onvoldoendes > 2
|   onvoldoendes <= 4
|   |   Geo-economie en economische geschiedenis  EIND <= 11: meer (7.0)
|   |   Geo-economie en economische geschiedenis  EIND > 11
|   |   |   Macro-economie  EIND <= 11
|   |   |   |   LEESTEST.TOTAAL <= 84: niet behaald (7.0)
|   |   |   |   LEESTEST.TOTAAL > 84: meer (3.0/1.0)
|   |   |   Macro-economie  EIND > 11: meer (4.0)
|   onvoldoendes > 4
|   |   BEURSSTUDENT = BEURSSTUDENT2: niet behaald (51.0)
|   |   BEURSSTUDENT = BEURSSTUDENT1
|   |   |   Financieel boekhouden <= 5: niet behaald (4.0)
|   |   |   Financieel boekhouden > 5: meer (2.0)
```

- **<u>IBk</u>**

  - ***<u>Voor het begin van de studies</u>***

```
Correctly Classified Instances          85               44.2708 %
Incorrectly Classified Instances       107               55.7292 %
Kappa statistic                          0.1338
Mean absolute error                      0.3728
Root mean squared error                  0.6044
Relative absolute error                 87.9925 %
Root relative squared error            131.3527 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
 13 13 12 |  a = meer
 23 44 19 |  b = normaal
 12 28 28 |  c = niet behaald
```

- ***Na resultaten 1ste trimester***

```
Correctly Classified Instances          87               45.3125 %
Incorrectly Classified Instances       105               54.6875 %
Kappa statistic                          0.1422
Mean absolute error                      0.3659
Root mean squared error                  0.5987
Relative absolute error                 86.3813 %
Root relative squared error            130.1195 %
Total Number of Instances              192


=== Confusion Matrix ===

  a  b  c   <-- classified as
 10 14 14 |  a = meer
 20 47 19 |  b = normaal
 12 26 30 |  c = niet behaald
```

- ***Na het eerste jaar***

```
Correctly Classified Instances         142               73.9583 %
Incorrectly Classified Instances        50               26.0417 %
Kappa statistic                          0.5902
Mean absolute error                      0.1782
Root mean squared error                  0.4132
Relative absolute error                 42.0722 %
Root relative squared error             89.8004 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
 18 10 10 |  a = meer
 12 68  6 |  b = normaal
  7  5 56 |  c = niet behaald
```

- **MultilayerPerceptron**

  - ***Voor het begin van de studies***

```
Correctly Classified Instances         103               53.6458 %
Incorrectly Classified Instances        89               46.3542 %
Kappa statistic                          0.2527
Mean absolute error                      0.3227
Root mean squared error                  0.5121
Relative absolute error                 76.1774 %
Root relative squared error            111.2935 %
Total Number of Instances              192

=== Confusion Matrix ===

  a  b  c   <-- classified as
  7 16 15 |  a = meer
 14 59 13 |  b = normaal
  7 24 37 |  c = niet behaald
```

  - ***Na resultaten 1ste trimester***

```
Correctly Classified Instances         121               63.0208 %
Incorrectly Classified Instances        71               36.9792 %
Kappa statistic                          0.4123
Mean absolute error                      0.2581
Root mean squared error                  0.4611
Relative absolute error                 60.9272 %
Root relative squared error            100.2124 %
Total Number of Instances              192


=== Confusion Matrix ===

  a  b  c   <-- classified as
 11 18  9 |  a = meer
 17 59 10 |  b = normaal
  5 12 51 |  c = niet behaald
```

- ***<u>Na het eerste jaar</u>***

```
Correctly Classified Instances         146               76.0417 %
Incorrectly Classified Instances        46               23.9583 %
Kappa statistic                          0.6237
Mean absolute error                      0.1594
Root mean squared error                  0.3692
Relative absolute error                 37.6209 %
Root relative squared error             80.2477 %
Total Number of Instances              192


=== Confusion Matrix ===

  a  b  c   <-- classified as
 19 12  7 |  a = meer
 12 72  2 |  b = normaal
 10  3 55 |  c = niet behaald
```

		OneR	Naive Bayes	J48	Ibk	MultilayerPerceptron
Bachelor 2	vooraanvang	42	55	55	44	54
	naresult1	72	66	70	45	63
	eindejaar	82	80	83	74	76

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Analyse van de slaagresultaten m.b.v. data-miningtechnieken

Richting: **master in de toegepaste economische wetenschappen: handelsingenieur in de beleidsinformatica**
Jaar: **2013**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Lemmens, Jeroen

Datum: **22/08/2013**