Voorspellen van Studiesucces

Met Machine Learning Technieken

Master project Business Analytics

Johnno Pastor

Begeleider: Evert Haasdijk
Tweede lezer: René Bekker
Amsterdam, juli 2016

Factulteit der Exacte Wetenschappen
De Boelelaan 1081
1081 HV Amsterdam

Stagebedrijf: Future Facts
Danzigerkade 19
1013 AP Amsterdam

Abstract

Het doel van een onderwijsinstelling is om zo goed mogelijk onderwijs te verzorgen. Eén van de maten waaraan onderwijskwaliteit wordt gemeten is het studiesucces van de leerlingen, dat kan worden gekwantificeerd door een combinatie van de verblijfsduur met of en welk diploma wordt behaald.

Dit onderzoek brengt in kaart hoezeer gebruik kan worden gemaakt van machine learning technieken bij het maken van voorspellingen over toekomstig studiesucces van leerlingen in het mbo. Er worden verschillende technieken en algoritmen gebruikt om een voorspelling te maken ofwel een leerling een opleiding zal afronden en hoe lang hierover wordt gedaan. Aan de hand van literatuurstudie wordt onderzocht welke factoren een rol spelen in studiesucces en hoe analytics kan worden toegepast om meerwaarde te creëren. Verder worden de gegevens van een onderwijsinstelling in het mbo tot een enkele dataset gevormd die verder wordt onderverdeeld in train, test en validatie sets, waarop machine learning modellen worden opgesteld waarvan de uitkomsten (met statistische tests) worden vergeleken. Bij het opstellen van de modellen worden verschillende gegevens gebruikt om te zien wat de voorspellende waarde is.

Bij het voorspellen of een leerling een diploma haalt, blijkt dat uit enkel de opleidingsgegevens al patronen kunnen worden geleerd. Een redelijk accurate voorspelling kan gemaakt worden enkel op basis van opleidingsgegevens: met een zekerheid van 68% kan worden voorspelt of een leerling zijn opleiding succesvol zal afsluiten; een voorspelling doen op basis van leerlinggegevens is lastiger, hierbij haalt het beste model een accuracy van 62.2%. Wanneer de gegevens worden gecombineerd verbeteren de resultaten en haalt het beste model een accuracy van 72%. Statische tests wijzen uit dat de beste modellen in alle gevallen significant beter presteren dan een simpele regel. Verder wordt aan de hand van een lift grafiek inzichtelijk gemaakt dat het mogelijk is een groep te onderscheiden waar de kans op uitval hoog is. Dit biedt mogelijkheden voor het gericht sturen van begeleiding. Het best presteren de boosting methodes, die iteratief kleine modellen opstellen om tot een uiteindelijk model te komen.

Aan de hand van regressie wordt een voorspelling gedaan over de duur ten opzichte van de verwachte duur. De resultaten die hierbij worden behaald zijn moeilijker te interpreteren dan die van classificatie. Er wordt een betere voorspelling gedaan dan het naïef schatten van het gemiddelde, maar harde uitspraken over en een interpretatie van de verbetering zijn moeilijk te maken; wel kan worden geconcludeerd dat de verbetering beperkt is en er geen accurate scheiding ontstaat van langstudeerders en leerlingen die hun opleiding afronden binnen de nominale duur.

Om de toegevoegde waarde van aanvullende gegevens over de vooropleiding en gegevens die beschikbaar zijn bij een momentopname na een halfjaar vast te stellen, worden aparte modellen opgesteld. Hierbij blijkt dat gegevens over de vooropleiding een klein maar niet statistisch significante verbetering opleveren, gegevens over het aantal keer dat een leerling van opleiding wisselt en ook de toetsresultaten laten voor sommige modellen weliswaar een verbetering zien, maar deze is wederom niet statistisch significant. Uit analyse van belangrijke variabelen blijkt onder andere dat variabelen afgeleid van open data bronnen geen belangrijke rol spelen.

Machine learning technieken kunnen worden toegepast om een voorspelling te doen over het toekomstig studiesucces van leerlingen. Hierbij speelt de data een grote rol, en ook op dit gebied liggen nog verdere mogelijkheden. Bij dit onderzoek zijn we een aantal problemen met betrekking tot de data tegengekomen. Door de kwaliteit van de data te verbeteren en de beschikbare data verder uit te breiden kan nog winst worden behaald.

Preface

Deze masterthesis is geschreven als afsluiting van de studie Business Analytics aan de Vrije Universiteit Amsterdam. Business Analytics is een multidisciplinaire studie die zich richt op het oplossen van kwantitatieve bedrijfsproblemen. De tweejarige Master wordt afgerond met een stage waarin Business Analytics wordt toegepast om een (bedrijfskundig)probleem uit de praktijk op te lossen. In deze thesis beschrijf ik het onderzoek dat ik heb uitgevoerd tijdens mijn stage bij Future Facts in opdracht van een onderwijsinstelling in het mbo.

Voor een half jaar heb ik stage gelopen bij het recent opgerichte Future Facts, een dochteronderneming van Hot ITem, een bedrijf dat zich bezig houdt met onder andere business intelligence en procesoptimalisatie. Future Facts gaat een stap verder en voert diepgaandere analyses uit en past onder andere machine learning technieken toe om waarde uit data te halen. In de stage zijn modellen ontwikkelt die kunnen worden gebruikt om een voorspelling te doen over toekomstig studiesucces van een leerling.

Ik wil graag een aantal personen bedanken voor de hulp bij mijn stage. Ten eerste wil ik mijn begeleider bij Future Facts, Joost de Jonge, erg bedanken voor de betrokkenheid bij het onderzoek en voor alle feedback. Verder wil ik Evert Haasdijk, mijn begeleider aan de VU, bedanken voor alle hulp en aanbevelingen. Als laatste wil ik ook René Bekker bedanken voor de rol die hij heeft gespeeld als tweede lezer.

Johnno Pastor
Juli 2016

Inhoudsopgave

Lijst met Afkortingen

Onderwijs afkortingen

bbl	Beroepsbegeleide Leerweg
bol	Beroepsopleidende Leerweg
bpv	Beroepspraktijkvorming
BRIN	Basisregistratie Instellingen
BRON	Basis Register Onderwijs
BVE	Beroepsonderwijs en Volwasseneeducatie
Crebo	Centraal Register Beroepsopleidingen
DUO	Dienst Uitvoering Onderwijs
EZ	(Ministerie van) Economische Zaken
FoV	Focus op Vakmanschap
lgf	Leerlinggebonden financiering
mbo	Middelbaar beroepsonderwijs
oac	Agrarisch opleidingscentrum
OCW	(ministerie van) Onderwijs Cultuur en Wetenschap
roc	Regionaal opleidingscentrum
SBU	Studiebelasting
VSO	Voortgezet speciaal onderwijs
WEB	Wet educatie en beroepsonderwijs

Machine Learning afkortingen

ANOVA . . . . . Analysis of Variance
AWNB . . . . . Averaged Weighted Naïve Bayes
BAN . . . . . Bayesian Network Augmented Naïve Bayes
B(-)GLM . . . . Bayesian (-) Generalized Linear Model
BN . . . . . Bayesian Network
CART . . . . . Classification and Regression Trees
DT . . . . . Decision Tree
GBM . . . . . Gradient Boosting Machine
GBN . . . . . General Bayesian Network
GLM . . . . . Generalized Linear Model
MAE . . . . . Mean Absolute Error
MDL . . . . . Minimum Description Length
mice . . . . . Multivariate imputation by chained equations
MSE . . . . . Mean Squared Error
kNN . . . . . k Nearest Neighbour
NB . . . . . Naïve Bayes
RBF . . . . . Radial Basis Function
RF . . . . . Random Forest
RMSE . . . . . Root Mean Squared Error
ROC . . . . . Receiver Operating Characteristic
SVM . . . . . Support Vector Machine
TAN . . . . . Tree Augmented Naïve Bayes

Lijst met termen

Basisberoep-sopleiding	Opleiding die leerlingen voorbereidt uitvoerende werkzaamheden te doen (bv. kapper of autotechnicus). Duurt 1-2 jaar.
Beroepsbegeleide leerweg	Leerweg in het mbo waarbij het grootste deel van theorische aard is en op een roc wordt gevolgd.
Beroepsopgeleide leerweg	Leerweg in het mbo waar de nadruk ligt op de praktijk; bestaat voornamelijk uit beroepspraktijkvorming en slecht een klein deel van het onderwijs is theoretisch en vind plaats op een roc.
Beroepspraktijk-vorming	Het deel van een mbo-opleiding dat bestaat uit werken en leren in de praktijk bij een erkent leerbedrijf.
Basisregistratie Instellingen	Een register met alle scholen en aanverwante instelling, wordt uitgegeven door het Ministerie van OCW en bijgehouden door DUO.
Basis Register Onderwijs	Een register met leerlinggegevens aangeleverd door onderwijsinstellingen in Nederland.
Crebocodes	Unieke code dat verbonden is aan een kwalificatiedossier van het centraal register beroepsopleidingen, wordt door onderwijsinstellingen gebruikt voor de administratie van onderwijs, examens en leerlingen.
Deelnemer	Leerling of student.
Dienst Uitvoering Onderwijs	Dienst die onderwijsdeelnemers en instelling informeert en financiert; heeft verschillende taken zoals dragen zorg voor tentamen en het erkennen van diploma's die onderwijs mogelijk maken.
Entreeopleiding	Opleiding voor jongeren zonder een diploma van een vooropleiding, bereidt jongeren voor op de arbeidsmarkt. Duurt 1 jaar.
Extranei	Meervoudsvorm van extraneus, een leerling die een examen afleggen zonder de betrokken school te bezoeken voor onderwijs.
Focus op Vakmanschap	Actieplan dat liep van 2011-2015 waarin werd gestuurd op het verhogen van de kwaliteit van het mbo, vereenvoudiging van het mbo als stelsel en het op orde brengen van de bedrijfsvoering.
MBO Raad	Brancheorganisatie van de onderwijsinstellingen in het mbo.
Middenka-deropleiding	Opleiding die leerlingen leert werkzaamheden volledig zelfstandig uit te voeren (bv. filiaalbeheerder of activiteitenbegeleider). Duurt 3 jaar.
Opleidingsdomein	Mbo opleidingen zijn geordend in zestien domeinen van opleidingen die zijn gericht op een bedrijfstak of groep van bedrijfstakken. Worden vastgesteld door het ministerie van OCW samen met de kwalificatiestructuur.

Prijsfactor	Wegingsfactor die aangeeft welk deel uit het macrobudget een opleiding krijgt toegerekend.
Regionaal opleidingscentrum	Opleidingscentrum dat opleidingen in het middelbaar beroepsonderwijs en volwasseneneducatie aanbiedt.

1. Introductie

1.1 Probleembeschrijving

Het doel van een onderwijsinstelling is om zo goed mogelijk onderwijs te verzorgen. Een van de succesmaten waaraan onderwijskwaliteit wordt gemeten is het studiesucces van de leerlingen. Studiesucces kan worden gekwantificeerd door een combinatie van de verblijfsduur met of een (en welk) diploma wordt behaald. Om studiesucces te bevorderen wordt al gestuurd op cijfers als leerlingenaantallen en het diplomarendement. Het onderwijs verbetert zich door in te zetten op professionalisering van docenten, verzuim van leerlingen op te volgen en door onderwijsprocessen te optimaliseren.

Studiesucces wordt daarnaast ook financieel beloond door de manier waarop de bekostiging is ingevuld. Voor het middelbaar beroepsonderwijs is de grootte van de bekostiging afhankelijk van het aantal diploma's en de tijd die nodig is om een opleiding af te ronden. Verder wordt studiesucces extra gestimuleerd door het ministerie met de regeling kwaliteitsafspraken mbo van 8 december 2014. Hierin is een resultaatafhankelijk budget opgenomen wat een extra stimulans geeft voor het verbeteren van onderwijsprestaties.

Opleidingsinstituten beschikken over informatiesystemen die inzicht geven in studievoortgang. Voor onderwijsinstellingen is het interessant wanneer zij naast de al beschikbare gegevens inzicht kunnen krijgen in de verwachte prestaties van een leerling. Een voorspelmodel opgesteld met machine learning technieken kan worden ingezet om een voorspellingen te doen en dit mogelijk te maken. De voorspellingen en inzichten die hieruit komen kunnen worden ingezet om studiesucces te bevorderen. De meerwaarde is hierbij afhankelijk van de gebruiker, de plaats in het proces en de toepassing.

1.2 Onderzoeksvragen

De hoofdonderzoeksvraag is als volgt geformuleerd:

"Wat is de toegevoegde waarde van interne, externe data en voorspellende modellen op beleid voor studiesucces?"

Om antwoord te geven op deze hoofdvraag, wordt er een antwoord gezocht op de volgende deelvragen:

1. Welke factoren worden in de literatuur gedefinieerd als het hebben van invloed op studiesucces en hoe corresponderen deze met beschikbare variabelen?

2. Hoe kunnen de uitkomsten van een voorspelmodel worden ingezet om toegevoegde waarde te creëren?

3. Welke externe data kan worden ontsloten en hoe kan deze worden gekoppeld aan de beschikbare interne data?

4. Hoe goed zijn de voorspellingen gemaakt door voorspelmodellen opgesteld met verschillende technieken en gebruik makend van verschillende data bronnen?

De eerste deelvraag zal aan bod komen bij de literatuurstudie naar belangrijke factoren bij studiesucces. Het inzicht in welke door onderzoek uitgewezen factoren een rol spelen bij studiesucces

zal worden meegenomen bij het opstellen van de modellen: welke beschikbare variabelen hebben volgens de literatuur voorspellende waarde en zullen daarom in de voorspelmodellen van belang zijn?

De tweede deelvraag gaat in op beleid, en zal terugkomen in de achtergrond en in het tweede deel van de literatuurstudie. De deelvraag bevat twee aspecten. Ten eerste de hoe-vraag: op welke manier kan een voorspelmodel een praktische bijdrage leveren. Het tweede aspect betreft de toegevoegde waarde: hoe groot (ten opzichte van een baseline) is de toegevoegde waarde. Hiervoor zal eerst worden gekeken naar hoe nu data en analytics worden ingezet en vervolgens naar de plaats waar en manier waarop een voorspelmodel en zijn uitkomsten kunnen worden ingezet. Verder zal in het literatuuronderzoek aan bod komen op welke manieren data mining reeds wordt toegepast.

De derde deelvraag gaat in op het gebruik van externe (open) data. Het gebruik van open data is een interessante invalshoek, en is verder interessant voor het stageverleenend bedrijf met het oog op herbruikbaarheid. Er zal worden gekeken naar welke open data beschikbaar is en hoe deze kan worden ontkoppelt en klaargemaakt zodat deze op een makkelijke manier opnieuw te gebruiken is.

Aan de hand van de laatste deelvraag worden de prestaties van verschillende modellen en technieken beoordeeld. Deze deelvraag gaat ook hand in hand met de eerste en derde deelvraag: er zal worden teruggekomen op het belang van de variabelen (uit open data) in de opgestelde modellen en hoe technieken presteren ten opzichte van elkaar.

1.3 Opbouw

Hoofdstuk 2 schets de achtergrond van het probleem door een overzicht te geven van het onderwijslandschap in Nederland, hierbij zal in het speciaal worden ingegaan op het mbo. Er wordt afgesloten met de plek van analytics en een voorspelmodel.
Hoofdstuk 3 behandelt literatuur en is tweedelig. Ten eerste zal onderzoek gedaan worden naar welke factoren invloed hebben op studiesucces. Ten tweede zullen huidige toepassingen van data mining en het maken van voorspellingen in het onderwijs worden besproken.
Hoofdstuk 4 gaat in op de data die gebruikt is voor dit onderzoek. Het hoofdstuk begint met een beschrijving van interne data verkregen uit de informatiesystemen en externe data uit open data bronnen. Vervolgens wordt besproken hoe tot de samengestelde dataset is gekomen dat wordt gebruikt als invoer voor het voorspelmodel. Er wordt afgesloten met een analyse van de samengestelde dataset.
Hoofdstuk 5 beschrijft de toegepaste technieken. In dit hoofdstuk worden machine learning algoritmen, preprocessing methoden en validatie technieken besproken.
Hoofdstuk 6 beschrijft de experimentele methodiek – hoe is het onderzoek opgezet en welke implementaties van de (Machine Learning) algoritmen zijn er gebruikt.
Hoofdstuk 7 presenteert de resultaten van verschillende modellen. In dit hoofdstuk zal een vergelijking gemaakt worden van de toegevoegde waarde van verschillende databronnen en variabelen, technieken en methoden.
Hoofdstuk 8, de conclusie, zal antwoord geven op de in dit hoofdstuk gestelde onderzoeksvragen.
Hoofdstuk 9 betreft de discussie en bespreekt aanbevelingen en verder onderzoek.

2. Achtergrond

Dit hoofdstuk zal een schets geven van de achtergrond. Ten eerste zal het onderwijsstelsel in Nederland in zijn algemeen worden besproken. Hierbij worden een aantal kerncijfers gepresenteerd en wordt besproken hoe de bekostiging ruwweg in elkaar steekt. Vervolgens wordt het middelbaar onderwijs (mbo) verder uitgewerkt. Tot slot wordt behandeld hoe analytics en een voorspelmodel (kunnen) worden ingezet.

2.1 Het Onderwijslandschap

Het Nederlandse onderwijsstelsel in Nederland is verdeeld naar leeftijdsgroep en vervolgens naar niveau. Figuur 2.1 geeft het onderwijsstelsel schematisch weer. In dit verslag zal de nadruk worden gelegd op de instellingen die onderwijs verzorgen voor de oudere leeftijdsgroep en daarmee de bovenste helft van Figuur 2.1.

Het secundair beroepsonderwijs of middelbaar beroepsonderwijs verzorgt beroepsgerichte opleidingen en heeft als doel leerlingen op te leiden voor een beroep. Na het behalen van een diploma gaat het overgrote deel van de afgestudeerden de arbeidsmarkt op, daarnaast is het mogelijk om opleidingen in het mbo te 'stapelen' of door te stromen naar het hoger beroepsonderwijs. Het hoger beroepsonderwijs (hbo) vormt samen met het wetenschappelijk onderwijs (wo) het hoger onderwijs. Het hoger beroepsonderwijs wordt verzorgt door hogescholen en is gericht op de beroepspraktijk. Het niveau ligt hoger dan op het mbo en het bereidt studenten dan ook voor op functies die meer zelfstandigheid en deskundigheid vereisen. Om te helpen bij de beeldvorming worden een aantal kerncijfers gepresenteerd in Tabel 2.1.

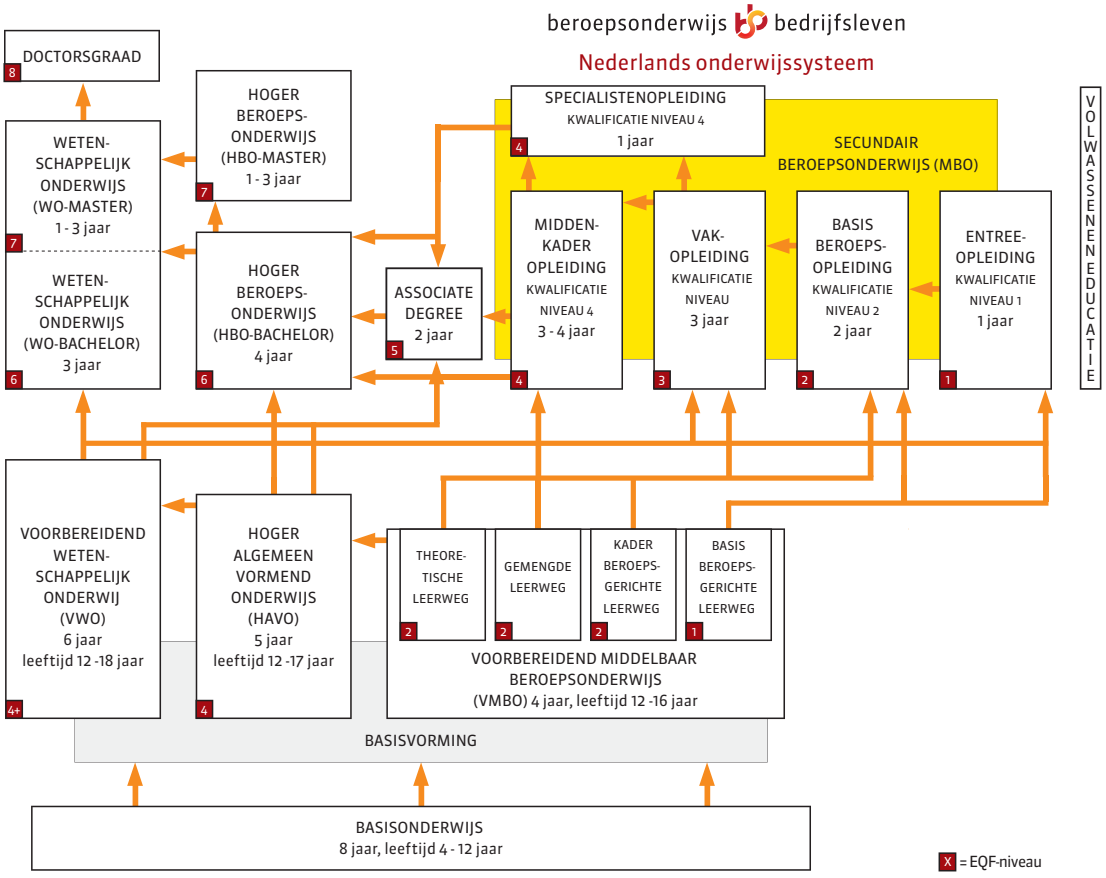
	mbo	hbo	wo
Instellingen (aantal)	66	37	19
Leerlingen (aantal)	494,000	446,457	256,506
Overheidsbijdrage (mld, euros)	4,926	3,769	5,325
Waarvan lumpsum (mld, euros)	3,596	2,917	4,923

Tabel 2.1: Cijfers onderwijsvormen 2014/'15, bron cijfers: CBS, 2015 [1]

2.1.1 Bekostiging Onderwijs door de Rijksoverheid

Het onderwijs in Nederland wordt deels bekostigd door de Rijksoverheid. Alle instellingen in het hoger onderwijs krijgen één bedrag van de overheid: de lumpsum. De lumpsum is bedoeld voor de financiering van personeel, materieel en huisvesting, en elke instelling kan deze naar eigen inzicht besteden. De grootte van de lumpsum is gebaseerd op het aantal inschrijvingen en het aantal verstrekte diplomas. Verder bestaat er in het hoger onderwijs nog een extra budget gebaseerd van het nakomen van prestatieafspraken. Ook voor het mbo bestaat de bijdrage deels uit een lumpsum. Mbo-instellingen krijgen daarnaast extra budget als ze kwaliteitsafspraken nakomen.

[1] http://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=03753



Figuur 2.1: Het Nederlandse onderwijssysteem, bron: s-bb [2]

2.1.2 Het middelbaar beroepsonderwijs

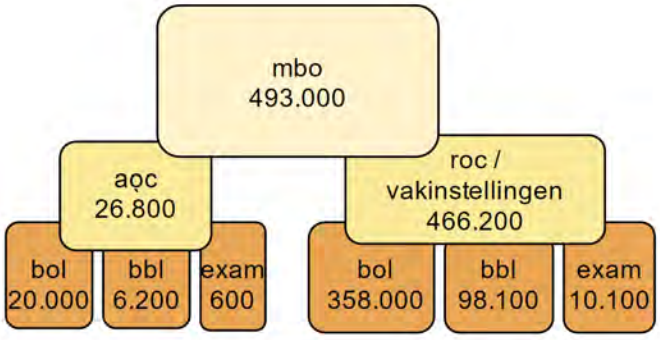
Het middelbaar beroepsonderwijs (mbo) heeft als doel leerlingen op te leiden voor een beroep. In het mbo zijn twee vormen in de indeling van onderwijs te onderscheiden: de beroepsopleidende leerweg (bol) en de beroepsbegeleidende leerweg (bbl). In de bol is de leerling het grootste deel van de opleiding op school, in de bbl ligt de nadruk op de praktijk. Naast dat het mbo een onderverdeling kent naar leerweg, is er ook een onderverdeling naar niveau. In totaal zijn er vier niveaus die ieder een eigen duur en benodigde vooropleidingen kennen (zie ook Figuur 2.1).

Er zijn verschillende instellingen die het middelbaar beroepsonderwijs verzorgen: Regionale Opleidingencentra (roc's), Agrarische Opleidingscentra (aoc's) en vakinstellingen. De meeste mbo-opleidingen worden gegeven op roc's, zoals opleidingen gericht op bouw, techniek, en zorg. Deze opleidingen en daarmee ook hun bekostiging vallen onder het ministerie van Onderwijs, Cultuur en Wetenschap (OCW). Groene opleidingen worden gegeven op aoc's en vallen onder het ministerie van Economische Zaken (EZ). Vakinstellingen verzorgen mbo-opleidingen voor een bepaalde branche.

Bekostiging middelbaar beroepsonderwijs

Er is een extra 400 miljoen euro beschikbaar gesteld om onderwijsinstellingen aan te moedigen zich te verbeteren op zes landelijk geformuleerde thema's, deze financiering valt uiteen in een investeringsbudget en resultaatafhankelijk budget. Het resultaatafhankelijk budget komt uit op

[2] https://www.s-bb.nl/studenten/diplomawaardering-en-onderwijsvergelijking/het-nederlandse-onderwijssysteem



Figuur 2.2: Leerlingenaantallen mbo, bron cijfers: DUO, 2015 [3]

[3] https://duo.nl/open_onderwijsdata/databestanden/mbo/Onderwijsdeelnemers/Deelnemers_mbo4.jsp

5.5% (211 miljoen euro) van het totale mbo-budget. In 2016 bedraagt het deel beschikbaar gesteld voor het verbeteren van studiewaarde 99 miljoen euro, en in 2017 en 2018 116 miljoen euro. Mbo-instellingen ontvangen deze extra middelen als zij of een verbetering realiseren of goede resultaten behouden.

De bijdrage vanuit het rijk bestaat naast het al benoemde extra budget uit een lumpsum. Mbo-instellingen worden bekostigd vanuit twee macrobudgetten: één voor roc's en vakinstellingen, en één voor aoc's. Per macrobudget gelden weer afzonderlijke budgetten voor entreeopleidingen en overige niveaus.

De bekostiging is geregeld in de Wet Educatie en Beroepsonderwijs (WEB) en is in 2015 op de schop gegaan met de regeling kwaliteitsafspraken. Voor mbo-opleidingen voor niveaus twee tot en met vier bestaat de bekostiging uit een input- (aantal ingeschreven leerlingen) en een outputbekostiging (diplomas). Voor de entreeopleidingen is er een apart budget met alleen een inputbekostiging. Voor de verdeling van het budget wordt voor elke instelling een som gemaakt van deelnemerswaarden en diplomawaarden en deze som bepaalt het relatieve aandeel in het totale macrobudget. Een uitleg over hoe de rijksbijdrage wordt berekend is te vinden in Appendix B.

Inputbekostiging: De deelnemerswaarde is afhankelijk van de volgende factoren: het aantal ingeschreven deelnemers, het niveau, het type leerweg, het aantal verblijfsjaren van de deelnemer (cascade) en een prijsfactor (reflecteren de kosten om de opleiding aan te bieden).

Outputbekostiging: De diplomawaarde van elke instelling voor niveaus twee tot en met vier wordt berekend door de waarden van elk afgegeven diploma op te tellen. De waarde van een diploma is afhankelijk van het niveau van de opleiding en of de deelnemer al eerder een mbo-diploma heeft behaald.

2.1.3 Het inzetten van analytics en een voorspelmodel

Onderwijsinstellingen worden aangemoedigd om studiesucces te bevorderen en zijn hier al mee bezig door professionalisering van personeel, verzuim op te volgen en door onderwijsprocessen te verbeteren. Uit vele voorbeelden blijkt dat onderwijsinstellingen bezig zijn met het aanpakken van vroegtijdig school verlaten, zo blijkt ook uit het jaarverslag ROC midden Nederland 2014: "Sterke sturing op VSV aan de hand van verschillende indicatoren, ... Het lukt steeds beter om vroegtijdig risicos te signaleren en de studenten succesvol te begeleiden ... Verzuim is een voorbode van uitval, door dit goed in beeld te hebben, tijdig te melden en zorgvuldige actie voorkomen we uitval. "

Hieruit blijkt dat de manier van sturen verschuift naar meer gestandaardiseerd en op cijfers gegrond. In Amerika gebeurd dit al op grotere schaal, het bekendste bedrijf wat zich hier mee bezighoudt is Ruffalo Noel Levitz. Zij bieden diensten die zich richten op het verhogen van het aantal inschrijvingen ('Strategic Enrollment Management') en het verbeteren van de retentiegraad ('Graduate Enrollment Management') en zetten hierbij data, analytics en assementments in.

Huidige toepassingen van analytics zijn vaak gericht op 'traditionele' business intelligence diensten als dashboards. Het onderwijsverslag 2013-2014 [Inspectie van het onderwijs, 2013] zegt hierover: "Interpretatie studievolgsystemen lastig - De meeste opleidingen beschikken over studievolgsystemen, maar deze zijn voor docenten niet altijd goed bruikbaar." De uitkomst van een voorspelmodel maakt de interpretatie mogelijk eenvoudiger.

In Nederland zijn er geen partijen die dit soort diensten op grote schaal aanbieden. Wel biedt het DUO het mbo diensten aan die grenzen aan analytics en waarmee "het bestuur gericht kan werken aan de loyaliteit en tevredenheid van (oud-)studenten, medewerkers en andere stakeholders"[4]. Hiertoe worden vier mogelijkheden geboden, waarbij met kwalitatief en kwantitatief onderzoek inzicht wordt gegeven in de verdere loopbaan van schoolverlaters, de prestaties van docenten, het imago van de instelling en waarom leerlingen juist wel of niet voor de instellingen hebben gekozen.

Hiermee richt het DUO zich vooral op het schoolniveau door stuur- en verantwoordingsprocessen te ondersteunen. Een andere mogelijkheid van een voorspelmodel is op leerlingniveau: een toepassing die hierbij voor de hand ligt en veel in de literatuur naar voren komt, is het onderscheiden van veelbelovende en minder belovende leerlingen. Het biedt de personen die begeleiding sturen een extra hulpmiddel om vroegtijdig problemen te signaleren en kunnen daarmee passende begeleiding bieden aan de leerlingen die dit nodig hebben. De plek waar zo'n model kan worden ingezet staat niet vast, wel is natuurlijk wenselijk zo vroeg mogelijk problemen te signaleren. Daarnaast is het mogelijk verder in het leerproces een nieuwe voorspelling te doen en hierbij bijvoorbeeld gebruik te maken van gegevens over aanwezigheid en toetsresultaten.

Bij deze toepassing gaat de interpretatie over de data die voorkomt uit het leerproces van een individuele leerling. De uitkomsten van een dergelijk model kunnen ook worden gebruikt om overzichten op te leveren die gebruikt kunnen worden op hoger niveau. Hiermee hoeft de toepassing niet gericht te zijn op de leerling of het leerproces: de uitkomsten kunnen worden geanalyseerd om op hogere schaal aanpassingen te maken in het schoolproces. Zo zou een voorspellend model kunnen liggen aan de basis van een (budget-)simulatiemodel, waarbij leerlingenaantallen worden gesimuleerd aan de hand van een voorspelde uitstroom. Of kan een opgesteld model worden geanalyseerd om antwoord te geven op vragen als: bevatten vooropleiding of toetsresultaten voorspellende waarde?

In Nederland loopt het hoger onderwijs iets voor in de toepassing van analytics. SURFnet, een organisatie die ICT-voorzieningen biedt ter verbetering van samenwerking in het (hoger) onderwijs en onderzoek, heeft de taak op zich genomen om met onderwijsinstellingen samen te werken in kennisdeling en het vastleggen van standaarden voor Learning Analytics. Hiervoor is de special interest group (SIG) Learning Analytics opgericht, een kennisgemeenschap om kennisuitwisseling te bevorderen. Eén van de initiatieven is om onderzoek te doen naar: "ICT-infrastructuur die de analyse uit verschillende databronnen mogelijk maakt. In 2016 doen we samen met een aantal instellingen ervaring op met het implementeren van learning-analyticsarchitecturen, -standaarden en -software om te onderzoeken hoe zo'n infrastructuur eruit moet zien."

[4] http://www.duo-onderwijsonderzoek.nl/middelbaar-beroepsonderwijs/

3. Literatuur

In dit hoofdstuk wordt de voor dit onderzoek bestudeerde literatuur besproken. In de eerste sectie zal onderzoek naar studiesucces en de factoren (variabelen) die hierin een rol spelen worden besproken, in de tweede sectie zullen toepassingen van data mining in het onderwijs worden besproken.

3.1 Studiesucces

Onderzoek Noorderpoort

Het onderzoek van Noorderpoort et al. [2013], "Big data van Hype naar actie", is een onderzoek dat in gaat op het gebruik van (big) data om vroegtijdig schoolverlaat aan te pakken. Hierbij is het doel om inzicht te creëren; er wordt geen machine learning toegepast om een voorspelling te doen over toekomstig studiesucces op individueel niveau. Het biedt een goed beginpunt voor het onderzoek naar factoren die van invloed zijn op studiesucces en biedt handvatten voor hoe een big data onderzoek in het onderwijs vorm te geven. Het literatuuronderzoek wordt benoemd als belangrijk startpunt. Hier wordt invulling aan gegeven door een lijst met risicofactoren uit de literatuur te geven, zie hiervoor Figuur 3.1.

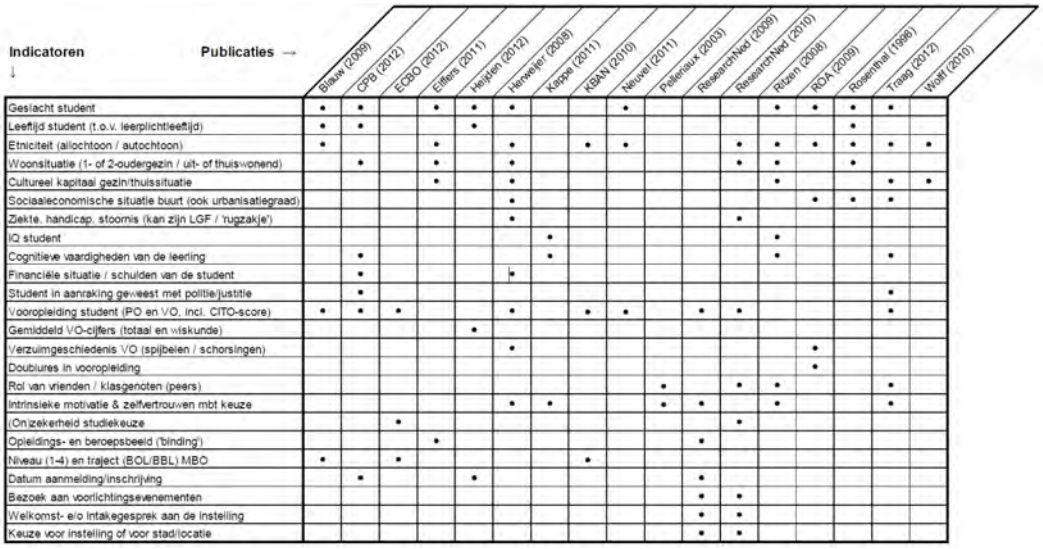
Het kwantitatieve onderzoek van Noorderpoort deelt studiesucces op in switch, duur en het behalen van een diploma. Het onderzoek heeft uitgewezen dat leeftijd gecorreleerd is met studiesucces: oudere leerlingen hebben een grotere kans op het behalen van een diploma en hebben een kleinere kans om vertraging op te lopen, wel is de kans op een switch groter. Verder lopen allochtonen relatief minder vertraging op. Het aantal adressen waar een leerling tijdens zijn studie woont heeft een negatief effect op het behalen van een diploma en resulteert over het algemeen in meer vertraging. Ook het opleidingsniveau en de leerweg variant zijn van invloed. Hierbij geldt dat hoe lager het niveau, hoe kleiner de kans op studiesucces. In de bbl en bol deeltijd wordt meer studiesucces behaald dan in de bol voltijd. Een begeleidend gesprek is negatief gecorreleerd met het behalen van diploma, dit beschrijft echter (hopelijk en hoogstwaarschijnlijk) geen oorzakelijk verband. Verder zijn ook nog het hebben van schulden en het later aanmelden voor een opleiding negatief gecorreleerd met het behalen van een diploma.

Behalve het aantal adressen en schulden zijn voor dit onderzoek soortgelijke gegevens beschikbaar.

Figuur 3.1 geeft een overzicht en laat enkel zien welke risicofactoren worden besproken in de literatuur; niet of zij daadwerkelijk van invloed zijn. Geslacht, leeftijd, etniciteit en vooropleiding zijn een aantal factoren die veel worden beschouwt. Minder prevalent zijn behaalde resultaten, informatie over doubleren, en verzuimgeschiedenis. Dit overzicht betreft zowel kwantitatieve als kwalitatieve onderzoeken. Omdat verder ook ieder onderzoek gebruik maakt van andere groepen leerlingen is het lastig een eenduidig beeld te schetsen van welke factoren in welke gevallen mogelijk van invloed zijn, onder andere omdat voor sommige factoren, zoals achtergrond, geen eenduidigheid bestaat.

Persoonlijkheid en studieduur

Kappe [2011] heeft onderzoek gedaan naar de invloed van persoonlijkheid op studiesucces van hogeschool studenten. Hieruit is gekomen dat consciëntieusheid van groot belang is, en zwaarder weegt dan intelligentie bij zowel de tijd-tot-afstuderen als het behalen van een diploma. Dit



Indicatoren ↓ / Publicaties →	Blauw (2009)	CPB (2012)	ECBO (2012)	Eiffers (2011)	Heijden (2012)	Herweijer (2008)	Kappe (2011)	KBAN (2010)	Nieuwet (2011)	Pellenaux (2003)	ResearchNed (2008)	ResearchNed (2010)	Ritzen (2008)	ROA (2009)	Rosenthal (1998)	Traag (2012)	Wolff (2010)
Geslacht student	•	•		•	•	•			•			•	•	•	•	•	
Leeftijd student (t.o.v. leerplichtleeftijd)	•	•			•										•		
Etniciteit (allochtoon / autochtoon)	•			•		•		•	•			•	•	•	•	•	•
Woonsituatie (1- of 2-oudergezin / uit- of thuiswonend)		•		•		•						•	•		•		
Cultureel kapitaal gezin/thuissituatie				•		•							•			•	•
Sociaaleconomische situatie buurt (ook urbanisatiegraad)						•								•	•	•	
Ziekte, handicap, stoornis (kan zijn LGF / 'rugzakje')						•						•					
IQ student							•						•				
Cognitieve vaardigheden van de leerling		•					•						•			•	
Financiële situatie / schulden van de student		•				•											
Student in aanraking geweest met politie/justitie		•														•	
Vooropleiding student (PO en VO, incl. CITO-score)	•	•	•			•		•	•		•	•				•	
Gemiddeld VO-cijfers (totaal en wiskunde)					•												
Verzuimgeschiedenis VO (spijbelen / schorsingen)						•								•			
Doublures in vooropleiding														•			
Rol van vrienden / klasgenoten (peers)										•		•	•			•	
Intrinsieke motivatie & zelfvertrouwen mbt keuze						•	•			•	•		•			•	
(On)zekerheid studiekeuze			•									•					
Opleidings- en beroepsbeeld ('binding')				•							•						
Niveau (1-4) en traject (BOL/BBL) MBO	•		•				•										
Datum aanmelding/inschrijving		•			•						•						
Bezoek aan voorlichtingsevenementen											•	•					
Welkomst- e/o intakegesprek aan de instelling											•	•					
Keuze voor instelling of voor stad/locatie											•	•					

Figuur 3.1: Indicatoren die worden besproken in het literatuuronderzoek van Noorderpoort, bron: [Noorderpoort et al., 2013]

persoonlijkheidskenmerk kenmerkt de meer georganiseerde, gedisciplineerde en hardwerkende student. Voor dit onderzoek zijn geen gegevens over persoonlijkheid beschikbaar, mogelijk kan het afnemen van een persoonlijkheidstest in de toekomst interessant zijn om mee te nemen bij het maken van een voorspelling.

Naast het onderzoek van Noorderpoort en Kappe is er vrij beperkt onderzoek gedaan naar factoren die van invloed zijn op studieduur. Ook Herzog [2006], een onderzoek dat in de volgende sectie behandeld zal worden, benoemt dit: Er zijn wel enkele regressie modellen en pad analyses die bijdragen aan het begrip over retentie, maar de hoeveelheid onderzoek gewijd aan tijd-tot-afstuderen is beperkt. Er is meer onderzoek naar vroegtijdig schoolverlaten, de risicofactoren die hier van belang zijn en welke preventie technieken helpen.

Etniciteit en delinquent gedrag

Traag et al. [2010] hebben risicofactoren onderzocht aan de hand van data uit het Voortgezet Onderwijs Cohort 1999 in samengang data uit Herkenningsdienstsystemen over delinquent gedrag. Hieruit is gebleken dat achtergrondkenmerken als geslacht en herkomst sterk samenhangen met zowel voortijdig schoolverlaten als jeugdcriminaliteit. Voor voortijdig schoolverlaten is er, na controle voor andere variabelen, geen significant verschil tussen allochtone en autochtone leerlingen. Wel van belang zijn de religie en burgerlijke staat van de ouders. Voor dit onderzoek zijn geen gegevens beschikbaar over de ouders van leerlingen.

Over de invloed van etniciteit op studiesucces bestaat in de literatuur geen eenduidigheid. Het onderzoek van Noorderpoort et al. [2013] heeft uitgewezen dat het allochtonen over het algemeen (minder) vertraging oplopen. Wolff et al. [2010] concluderen op basis van gegevens uit het hoger onderwijs daarentegen dat etnische afkomst een negatief effect heeft op studievoortgang, ook wanneer hierbij gecorrigeerd wordt voor het opleidingsniveau van de ouders.

Pijpers [2010] heeft een multivariate analyse uitgevoerd op een dataset met alle leerlingen tussen de 12 en 23 jaar in het vo, mbo, en vavo in de Gemeentelijke Basis Administratie in het cohort 2007/2008. Of leerlingen vroegtijdig uitvallen hangt af van een combinatie van factoren. Hierbij is naast leeftijd het leeftijdsverschil tussen leerlingen en hun klasgenoten een belangrijke factor, eveneens als het aantal keren dat een leerling verdacht is geweest van een misdrijf. Interacties tussen deze en overige factoren zijn in hoge mate bepalend voor het risico op voortijdige schooluitval. De relatief hogere uitval onder allochtone jongeren kan hier voor een groot deel verklaard worden door

een gevolg van andere factoren dan herkomst.

Schoolverzuim

Baat [2010] heeft gekeken naar het aanpakken van schoolverzuim omdat eerder onderzoek heeft uitgewezen dat verzuim een voorloper is van voortijdig schoolverlaten. Voor dit onderzoek is helaas slechts beperkte informatie beschikbaar over verzuim: voor minder dan de helft van de deelnemers is er enkel een gemiddelde presentie aanwezig (welke daarnaast vaak niet gaat over het begin van de studie). Baat [2010] concludeert dat de effectiviteit van coaching het grootst is voor de groep waar vroegtijdig schoolverlaten in hogere mate voorkomt, zoals leerlingen met een sociaaleconomische achterstand. Dit onderlijnt de mogelijke meerwaarde: als het mogelijk is de groep waar de kans op uitval hoog is te onderscheiden, dan kan deze effectief begeleiding worden geboden.

Onderwijsgeschiedenis en opleiding

Wijk et al. [2012] hebben grootschalig onderzoek gedaan naar de invloeden van onderwijsgeschiedenis (loopbaan) en de kenmerken van een opleiding zoals niveau, intensiteit en leerweg op voortijdig schoolverlaten. Hiervoor is gebruik gemaakt van informatie over de zeven jaar voorafgaand aan uitval aan de hand van het onderwijsnummer voor alle leerlingen die op 1 oktober 2009 een geldige inschrijving hadden.

Het onderzoek heeft uitgewezen dat de loopbaan een beduidende invloed heeft op uitval. Tijdens het eerste leerjaar van het mbo geldt dat nog minder dan 50% een reguliere leerweg heeft gevolgd: leerlingen blijven zitten, stapelen of wisselen van opleiding en niveau. De meeste afwijkende leerloopbanen vertonen relatief meer voortijdige uitval. Dit gaat echter niet altijd op, enkele afwijkende loopbanen vertonen juist weinig voortijdig schoolverlaters. Ook in de reguliere loopbanen zijn patronen te ontdekken, zo vallen bijvoorbeeld vmbo-bl-leerlingen vaker uit als zij doorstromen naar bbl niveau 2 (9,8%) dan als zij doorstromen naar bbl niveau 3 (6,7%).

Verder is gebleken dat voor alle niveaus geldt dat leerlingen in de bbl vaker uitvallen dan leerlingen in de bol. Of dit komt door verschillen tussen de leerwegen op zich of door onderliggende kenmerken, staat niet vast. Verder blijkt dat leerlingen met een reguliere loopbaan relatief vaker uitvallen in de bbl dan in de bol. En leerlingen die tijdens hun leerloopbaan van leerweg wisselen, vaker uitvallen dan de gemiddelde leerling. Ook hierbij is het weer lastig te bepalen wat de onderliggende oorzaak is, en of de wissel mogelijk enkel een signaal betreft.

Ook de invloed van de vooropleiding is onderzocht: hieruit is gebleken dat leerlingen met een lage vooropleiding in beide leerwegen en op alle niveaus vaker uitvallen in het eerste verblijfsjaar. Het verschil in uitval is echter meestal klein. Er wordt dan ook gewaarschuwd dat een strengere selectie tot gevolg kan hebben dat meer leerlingen onterecht worden afgewezen dan terecht op een passender niveau worden geplaatst.

Onderverdeling risicofactoren

Rooij et al. [2010] van ResearchNed hebben in opdracht van het Ministerie van OCW literatuuronderzoek uitgevoerd naar risicofactoren die ten grondslag liggen aan voortijdig schoolverlaten. Hierin wordt een interessante opdeling gemaakt naar 'klassieke' risicofactoren t.o.v. factoren die 'geruisloze' vsv'ers kenmerkt. Verder worden deze factoren verder onderverdeeld in persoonlijke-, gezins-, omgevings- en schoolfactoren.

Dit onderzoek schetst voornamelijk een overzicht en bespreekt voor een groot deel indicatoren die al eerder benoemd zijn en komt hierbij niet op nieuwe inzichten. Verder komen factoren ter sprake die moeilijk mee te nemen zijn als invoer voor een voorspelmodel en daarnaast moeilijk te beïnvloeden zijn: schoolfactoren als relaties, steun, of pestproblematiek en moeite met het vinden van een stage.

Conclusie

De onderzoeken die hier besproken zijn bieden inzicht in welke variabelen van belang kunnen zijn. In vergelijking met sommige andere onderzoeken zijn de gegevens die voor dit onderzoek beschikbaar zijn vrij beperkt. Twee factoren die veel naar voren komen maar helaas missen voor dit onderzoek, zijn delinquent gedrag en gegevens over de ouders. Studieresultaten (zowel tussenresultaten als voorgaande schoolresultaten) komen vrijwel niet voor in de literatuur, terwijl deze een interessante indicator kunnen zijn voor toekomstig studiesucces. Voorgaande resultaten kunnen ook dienen als proxy om variabelen te weerspiegelen die ten grondslag liggen aan studiesucces, zoals de karaktereigenschap consciëntieusheid. Over de invloed van afkomst bestaat geen eenduidigheid. De invloed van deze factor wordt vaak verklaart door een combinatie van andere onderliggende kenmerken zoals opleidingsniveau van de ouders. Het grootschalig onderzoek naar de invloed van loopbaan en de kenmerken van opleidingen laat interessante resultaten zien. Zo zijn er verschillen tussen de leerwegen en niveau op zich, maar blijkt ook dat de loopbaan vaak een grote rol speelt. Vandaar dat het spijtig is dat de vooropleidingsgegevens slechts beperkt (en in beperkte vorm) beschikbaar zijn.

3.2 Learning Analytics and Knowledge & Educational Data Mining

Stromingen

Er zijn twee stromingen die zich bezighouden met het gebruik van data en analytics in het onderwijs: 'Learning Analytics and Knowledge (LAK)' en 'Educational Data Mining (EDM)'. Siemens and Baker [2012] geven een beschrijving van beide velden. Learning analytics wordt gedefinieerd als het proces waarbij data over student karakteristieken en leergedrag is verzameld en geanalyseerd om zowel het begrip als het leren en de leeromgeving te verbeteren. Educational Data Mining houdt zich bezig met het ontwikkelen van methodes om data uit een educatieve setting te verwerken en dit te gebruiken om studenten en de leeromgeving beter te begrijpen.

EDM richt zich hierbij meer op processen en modellen met een geautomatiseerde toepassing, LAK legt de nadruk op het verbeteren van informatievoorziening en het ontwikkelen van modellen om onderwijzers te informeren en te ondersteunen bij het maken van beslissingen [Siemens and Baker, 2012]. In de context van dit onderzoek past de LAK stroming beter.

Toepassingen van data mining

Er is getracht onderzoeken te bespreken die ofwel veel overeenkomsten vertonen of interessant zijn vanwege de gebruikte data, technieken en/of opzet.

Heijden et al. [2010] coderen studiesucces als ordinale variabele bestaande uit vier categorieën: de minste studiewaarde wordt toegekend aan leerlingen die vertrekken in het eerste jaar, de volgende categorie betreft leerlingen die vertrekken na jaar één of langer dan vier jaar studeren, en de categorieën met de hoogste studiewaarde bevatten leerlingen die een diploma halen in jaar vier en jaar drie. Leerlingen die lang doen over hun opleiding worden samengeplaatst met uitvallers, terwijl de populatie en daarmee gelijkenissen in de groepen niet per se overeen zullen komen (een langstudeerder zal zich waarschijnlijk laten kenmerken door andere eigenschappen dan een leerling die vroeg uitstroomt). Mijns inziens is het een subjectieve indeling waarmee twee verschillende groepen worden samengevoegd, wat de resultaten waarschijnlijk niet ten goede zal komen.

De voorspellingen worden gedaan aan de hand van een regressiemodel en maken gebruik van gegevens van de Faculteit Sociale Wetenschappen Universiteit Utrecht. Er worden door de tijd heen voorspellingen gemaakt waarbij (logischerwijs) de resultaten van tentamens een steeds belangrijkere rol krijgen. Dit zal het geval zijn omdat deze een afspiegeling zijn van de voortgang in het leertraject. Vanuit data mining perspectief is dit onderzoek beperkt: zowel de sample size als het aantal variabelen is vrij klein en ook bijvoorbeeld een onderverdeling in een train en test

set mist. Hiermee is het onderzoek methodologisch onzuiver, en zullen de resultaten in dit licht worden beschouwt.

Dekker et al. [2009] gaan op machine learning gebied verder en gebruiken Decision Trees, logistische regressie en Random Forest modellen om te voorspellen of TUe studenten hun propedeuse binnen drie jaar halen. Ook hier wordt dit gedaan met en zonder gegevens over de studieloopbaan. Wanneer er geen gegevens over resultaten worden meegenomen presteren de modellen tegenvallend: ze presteren niet significant beter dan een 'one-rule' classifier (met als variabele "VWO Science mean") die een accuracy haalt van 68%. Wanneer resultaten worden meegenomen is er wel een significant verschil ten opzichte van de one-rule classifier met 76% tegenover het best presterende Decision Tree model wat 81% goed voorspeld.

Herzog [2006] heeft twee voorspellingen gemaakt voor deelnemers van de Carnegie universiteit, één over de retentie en één over de tijd-tot-afstuderen. Decision Trees en Neurale Netwerken zijn gebruikt om te voorspellen of iemand na een jaar terugkomt voor een tweede jaar, een andere studie kiest of uitvalt. De tijd-tot-afstuderen wordt voorspeld voor 15,457 afgestudeerden tussen 1995 en 2005, en is in de volgende categorieën onderverdeeld: diploma binnen drie of minder jaar, vier jaar, vijf jaar of meer dan vijf jaar (zodat de afhankelijke variabele gebalanceerd is). De retentie op jaarbasis wordt enkel voor het eerste jaar voorspeld, de tijd tot afstuderen enkel voor afgestudeerde leerlingen.

De resultaten van het voorspellen van de retentie vallen tegen. Decision Trees en Neurale Netwerken presteren niet significant beter ten opzichte van het regressie model, wat als uitgangspunt genomen is. Ook wanneer informatie over studievoortgang wordt meegenomen wordt er niet significant beter gepresteerd: de accuracy neemt weliswaar met 10% toe, maar dit gebeurd ook bij het regressie model. Met de rijkheid van de beschikbare variabelen is dit vrij teleurstellend; er zijn demografische en vooropleiding gegevens beschikbaar, academische en campus ervaringen en gegevens over de financiële steun. De auteur haalt zelf aan dat de resultaten tegenvallen, een oorzaak hiervoor wordt niet besproken. Mogelijk beschrijft een lineair model de relaties tussen variabelen al goed en zijn de andere algoritmen niet in staat om niet-lineaire verbanden tussen variabelen te vinden die die iets zeggen over studiesucces.

Voor het voorspellen van de tijd-tot-afstuderen is ook informatie over gevolgde vakken beschikbaar. De beste resultaten worden behaald door het Neurale Netwerk met drie hidden layers en de C5.0 Decision Tree. Opvallend zijn de wisselende resultaten van de verschillende Neurale Netwerken; dit geeft het belang aan van het proberen van verschillende parameters bij verschillende technieken. Het is ook een indicatie dat de methode een lage bias maar hoge variantie vertoont: dit is een mogelijk indicatie dat er sprake is van overfitting: de resultaten dragen niet over en dit geeft minder vertrouwen in nieuwe voorspellingen.

Cortez and Silva [2008] maken voorspellingen over het succes in twee belangrijke toetsen aan de hand van classificatie in een binaire onderverdeling (succes Ja/Nee) en een onderverdeling in vijf categorieën en aan de hand van regressie waarbij de numerieke uitkomst tussen de nul en 20 voorspelt. De data die hiervoor gebruikt wordt is afkomstig van twee scholen in Portugal in het schooljaar 2005/2006 en zijn persoonsgegevens uitgebreid met vragen- en cijferlijsten. Uit de cijferlijsten zijn gegevens beschikbaar over bijvoorbeeld familie, het werk van ouders en alcoholgebruik.

Eén voorspelling wordt gedaan op basis van alleen achtergrondgegevens en de twee andere voorspellingen maken gebruik van de cijfers uit periode 1 en periode 1 en 2. De volgende algoritmen worden toegepast: Neurale Netwerken, SVM, Decision Tree en Random Forest. Wanneer een voorspelling gemaakt wordt op basis van achtergrondgegevens alleen is de accuracy nauwelijks beter dan de referentie waarde waarbij de meest voorkomende klasse wordt voorspeld. Dit resultaat valt tegen, vooral omdat er een ruime set aan variabelen beschikbaar is.

Voorspellen van resultaten in het op afstand leren

Veel toepassingen in de literatuur spitsen zich toe op het op 'afstand' leren. Leerkrachten hebben geen direct contact met leerlingen en daarom meer moeite met het onderscheiden van mogelijke problemen, verder zijn er vaak rijkere gegevens over het online gedrag en gewoontes, zoals het klikgedrag en tijden waarop een platform wordt gebruikt. Een voordeel van deze omgevingen is

dat op het eenvoudiger is om ondersteuning te bieden op de juiste tijd en plek.

Annika Wolff and Zdenek Zdrahal [2012] spelen hierop in: er worden twee voorspellingen gemaakt over resultaten van 7,701 studenten voor een module waarbij veel gebruikt wordt gemaakt van een virtuele leeromgeving. Er wordt voorspeld of de volgende opdracht met een voldoende wordt beoordeeld en of het vak uiteindelijk gehaald wordt. Bij beide voorspellingen presteren Decision Trees beter dan andere technieken. Op of de volgende opdracht wordt gehaald wordt 'goed' gepresteerd: er wordt een precision (het percentage werkelijke uitvallers ten opzichte van het aantal als zodanig aangemerkt) behaald tussen 0.77 en 0.98. Hoe goed dit precies is, is moeilijk in te schatten; het voorspellen van een 'enkele stap vooruit' lijkt geen moeilijke opgave gegeven de vorige resultaten en er wordt geen vergelijking met een baserate gemaakt. Hetzelfde geldt voor de voorspelling van het eindresultaat: hierbij blijkt wel dat naarmate verder gevorderd is in de module het moeilijker is goed te voorspellen wie er nog gaan uitvallen. Andere interessante inzichten zijn dat leerlingen consistent presteren totdat zij een probleem tegenkomen en dat de beste voorspelling niet gemaakt wordt door een vergelijking met een gemiddelde leerling te maken maar door te kijken naar veranderingen in het gedrag. Deze inzichten zijn erg interessant en te gebruiken in dit onderzoek in hoe om mogelijk om te gaan met toetsresultaten. Patronen in de toetsresultaten van een enkele leerling zullen mogelijk interessanter zijn dan een vergelijking van een cijfer met het gemiddelde.

Kotsiantis et al. [2004] voorspellen het wel of niet halen van de eindtoets en maken gebruik van verschillende machine learning algoritmes als Decision Tree, Neurale Netwerken, Naïve Bayes, SVMs en logistische regressie. Hierbij wordt in twee van de drie maatstaven (accuracy en sensivity) het best gepresteerd door het Naïve Bayes algoritme. De precisie is 62% wanneer gebruik gemaakt wordt van alleen de demografische gegevens en 82% net voor de eindtoets. Ook hierbij worden geen baserates gegeven om een vergelijking te maken, wat het moeilijk maakt de meerwaarde in te schatten.

Conclusie

Het algemene beeld is dat wanneer er alleen gebruik wordt gemaakt van demografische gegevens er vaak geen significant betere voorspelling gedaan kan worden ten opzichte van een simpele classifier. Dit schets een dim vooruitzicht: de achtergrondgegevens beschikbaar voor dit onderzoek zijn in vergelijking een stuk schraler. Wanneer er gebruik wordt gemaakt van resultaten schiet dit beeld om. In veel van de gevallen kunnen modellen deze gegevens gebruiken om tot een significant betere voorspelling te komen ten opzichte van een referentiemodel.

Er worden veel verschillende soorten voorspellingen gemaakt: een binaire classificatie van succes, een onderverdeling in categorieën in vele smaken, en ook de aannames zijn vaak (net) anders. Dit alles, samen met dat de doelgroep en het percentage wat succesvol is (de baserate) steeds verschilt, maakt het moeilijk een vergelijking te maken en een eenduidig beeld te schetsen van wat er precies aan prestaties te verwachten is.

Bij de toepassingen in het op afstand leren worden voornamelijk toetsresultaten en eindcijfers voorspelt. Op dit 'lagere' niveau zijn de voorspellingen nauwkeuriger, en spelen voorgaande toetsresultaten een belangrijke rol.

Tussen de onderzoeken worden verschillende algoritmes toegepast. Voornamelijk Decision Trees worden veel toegepast en presteren in veel onderzoeken redelijk. Dat Decision Trees zoveel gebruikt worden is waarschijnlijk omdat het een zo een algoritme is wat direct toepasbaar is: het kan goed omgaan met verschillende soorten variabelen, is in wiskundig opzicht niet ingewikkeld en het opgestelde model is vrij tastbaar. Verder blijkt dat prestaties van technieken tussen de onderzoeken erg verschillen; een techniek die over alle onderzoeken genomen het best presteert is er niet. Prestaties van algoritmen moeten worden beschouwd op een case-by-case basis, prestaties zijn hierbij afhankelijk van de (vorm van de) data. Voor dit onderzoek zal het dus ook van belang zijn hier in te experimenteren.

4. Data

4.1 Data beschrijving

In dit hoofdstuk zal de data die gebruikt is in het onderzoek worden beschreven. De data die is gebruikt komt van meerdere bronnen, zowel intern (Da Vinci college) als extern (open data).

4.1.1 Interne data

De eerste databron is het datawarehouse onderwijs en het studentinformatiesysteem EduArte. Gegevens zijn op verschillende momenten geëxporteerd als Excel bestanden. Het eerste data bestand is geëxporteerd op 9 juli 2015 en is platgeslagen op leerlingenniveau: elk van de in totaal 13,832 regels bevat leerlinggebonden gegevens. Het tweede bestand bevat verbintenissen en is geëxporteerd op 17 september en telt 24,237 regels. Het bevat voor elke deelnemer per jaar voor welke opleiding hij of zij staat ingeschreven. Het derde bestand is geëxporteerd op 18 februari 2016 en bevat de aanwezigheid op leerlingniveau per week uitgesplitst op vak van het cohort 2015/2016. Het vierde bestand is geëxporteerd op 13 April 2016 en bevat meer dan 110,000 resultaten (zowel numerieke als categorische toetsresultaten, maar ook stage-uren en presentie resultaten). De aanwezige variabelen in deze bestanden zijn te vinden in Appendix A.1.

Toetsresultaten

Logischerwijs zullen toetsresultaten in verband staan met studiesucces: leerlingen zullen voldoendes moeten halen om uiteindelijk een opleiding af te ronden. Er zijn toetsresultaten beschikbaar uit drie bronnen: de eerste resultaten zijn geëxporteerd samen met de leerlinggebonden gegevens en zijn summatieve resultaten, de tweede zijn de resultaten van losse niveautesten en de derde resultaten zijn meer dan 110,000 numerieke en categorische toetsresultaten, maar ook stage-uren en presentie resultaten.

Er zijn 42,000 summatieve toetsresultaten van 4,800 unieke deelnemers beschikbaar. Hiervan gaan er 13,000 over kerntaken, de overige resultaten zijn van vakken als Nederlands, Engels en rekenen. De resultaten van deze summatieve toetsen komen vaak pas later in de opleiding beschikbaar en worden vermeld op het diploma. De niveautesten zijn adaptieve meerkeuzetoetsen, en bestaan uit een niveau en een verdere specificatie aan de hand van een percentage. De laatste resultaten zijn van verschillende soorten (formatieve) toetsen en metingen. Deze resultaten zijn eenduidig en beslaan ook niveautoetsen uit het tweede bestand. Dit is het meest omvangrijke bestand met meer dan 110,000 unieke toetsresultaten en is daarom gebruikt om variabelen van te vormen voor het gecombineerde bestand.

4.1.2 Externe data

Het aantal beschikbare variabelen per deelnemer is verder uitgebreid door gebruik te maken van open data. Hiervoor zijn de gebruikte koppelpunten de crebocodes die aan elke opleiding worden toegewezen en de postcode.

Open data

De rijksoverheid[1] omschrijft open data als data die openbaar is, waar geen auteursrecht op berust, die bekostigd is uit publieke middelen en bij voorkeur voldoet aan 'open standaarden' om gebruik en toegang makkelijk te maken. Deze definitie is vrij nauw: enkel publiekelijk bekostigde data wordt aangemerkt als open data. De Open Knowledge Foundation [2016] geeft de volgende 'Open Definitie': "Open data is data die vrij gebruikt kan worden, hergebruikt kan worden en opnieuw verspreid kan worden door iedereen – onderworpen enkel, in het uiterste geval, aan de eis tot het toeschrijven en gelijk delen."

Frankowski et al. [2015] maken de waarneming dat de meeste definities terug te leiden zijn tot drie kenmerken: open data is technisch, juridisch en economisch open. Technisch open houdt in dat data in een formaat beschikbaar is wat goed door computers wordt ondersteund, juridisch open houdt in dat de data door iedereen gebruikt kan worden (is niet juridisch beschermd), en economisch open houdt in dat data gratis of tegen een (lage) kostprijs beschikbaar is. Deze drie elementen spelen een rol bij open data, maar kunnen beter gezien worden als dimensies dan voorwaarden. De gebruikte externe databronnen zullen worden beoordeeld over deze dimensies.

Gebruikte open data

Als eerste koppelpunt is de postcode gebruikt. Aan deze postcode zijn twee bestanden gekoppeld: postcode data van `www.postcodedata.nl/` en SES statusscores.

De postcode data is beschikbaar op `www.postcodedata.nl/download` en bevat per 6-positie postcode de straatnaam, stad, gemeente, provincie en de geografische locatie (breedte- en lengtegraad) en waarop deze gebaseerd is. Deze data wordt maandelijks bijgewerkt en kan worden gedownload of worden opgevraagd met een API.

De gegevens zijn samengesteld uit "de betrouwbaarste bronnen": Publieke Dienstverlening Op de Kaart (PDOK), de Kamer van Koophandel, het Centraal Bureau voor de Statistiek en de BAG van het Kadaster. Er wordt gesteld dat de gegevens uit deze bronnen voldoen aan de open data standaarden. Garanties over de betrouwbaarheid en beschikbaarheid worden niet gegeven; ook is niet bekend welke gegevens uit welke bronnen komen en hoe de data wordt samengesteld.

De statusscores zijn bepaald door het SCP en weerspiegelen de sociale status van wijken (4-positie postcode) met meer dan 100 huishoudens. De statusscores volgen uit een principale componentenanalyse waarin het gemiddelde inkomen, het percentage mensen met een laag inkomen, het percentage laag opgeleiden en het percentage mensen dat niet werkt worden samengesteld tot één score (het eerste principale component). De hiervoor gebruikte gegevens zijn afkomstig van EDM BV dat op zijn beurt ook gebruik gemaakt van verschillende publieke en private databronnen. Per 4-positie postcode is de indicatieve naam beschikbaar samen met de statusscore, de bijbehorende rangorde, het aantal huishoudens en het bevolkingsaantal. Elk gebied omvat ongeveer 1,825 huishoudens en is beschikbaar (mits het minimum van 100 huishoudens wordt gehaald) voor de jaargangen 1998, 2002, 2006 en 2014; het zou dus mogelijk zijn om een trend in sociale status te bepalen.

In het verbintenissen bestand staat een crebocode en daarnaast al enige informatie over de bijbehorende opleiding. Informatie over een opleiding is verder uitgebreid door een koppeling te maken met de koppeltabel beschikbaar gemaakt door DUO[2]. Dit bestand bevat gegevens uit de koppeltabel SBB 2014/2015 zoals gepubliceerd op `http://kwalificatiesmbo.nl/` en is gecombineerd met gegevens uit het Creboregister, de gegevens zijn gebaseerd op CREBO (Centraal Register Beroepsonderwijs) afkomstig uit BRIN (Basisregister Instellingen). Per crebocode zijn onder andere bekend: de kwalificatie naam, domeinnaam, prijsfactor, en verder het aantal studiebelastingsuren en aan welk kenniscentrum de opleiding toebehoort.

De SES statusscores worden eens in de vier jaar berekend. De databronnen die worden gebruikt om tot deze gegeven te komen zijn niet bekend. Het SCP gegevens gebruikt van EDM BV, echter welke bronnen zij hiervoor gebruiken is ook op aanvraag niet beschikbaar. De postcodedata wordt

[1] `data.overheid.nl/`

[2] `duo.nl/open_onderwijsdata/databestanden/mbo/Crebo/`

iedere maand vernieuwd en is samengesteld uit de eerder benoemde "betrouwbaarste bronnen". De crebokoppeltabel wordt ieder jaar beschikbaar gemaakt door het DUO in samenwerking met andere grotere instellingen in het beroepsonderwijs.

Koppelpunt	Postcode (6-delig)	Postcode (4-delig)	Crebocode
Bron	www.postcodedata.nl/	SES Statusscores	DUO Crebokoppeltabel
Manier van verkrijgen	Download & API	Op aanvraag	Download
Open data Dimensies			
Technisch open	xls, sql & csv formaat	xls, dat, sav & csv formaat	xls & csv formaat
Juridisch open	Ja	Ja	Ja
Economisch open	Gratis	Gratis	Gratis

Tabel 4.1: Gebruikte open data bronnen; het koppelpunt en beoordeling in de open data dimensies

Mogelijke uitbreidingen open data bronnen

Er zijn een aantal alternatieve data bronnen niet gebruikt die mogelijk wel toegevoegde waarde kunnen leveren.

Eén mogelijkheid is Studie in Cijfers[1] (ook bekend als 'studiebijsluiter'). Studie in Cijfers geeft een objectief beeld over een opleiding en de kansen van die opleidingen op de arbeidsmarkt. Daarnaast geeft het aan hoe een opleiding bij een bepaalde onderwijsinstelling het doet in vergelijking met andere instellingen. Een nadeel is dat niet alle scholen meedoen[1] en de data die gebruikt wordt voor dit onderzoek afkomstig is van één van de scholen die niet meedoet. Daarnaast hebben we geen locatie gevonden waar deze bijsluiters in een gestructureerde vorm opvraagbaar zijn en staat op de bijsluiter enkel de opleidingnaam, die vaak soms kleine afwijkingen kent (zoals een nadere toelichting tussen haakjes, een afkortingen en 'en' in plaats van &) en geen crebocode.

Een tweede bron is de keuzegids[2]. Deze keuzegids wordt elk jaar beschikbaar gemaakt en geeft voor een kostprijs een overzicht van mbo-studies per vakgebied en per niveau ingedeeld in regio's. Hierbij wordt informatie gegeven over mogelijke beroepen, baankansen, het gemiddelde startsalaris maar ook informatie over de studie zoals doorstroom en het oordeel van leerlingen en de onderwijsinspectie. De gids is echter niet zo opgesteld dat deze eenvoudig programmatisch kan worden doorzocht om informatie uit af te leiden en ook de online versie kent (nog) geen zoeksysteem. Daarnaast zijn ook hier enkel de (net afwijkende) opleidingsnamen beschikbaar en geen crebocodes, wat een koppeling verder bemoeilijkt.

4.2 Data analyse

In deze sectie zullen we verder ingaan op de gebruikte data. We zullen beginnen met een beperkte analyse van de losse bestanden, gevolgd door hoe de data voor het voorspelmodel is samengesteld en een analyse van de gecombineerde data waarna we afsluiten met een toelichting op enkele aannames.

De analyse van de aparte bestanden is beperkt gehouden omdat de dataset die wordt gebruikt voor het voorspelmodel van meeste interesse is. Bij het samenstellen wordt eerst een selectie gemaakt, daarnaast komen leerlingen één keer voor in het leerlingbestand, kunnen zij meerdere keren voorkomen in de verbintenissen en worden deze weer opgedeeld in hetzelfde of minder aantal voorbeelden voor de gecombineerde dataset.

[1] https://www.s-bb.nl/onderwijs/studie-cijfers
[2] http://www.keuzegids.org/

4.2.1 Losse datasets

Van de 13,832 leerlingen met leerlinggegevens komen er 12,426 terug in het bestand met de verbintenissen. De 1,406 missende leerlingen die niet in de verbintenissen voorkomen zijn ingestroomd in het cursusjaar 2010/2011; voor deelnemers waar wel verbintenissen van bekend zijn maar geen leerlinggegevens is geen patroon te vinden. Een verdere discrepantie zit in de eerste opleiding te vinden in de leerlinggegevens: 1,012 van de eerste opleidingen komen niet terug in de verbintenissen.

De achterliggende oorzaak van deze missende verbintenissen is onbekend. Dat er verbintenissen missen blijkt ook uit het percentage stapelaars: volgens het onderwijsverslag 2012/2013 zou ongeveer 29 procent van de gediplomeerde studenten het mbo in met meer dan één diploma verlaten. De hier beschikbare gegevens laten echter een percentage van onder de 10% zien. Er is niet te zeggen of de verbintenissen die missen gaan over succesvolle of niet-succesvolle leerlingen. Dit heeft tot gevolg dat de voorspellingen die gemaakt zullen worden over toekomstige leerlingen te positief of negatief kunnen zijn. Verder is het mogelijk dat er patronen zitten in de missende verbintenissen die nu niet worden opgepakt. Door de onzekerheid over de missende data is het echter lastig hier harde uitspraken over te doen.

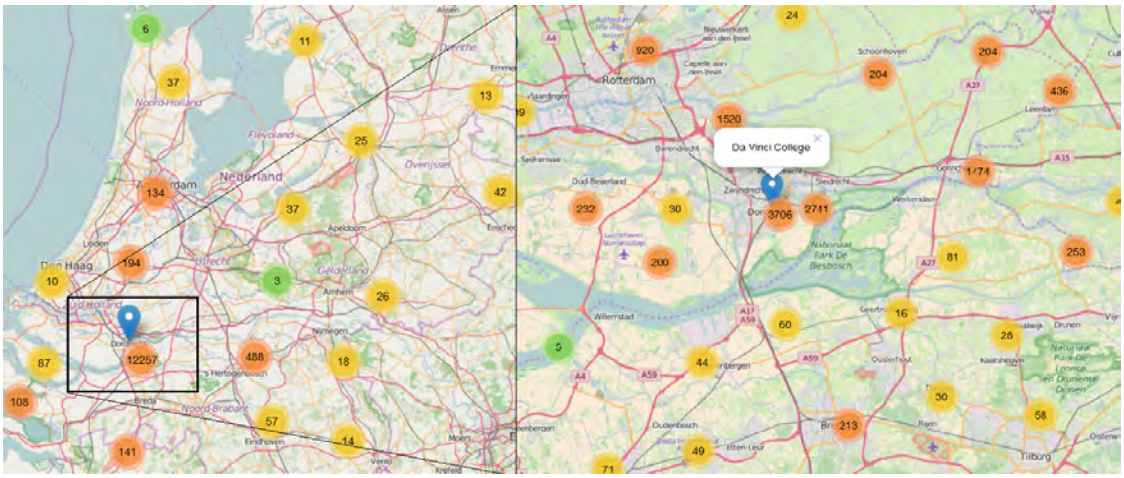
Leerlinggegevens

Per leerling zijn gegevens beschikbaar zoals geslacht, leeftijd, postcode en nationaliteit. Deze gegevens worden uitgewisseld met het BRON en hierbij gevalideerd, dit geeft dus een zekere waarborg over de correctheid van deze gegevens.

Per leerling is een 6-cijferige postcode beschikbaar; voor tien leerlingen is enkel het 4 cijferig deel beschikbaar, verder zijn er twee incorrecte postcodes: “SW 17 9 N” en “B-2381”. Deze verkeerde waarden zijn vervangen door de meest voorkomende postcode. Statusscores zijn enkel beschikbaar voor wijken met meer dan 100 inwoners, bij leerlingen waar de statusscore mist is deze ingevuld met de mediaan.

De postcodes zijn met gegevens van `www.postcodedata.nl/` omgezet in een locatie. De meeste leerlingen die onderwijs volgen komen uit de buurt, zoals te zien in Figuur 4.1. Hierdoor zal de toegevoegde waarde van de geografische gegevens hoogstwaarschijnlijk minder zijn dan wanneer de leerlingen meer verspreid zouden zijn.

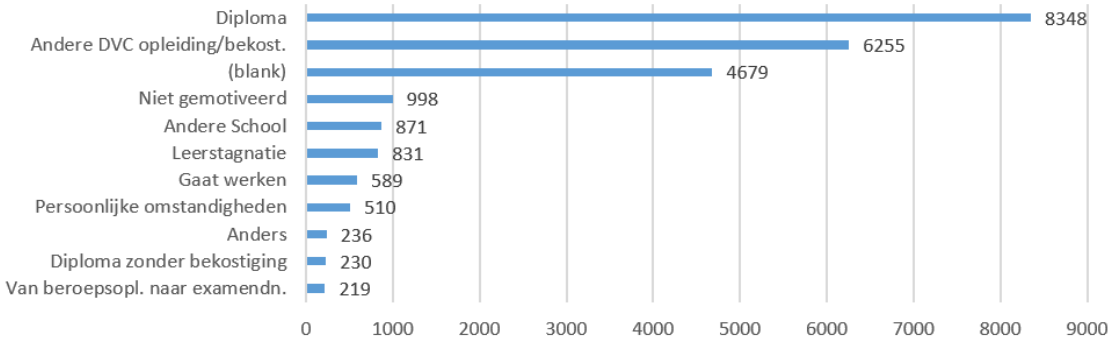
Verder kenmerkt het leerlingenbestand zich als volgt. De meerderheid is mannelijk (55%) en en valt in de leeftijdsgroep van 17 tot 23 jaar (75%). Het merendeel heeft als nationaliteit Nederlandse (10,459), gevolgd door Turkse (972), Nederlands Antilliaanse (657) en Marokkaanse (614). Er zijn 118 unieke geboortelanden, het overgrote deel komt uit Nederland (11,920), gevolgd door de Nederlandse Antillen (368) en Afghanistan (146). Een klein aandeel is missend: voor het geboorteland missen 90 waarden, voor de nationaliteit 88.



Figuur 4.1: Woonplaats leerlingen

Verbintenissen

Het verbintenissen bestand bevat per deelnemer de verbintenissen, of deze nog loopt of wanneer en waarom deze is beëindigd. De verdeling van de reden voor uitschrijven met meer dan 200 vermeldingen is te vinden in Figuur 4.2, hierbij zijn de 4,679 (blank) waarden de lopende verbintenissen. Een groot deel van de redenen voor uitschrijven is "Andere DVC opleiding/bekost.", deze wordt toegekend niet alleen bij het wisselen van opleiding maar ook bij specialiseren, het wisselen van leerweg of intensiteit (voltijd - deeltijd). Het aantal verbintenissen per deelnemer is te vinden in Tabel 4.2. Hier zijn ook leerlingen in opgenomen met een nog lopende verbintenis.



Figuur 4.2: Frequentie redenen voor beëindiging verbintenis

Verbintenissen	1	2	3	4	5	6	7	8
Aantal	10,603	4,083	1,304	286	67	8	3	1

Tabel 4.2: Aantal verbintenissen per deelnemer

Opleidingen worden op twee manieren onderverdeeld en deze manieren kennen een zekere overlap. Ten eerste is er de interne organisatiestructuur, deze is eerst onderverdeeld naar sector, vervolgens naar team en als laatste naar opleidingscluster en ten tweede is er de externe crebo-structuur: deze maakt een onderverdeling naar hoofdgroep, subgroep, en vervolgens beroepsopleiding. Per verbintenis beschrijven categorische variabelen hoe de opleiding hierin is gepositioneerd.

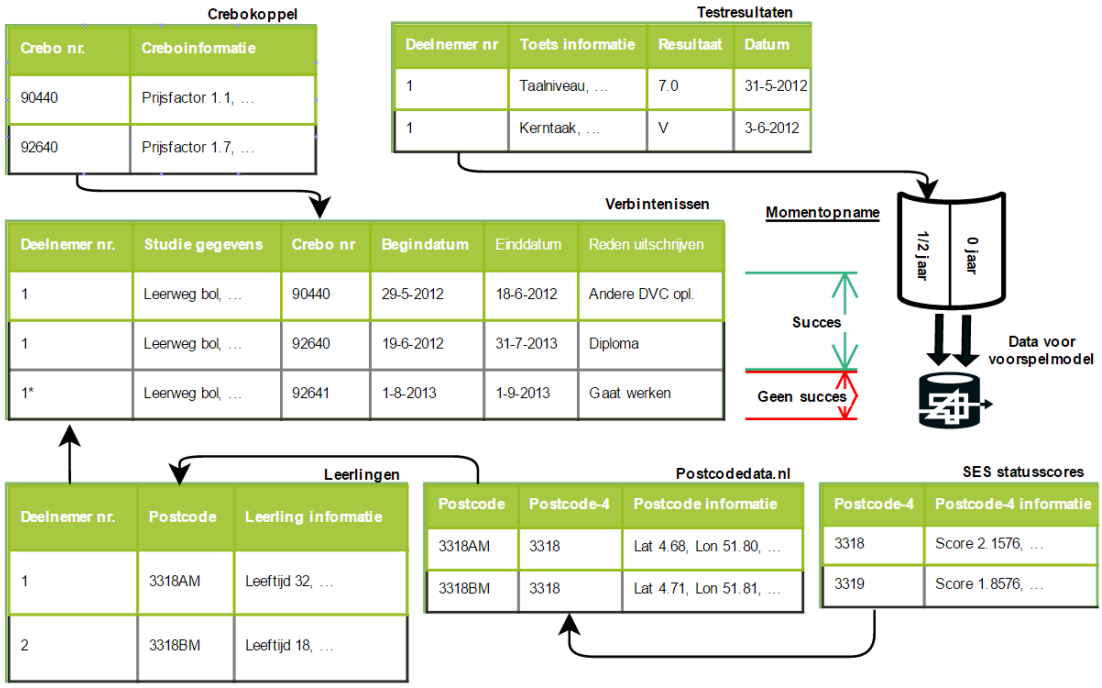
4.2.2 Klaarmaken samengestelde dataset

In Figuur 4.3 is weergeven hoe tot de samengestelde dataset is gekomen. In het klaarmaken van de data voor het model zijn een aantal aannames gemaakt. Ten eerste worden enkel de deelnemers meegenomen die voorkomen in zowel de verbintenissen als de leerlinggebonden export. De variabelen in het leerlinggebonden bestand zijn niet te herleiden zijn tot de variabelen die een verbintenis opmaken. Vandaar dat de verbintenissen die niet overeenkomen niet zijn toegevoegd. Per deelnemer worden alle verbintenissen tot het behalen van een diploma samengenomen op basis van einddatum. De verbintenissen die hierna volgen worden toegedeeld aan een 'nieuwe' deelnemer. Er zijn 1,102 stapelaars opgedeeld, hiervan komen 433 van niveau één, 429 van niveau twee, 236 van niveau drie en doen vier een nieuwe opleiding op niveau vier.

Zolang één van de redenen van uitschrijven van de samengenomen verbintenissen gelijk is aan het behalen van een diploma (zonder bekostiging), is de afhankelijke variabele 'succes'. De duur ratio die wordt voorspeld is het aantal dagen tussen de begindatum van de eerste verbintenis en de einddatum van de laatste verbintenis, gedeeld door het verschil in dagen tussen de begindatum en de verwachte einddatum van de laatste verbintenis.

Er is gebruik gemaakt van de verwachte einddatum en niet de nominale studieduur. Dit geeft een beter beeld omdat hierin de vooropleiding wordt meegenomen, zo zijn er opleidingen met

als nominale duur vier jaar, maar met een werkelijke duur van één jaar gegeven een bepaalde vooropleiding. Met randgevallen zoals het tegelijkertijd volgen van meerdere opleidingen en waarbij de tweede studie wordt afgerond voor de eerste studie, wordt niet expliciet rekening mee gehouden.



Figuur 4.3: Overzicht hoe de samengestelde dataset wordt klaargemaakt

Aan de hand van Figuur 4.3 zullen we een voorbeeld geven van hoe de data wordt gecombineerd. De eerste twee kolommen in het verbintenissen bestand worden gecombineerd tot een enkel voorbeeld, de derde kolom wordt opgenomen met een aangepast deelnemernummer. Bij een eerste momentopname aan het begin van de studie worden leerlinggegevens in combinatie met de eerste opleiding gebruikt. Bij volgende momentopnames wordt de studie gebruikt die op dat moment gevolgd wordt, samen met empirische variabelen zoals het aantal overeenkomsten. Ook worden er variabelen afgeleid van de tot dan toe behaalde resultaten. Dit betekent voor het voorbeeld in Figuur 4.3 dat bij een momentopname na een half jaar de eerste deelnemer zal worden opgenomen met crebocode 92640 en met als aantal overeenkomsten twee, de 'tweede' (1*) deelnemer zal niet terugkomen omdat deze binnen een half jaar is uitgestroomd.

Er zijn een aantal variabelen gevormd die iets zeggen op opleiding- of teamniveau. Gemaakte variabelen zijn de grootte van de opleiding, de grootte van het team en de begrote loonkosten per FTE per team per jaar. Ook de gemiddelde duur en het gemiddelde succespercentage zijn toegevoegd. Daarnaast is per team ook het gemiddelde aanwezigheidspercentage en het percentage ongeoorloofd afwezig toegevoegd. Andere variabele zijn afgeleid door beschikbare gegevens te combineren. Zo is het aantal studiebelastingsuren wat staat voor de opleiding gedeeld door de verwachte duur, hetzelfde is gedaan met het aantal gesprekken m.b.t. studievoortgang en verzuim. Verder is er nog een binaire variabele opgenomen die aangeeft of een studie een verzwaarde intake kent. Of een studie een verzwaarde intake kent is te vinden op de site, maar is ook opgenomen in het verbintenissen bestand.

Uit de resultaten zijn de volgende variabelen afgeleid: het aantal gemaakte toetsen, het aantal extra pogingen (herkansingen), het laagste en hoogste resultaat, de standaarddeviatie en het gemiddelde, de gemiddelde afwijking van het gemiddelde over een toets genomen, het percentage van de toetsen dat voldoende is en de toets frequentie: het aantal toetsen gedeeld door het aantal dagen tussen de eerste en de laatste toets.

4.2.3 Analyse samengestelde dataset

Kanttekening samengesteld bestand

We beginnen met een kanttekening bij de samengestelde dataset. Over het algemeen doen leerlingen die stoppen hier korter over dan leerlingen die uitstromen met een diploma. Verder hebben leerlingen die later beginnen en een lange opleiding een grotere kans om nog bezig te zijn in vergelijking met leerlingen die een kortere opleiding volgen en/of al lang bezig zijn. Dit heeft tot gevolg dat het moment van beginnen aan de opleiding en de opleidingsduur van invloed zullen zijn op het succespercentage. Dit komt door de manier waarop de data wordt klaargemaakt (er is een 'afkapping' op het moment van exporteren) en omdat een beperkte historie beschikbaar is. Dit heeft tot gevolg dat het kan lijken alsof variabelen voorspellend vermogen hebben, terwijl dit enkel ligt aan het feit dat zij een andere 'baserate' kennen vanwege de samenstelling van wanneer er is begonnen met een opleiding en hoe lang de opleiding duurt.

Dit kan het best worden geïllustreerd aan de hand van een variabele die aangeeft of de vooropleiding beschikbaar is. Op basis van de verdeling succes/niet succes binnen de verschillende waarden voor deze variabele lijkt het een goede voorspeller: de waarde "Niet Beschikbaar" correspondeert met een succespercentage van 62.1%, "Nee" met 51.0% en "Ja" met een percentage van 41.9%. Dit lijkt vreemd: het niet beschikbaar zijn van gegevens correspondeert met een hoger succespercentage. Wanneer de conditionele kans wordt bepaald en rekening gehouden wordt met de samenstelling van het instroomcohort ontstaat een heel ander beeld. Op basis van het instroomcohort is het verwachte verwachte succespercentage voor de groep waar de variabele de waarde "Niet Beschikbaar" aanneemt 64.6%, voor de groep "Nee" 48.2% en voor "Ja" 48.9%. De variabelen die hier het meest door worden beïnvloed zijn de variabelen die direct iets zeggen over de duur en het cohort en verder de variabelen die gaan over de vooropleiding.

Er zijn verschillende methoden onderzocht om met de 'verschuiving' van variabelen om te gaan. Ten eerste is geprobeerd bij het maken van voorspellingen algoritmen te gebruiken die kunnen omgaan met missende waarden. Echter bleek dat deze technieken uit de aanwezigheid van een variabele afleiden dat een leerling uit een recenter cohort komt of een langere studie volgt, wat tot gevolg heeft dat de kans op succes (de 'baserate') laag is. Dit geleerde patroon zal echter niet generaliseren.

Verder is er in de preprocssing getracht de vaak missende waarden te imputeren met onder andere kNN imputatie en mice. Ook hier vielen de resultaten tegen vanwege het grote aantal missende waarden, hierop zal worden terugkomen in Sectie 5.2. Ook is geprobeerd met behulp van factor analyse technieken een onderliggende structuur (de zogenaamde latente variabelen) inzichtbaar te maken. Dit leverde echter geen eenduidige structuur op met de onderliggende cohorten en vermoeilijkt enkel de interpretatie verder. Een nog andere mogelijkheid was om per cohort een model op te stellen, of verschillende modellen op te stellen met enkel de gegevens die wel beschikbaar zijn. Uiteindelijk is gekozen voor de laatste optie. Hoewel het mogelijk is dat de gevonden patronen niet generaliseren omdat de modellen gebaseerd worden op een bepaalde groep, is de interpretatie het eenvoudigst en maakt het een directe vergelijking van resultaten eenvoudiger.

		Opleiding duur in maanden			
		12	**24**	**36**	**48**
Cohort	**2009/2010**	-	$3/3 = 1$	$5/7 = 0.714$	$21/27 = 0.778$
	2010/2011	$68/78 = 0.872$	$149/183 = 0.814$	$232/290 = 0.8$	$201/249 = 0.807$
	2011/2012	$349/460 = 0.759$	$838/1314 = 0.638$	$406/650 = 0.625$	$237/468 = 0.506$
	2012/2013	$336/472 = 0.712$	$807/1394 = 0.579$	$368/676 = 0.544$	$89/243 = 0.366$
	2013/2014	$267/352 = 0.759$	$616/1345 = 0.458$	$115/358 = 0.321$	$14/148 = 0.095$
	2014/2015	$153/234 = 0.654$	$136/472 = 0.288$	$18/255 = 0.071$	$1/51 = 0.02$

Tabel 4.3: Kans op succes gegeven vooropleiding en studieduur van leerlingen die al klaar zijn; over het algemeen ligt deze lager bij recentere cohorten en bij opleidingen met een langere duur.

Niveau	Aantal	Succes %	Gem. duur	Gem. ratio
1	1529	73.6%	385.66	0.97
2	3411	52.4%	596.71	0.85
3	2432	53.2%	661.87	0.75
4	2095	50.5%	708.24	0.58

Tabel 4.4: Invloed van niveau op studiesucces en duur

Afstand	Aantal	Succes #	Succes %
0 - 2 km	1455	765	52.6%
2 - 5 km	2147	1171	54.5%
5 - 25 km	4174	2336	56.0%
25+ km	1691	979	57.9%

Tabel 4.5: Invloed van afstand op studiesucces

Voorspellingen

Naast succes wordt ook de duur ratio voorspeld, dit is de duur ten opzichte van de verwachte duur. Een ratio van minder dan één houdt in dat de leerling zijn opleiding sneller afrondt dan gepland, een ratio van boven de één geeft aan dat de leerling vertraging oploopt. De verdeling van deze ratio is visueel weergegeven in Figuur 4.4, hier is echter al wel een opdeling gemaakt op basis van intensiteit en leerweg.

Over alle leerlingen genomen is het percentage studiesucces 55.5%. Afhankelijk van de leerweg, richting en het opleidingsniveau ligt dit percentage net anders. Zo ligt het succespercentage van deeltijdleerlingen met 57.1% een stuk hoger dan de voltijdleerlingen (52.8%), uit de groep van 292 examendeelnemers haalt maar liefst 94.5% een diploma. De leerweg is van minder invloed: het succespercentage ligt bij leerlingen uit de bol op 54.5% en bij leerlingen uit de bbl op 57.6%. De invloed van niveau op studiesucces is te zien in Tabel 4.4. Voor de entreeopleidingen op niveau één ligt dit percentage erg hoog: 73% haalt een diploma. Voor de overige niveaus zijn de succespercentages nagenoeg hetzelfde.

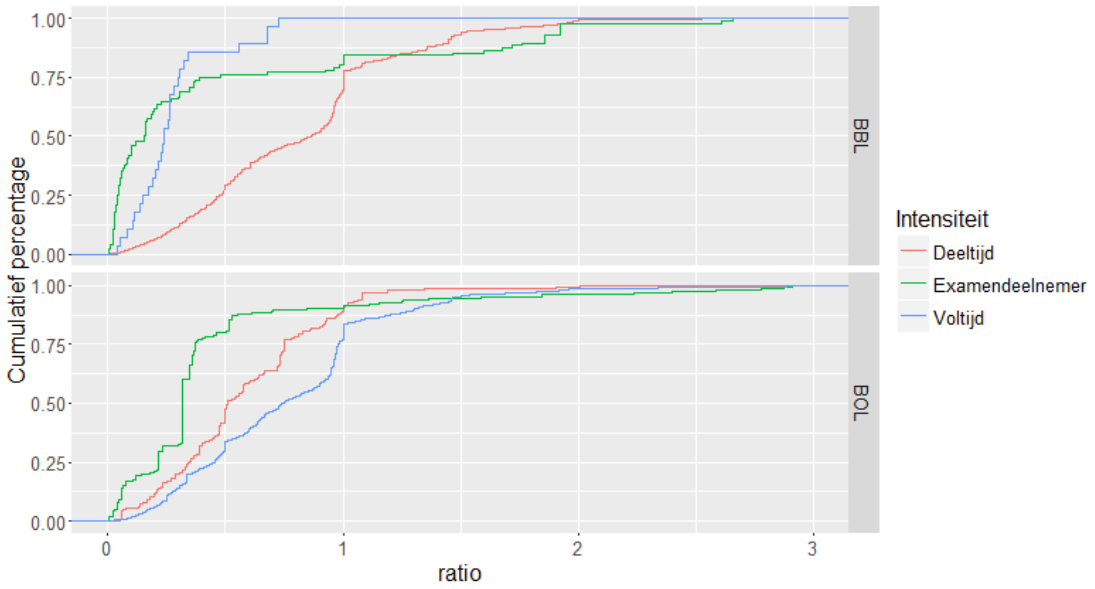
Verder blijkt dat vrouwelijke leerlingen vaker een diploma dan mannelijke (58.7% t.o.v. 53.0%). Uit gemeenten waar meer dan 75 leerlingen vandaan komen om onderwijs te volgen presteren de leerlingen uit Zaltbommel en Zederik het beste, met een succespercentage van respectievelijk 73.3% en 72.1%. Verder lijkt ook afstand tot de instelling van invloed te zijn: hoe verder weg een leerling woont hoe hoger het percentage van leerlingen dat een diploma haalt wordt, zie Tabel 4.5.

Voor de duur zien we in Tabel 4.4 dat een hoger niveau correspondeert met een langere duur. Ook de intensiteit is van invloed: Examendeelnemers doen gemiddeld 173 dagen over hun opleiding en hebben een gemiddelde ratio van 0.49. Deeltijdleerlingen hebben een gemiddelde duur van 587 t.o.v. voltijdleerlingen met 633 dagen, de ratio ligt bij beiden rond de 0.8. Dit is visueel weergegeven in Figuur 4.4: hierin is het cumulatieve aandeel van de ratio geplot voor iedere combinatie van leerweg en intensiteit. Hierbij valt op dat een groot deel van de examendeelnemers sneller klaar is dan gepland en er vervolgens een aantal zijn die er (veel) langer over doen dan gepland. Verder waar in de bbl de voltijdleerlingen vrijwel altijd sneller klaar zijn, lijkt het omgekeerde te gelden voor de bol: hier zijn de deeltijdleerlingen in vergelijking met de voltijdleerlingen relatief sneller klaar dan gepland.

4.2.4 Beperkt beschikbare gegevens

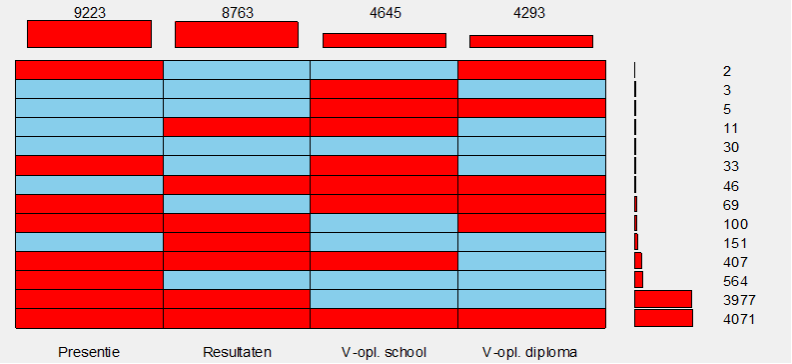
Een aantal variabelen zijn slechts beschikbaar voor een selecte groep leerlingen. Hieronder vallen de variabelen die gaan over vooropleiding, en variabelen die worden afgeleid bij de momentopname na een half jaar. Bij de momentopname na een half jaar is meer informatie beschikbaar voor een beperkt deel van de leerlingen. Van de deelnemers die na een halfjaar nog bezig zijn, zijn de combinaties van aanwezige variabelen weergeven in Figuur 4.5. In dit figuur geven de onderste kolommen de grootste groepen aan: voor 4,071 van de leerlingen is geen informatie beschikbaar over presentie, noch over resultaten en vooropleiding. Voor 3,977 leerlingen zijn beide gegevens over de vooropleiding beschikbaar. Voor slechts 706 leerlingen zijn gegevens over de resultaten beschikbaar en voor slechts 246 leerlingen is informatie over presentie aanwezig na een half jaar. Voor slechts 30 leerlingen zijn er geen missende waarden.

Naast dat de resultaten aanwezig zijn voor een zeer selecte groep, bestaat deze groep vrijwel enkel uit leerlingen uit recentere cohorten. De groep waarvoor gegevens over de vooropleiding



Figuur 4.4: Ratio ($= \frac{\text{duur}}{\text{verwachte duur}}$) afhankelijk van leerweg (boven: bbl, onder: bol) en intensiteit

beschikbaar zijn betreffen enkel de leerlingen die niet voor cohort 2012/2013 van start zijn gegaan. Van de 706 leerlingen met toetsresultaten komen er 273 uit cohort 2014/2015 en 236 uit cohort 2013/2014. In deze groep zijn de leerlingen die klaar zijn vaak uitgestroomd zonder diploma, waardoor het percentage studiesucces slechts op 34.4% ligt. Informatie over aanwezigheid is voor nog een kleiner deel beschikbaar, de samenstelling van deze selectie is echter wel meer gebalanceerd verdeeld over de cohorten en is het slagingspercentage in deze groep ligt op 67.1%.



Figuur 4.5: Aanwezigheid van combinaties van variabelen bij de momentopname na een half jaar

Datakwaliteit

De kwaliteit van de data is belangrijk. Als er veel missende waarden zijn of gegevens niet kloppen, dan zal het leren van patronen minder goed gaan. Het belang van (het nadenken over) data kwaliteit wordt ook benadrukt door Chapman [2005]: (Te) vaak wordt data gebruikt zonder kritieke blik en na te denken over de fouten die de data bevat en de gevolgen hiervan.

Datakwaliteit kan worden gezien aan de hand van verschillende attributen, [Pipino et al., 2002] geven hiervoor een lijst aan dimensies zoals beschikbaarheid, accuraatheid en compleetheid. Een aantal hiervan zijn hier niet van toepassing omdat het gaat om al geëxporteerde, geaggregeerde data. Strong et al. [1997] benadrukken dan ook dat data kwaliteit gezien moet worden vanuit

het perspectief van de gebruiker. De karakteristieken van hoge kwaliteit data kunnen hierbij worden onderverdeeld in de categorieën intrinsieke- en contextuele kwaliteit, toegankelijkheid en interpreteerbaarheid.

Intrinsieke kwaliteit gaat over conflicterende data en de kwaliteit van de data op zichzelf, en is al aan bod gekomen in deze sectie. De leerlinggegevens en algemene gegevens over de opleiding bevatten slechts een klein aantal missende waarden. Voor de resultaten en presentie is slecht een beperkt deel aanwezig. Verder zit er een discrepantie tussen zowel de deelnemers als de verbintenissen in de verschillende bestanden.

Toegankelijkheid gaat over technische toegankelijkheid maar ook volume. De data is al geëxporteerd in een toegankelijk formaat, en de hoeveelheid bezorgt ook geen problemen. Contextuele data kwaliteit gaat over missende variabelen en niet goed gedefinieerde of gemeten data. In de context van data mining zijn meer variabelen wenselijk. Voor dit onderzoek zouden dit gegevens betreffen als het aantal verschillende adressen of meer achtergrondgegevens. Ook zouden bepaalde variabelen van meer waarde zijn in een andere vorm, een voorbeeld hiervan zijn de bijzonderheden (adhd, handicap, etc.) in plaats van het aantal bijzonderheden.

De gebruikte gegevens over de leerlingen zullen ook beschikbaar zijn bij andere instellingen; het zijn de gegevens die moeten worden geleverd aan bevoegd gezag volgens de WEB (Artikel 2.5.5a.). De leerlinggegevens zijn afkomstig uit het Basisregister Onderwijsnummer (BRON) van DUO. De kwaliteit van deze gegevens wordt geborgd door het Programma van Eisen Beroepsonderwijs en Volwasseneneducatie waarin is vastgelegd hoe onderwijsgegevens op een eenduidige manier worden ontsloten. De gegevens van een leerling zijn ofwel gekoppeld aan een BSN of wanneer de leerling geen BSN wil doorgeven, een onderwijsnummer.

4.2.5 Aannames en toelichting gemaakte keuzes

Studiesucces bestaat uit meerdere dimensies die op verschillende manieren kunnen worden uitgelegd. In Sectie 3.2 zijn verschillende toepassingen en mogelijkheden besproken. Voor dit onderzoek is de keuze gemaakt om een tweedelige voorspelling te doen. Er is voor gekozen om een fractie te voorspellen omdat deze voor elke leerling in hetzelfde bereik ligt, verder is dit in lijn met de interpretatie van de verwachtingen over een enkele leerling: zal er korter of langer over de opleiding worden gedaan? Met de combinatie van de twee voorspellingen kan de (voorspelde) studiewaarde worden gekwantificeerd. Hierbij moet de kanttekening geplaatst worden dat een onderlinge afhankelijkheid bestaat tussen de afhankelijke variabelen: wanneer iemand een diploma haalt is de kans groter dat hier langer over gedaan wordt.

Er is gekozen om een voorspelling te maken op twee momenten met de informatie die dan beschikbaar is. De eerste voorspelling is aan het begin van de opleiding en de tweede na een half jaar. Dit tweede moment is gekozen zodat het niet te vroeg valt en er weinig nieuwe informatie beschikbaar is, maar ook niet te laat, zodat er nog kan worden gehandeld op basis van de uitkomsten. Verder zijn er geen gegevens die na deze opname beschikbaar komen die een grote impact zullen hebben op de voorspelling, ook is een steeds grotere groep uitgestroomd en neemt daarmee de meerwaarde af.

Voor een aantal gegevens, zoals aan het aantal gesprekken m.b.t. afwezigheid en of de leerling wordt begeleid, is niet bekend welke periode zij beslaan, of wanneer zij beschikbaar komen. Het betreffen een waarneming op een punt in de tijd en er is geen vaste plek in waar deze variabelen van waarde veranderen. Bij een momentopname op een willekeurig tijdstip zal een deel van deze gegevens zijn ingevuld en te gebruiken zijn in een voorspelmodel. Om toch een vergelijking te zou ervoor gekozen kunnen worden deze variabelen op te nemen bij de momentopname, ondanks dat niet zeker is of deze dan daadwerkelijk (in deze vorm) beschikbaar zouden zijn. Er is uiteindelijk voor gekozen deze gegevens niet te gebruiken; omdat onbekend is welke periode de aantallen beslaan zijn ze minder bruikbaar, verder zal de variabele geen oorzakelijk verband beschrijven; de groep die begeleiding krijgt, krijgt dit voor een reden.

Onderverdeling variabelen in categorieën

De variabelen die gebruikt worden zijn onverdeeld in acht categorieën, de onderverdeling is terug te vinden in Appendix A.3.

Bij de momentopname aan het begin van de studie zijn variabelen uit drie van de acht categorieën beschikbaar: de variabelen die gaan over de opleiding, leerlinggebonden variabelen en variabelen die het gemiddelde studiesucces van een opleiding uitdrukken. Deze variabelen zijn beschikbaar voor vrijwel elke leerling, en bevatten slechts een zeer beperkt aantal missende waarden.

Voor de leerlinggebonden variabelen zijn gegevens beschikbaar zoals het geslacht, de leeftijd, de nationaliteit en het geboorteland en verder een aantal gebaseerd op woonplaats zoals de postcode met bijbehorende lengte- en breedtegraad, statusscore en de afstand tot de instelling.

De opleidingsgegevens betreffen de variabelen die de opleiding plaatsen in de interne- en crebo structuur. Een aantal variabelen zijn hierbij niet gebruikt; deze zijn op een te laag niveau zoals de crebocode en de kwalificatienaam. Deze keuze is gemaakt omdat het het lastig maakt het model in te zetten voor nieuwe opleidingen. De derde categorie zijn de variabelen die iets zeggen over het gemiddelde studiesucces binnen een opleidingsteam.

De vierde categorie zijn de variabelen die gaan over de vooropleiding, zoals de code, of een diploma behaald is, de categorie en de tijd tussen de vooropleiding en de eerste aanmelding bij de instelling.

De andere vier categorieën betreffen variabelen die pas na een half jaar of langer beschikbaar komen. Wanneer deze categorieën worden meegenomen bij het maken van een model worden overige variabelen aangepast zodat zij gaan over de opleiding die gevolgd wordt tijdens de momentopname.

De eerste groep variabelen drukt het wisselen van opleiding van een leerling uit zoals het aantal verschillende crebocodes en interne teams. Ten tweede is er de groep variabelen die de behaalde resultaten omschrijven. De derde groep gaat over presentie: het aanwezigheidspercentage, het percentage lesuitval en het percentage geoorloofd afwezig. De laatste groep bevatten de variabelen die gaan over begeleiding: het aantal gesprekken m.b.t. studievoortgang en verzuim, en of de leerling in een traject zit bij het service centrum of het zorgteam.

5. Technieken

In dit hoofdstuk zullen de toegepaste technieken worden besproken. We beginnen met het geven van een korte introductie van de verschillende Machine Learning algoritmen die zijn gebruikt. Vervolgens behandelen we preprocessing stappen die kunnen worden toegepast op de data voordat deze aan een model wordt gegeven. Er wordt afgesloten met de validatie: hoe kan een Machine Learning onderzoek goed worden opgezet en op welke manieren kunnen de prestaties worden beoordeeld en getest.

5.1 Machine Learning Algoritmen

Machine learning algoritmes worden gebruik om patronen in data te ontdekken. In de huidige context worden algoritmes besproken die kunnen worden toegepast in supervised classificatie en/of supervised regressie. Supervised houdt in dat het algoritme leert van voorbeelden waarvan bekend is wat de uitkomst is. Voor m trainingsvoorbeelden zijn de input variabelen of features $x = x_1, \ldots, x_n$ bekend van een leerling (leeftijd, vooropleiding) en daarnaast een output of target variabele y die wordt voorspeld. Bij classificatie wordt er een klasse voorspeld (hier de binaire variabele studiesucces, $y \in \{-1, 1\}$), bij regressie een numerieke waarde (hier de variabele studieduur).

Bij de keuze voor algoritmes is rekening gehouden met een aantal voorkeuren. De algoritmes worden bij voorkeur vaker toegepast, hebben een beschikbare implementatie in R, en kunnen overweg met de data en het soort variabelen dat beschikbaar is. Verder gaat voorkeur uit naar technieken die zijn gebruikt in de toepassingen besproken in Hoofdstuk 3.

In deze sectie zullen we ingaan op de (wiskundige) achtergrond en voor- en nadelen van de gekozen algoritmen. Per techniek is een verwijzing opgenomen naar literatuur waarin de concepten een uitgebreidere introductie krijgen.

5.1.1 Naïve Bayes Classifiers

Naïve Bayes classifiers [Lewis, 1998] zijn een klasse van algoritme dat het theorema van Bayes gebruiken en de sterke naïve aanname doen dat verklarende variabelen onafhankelijk zijn gegeven de klasse.

Naïve Bayes is een probabilistisch model wat conditionele kansen berekend voor elke mogelijk uitkomst op basis van de bekende waarde van de variabelen: $p(y|x_1, \ldots, x_n)$. Vaak is het aantal variabelen groot, of kunnen variabelen veel verschillende waarden kan aannemen. De conditionele 'posterior' kans wordt daarom geschreven als $p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)\ p(y)}{p(\mathbf{x})}$. Hierbij kan de deler als constante worden gezien die niet afhankelijk is van y. Onafhankelijkheid wordt veronderstelt zodat: $p(y|\mathbf{x}) = (\prod_{i=1}^{n} P(x_i|y))\, P(y)$.

Naïve Bayes classifiers ontstaan door een beslisregel in te voeren. De meestgebruikte regel is om de klasse te kiezen met de grootste a posteriori kans, ook wel de 'MAP' decision rule genoemd: $\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}}\ p(y) \prod_{i=1}^{n} p(x_i|y)$.

Als een variabele een waarde niet aanneemt in combinatie met een klasse, dan zal deze waarde nooit hiermee in verband worden gebracht in een voorspelling, de a posteriori kans is 0. Dit is niet altijd wenselijk en wordt opgelost door gebruik te maken van de Laplace schatter: bij elke noemer wordt μ opgeteld en bij elke deler het aantal klassen maal μ.

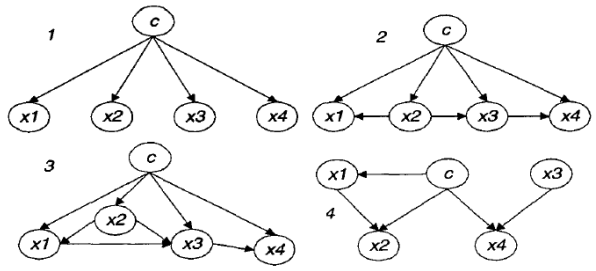
Voordelen van Naïve Bayes classifiers zijn dat ze robuust zijn met betrekking tot irrelevante variabelen, ze eenvoudig te implementeren zijn, ze kunnen omgaan met missende waarden, en het trainen en het classificeren met een getraind model efficiënt is. Een nadeel is de sterke onafhankelijkheidsaanname. Verder kunnen ze enkel direct omgaan met categorische variabelen. Om $p(x \mid y)$ te bepalen voor numerieke variabelen moet in de preprocessing worden gediscretiseerd of moet een onderliggende distributiefunctie worden geschat. In de praktijk blijkt dat Naïve Bayes classifiers vaak minder presteren dan andere algoritmen [Caruana and Niculescu-Mizil, 2006]. Wel worden ze veel toegepast bij en in de categorisatie van teksten.

Met de onafhankelijkheidsaanname kan worden omgegaan op twee manieren: ten eerste kan feature selectie worden gebruikt om enkel onafhankelijke variabelen te kiezen, de tweede mogelijkheid is om de afhankelijkheid aanname te relaxeren; Cheng and Greiner [1999] vergelijken een aantal mogelijkheden. Een eerste manier is om variabelen te groeperen (zonder overlapping) en te veronderstellen dat de gevormde groepen onafhankelijk zijn. Een andere manier is om gewichten aan variabelen toe te kennen zodat de negatieve conditionele log likelihood wordt geminimaliseerd [Zaidi et al., 2013]. Een andere manier is om conditionele afhankelijkheid toe te laten in een bepaalde mate en de structuur hiervan te omschrijven in een bayesiaans netwerk.

Bayesian Network Classifiers

Het voordeel van Bayesian Network (BN) classifiers [Cheng and Greiner, 1999] ten opzichte van Naïve Bayes is dat zij vaak beter presteren terwijl ze in rekenkundig opzicht niet veel duurder zijn. Veel van de voordelen van NB classifiers komen terug in de BN classifiers. Zo kunnen BN classifiers ook omgaan met missende waarden door de marginale waarschijnlijkheid te gebruiken (de delen van de voorbeelden die niet missen). Verder zijn ook BN classifiers robuust met betrekking tot irrelevante variabelen.

Voor BN classifiers wordt een netwerk met afhankelijkheidsrelaties opgesteld. Dit wordt gedaan aan de hand van de minimum description length (MDL) waarbij simpelere beschrijvingen (netwerken) voorkeur krijgen. De toegevoegde waarde wordt bepaald met 'mutual information tests': de toename in informatie wordt beschouwd conditioneel op de klasse wanneer variabelen samen worden gezien ten opzichte van wanneer niet.



Figuur 5.1: Mogelijke BN structuren. 1: Naïve Bayes, 2: Tree, 3: Network, 4: General. Bron: [Cheng and Greiner, 1999]

In Figuur 5.1 worden verschillende structuren van een BN weergeveb. De eerste structuur geeft een Naïve Bayes classifier weer, hierbij ligt de classificatie node boven die van de variabelen. De tweede graaf is een Tree Augmented Naïve Bayes (TAN) classifier, waarbij elke variabele ook afhankelijk kan zijn van maximaal één andere variabele. De derde is een Bayesian Network Augmented Naïve Bayes (BAN) classifier, waarbij een variabele afhankelijk kan zijn van meerdere andere variabelen. De laatste structuur is van een General Bayesian Network (GBN) classifier, hierin wordt de classificatie node niet gezien als speciaal. Op de laatste structuur na is de node die correspondeert met de klasse een ouder van de anderen, dit zorgt ervoor dat alle variabelen worden meegenomen bij het classificeren. De schatting die gemaakt wordt lijkt nog erg sterk op die van een Naïve Bayes classifier, enkel de extra afhankelijkheid moet in ogenschouw worden genomen: $P(y, x) = P(y) \prod_{i=1}^{n} P(x_i \mid y, \pi(x_i))$ waarbij $\pi(x_i)$ de node boven x_i is.

Weighted Naïve Bayes

Een andere manier om met de onafhankelijkheidsaanname is met Weighted Naïve Bayes [Zaidi et al., 2013]. Hierbij wordt de schatting $p(y|\mathbf{x}) = P(y)\prod_{i=1}^{n} P(x_i|y)$ vervangen door een gewicht aan elke variabele toe te wijzen, dit kan in het meest algemene geval op basis van de waarde van de variabele: $P(y)\prod_{i=1}^{n} P(x_i|y)^{w_{i,x_i}}$, of op basis van de variabele: $P(y)\prod_{i=1}^{n} P(x_i|y)^{w_i}$.

De voorgestelde methode van Zaidi et al. [2013], WANBIA$^{\text{CLL}}$, kent een gewicht toe door de conditionele log likelihood te minimaliseren. Het gewicht dat wordt toegekend wordt dus niet bepaald op basis van de voorspellende waarde van een variabele, maar op basis van de algemene prestaties van de classifier. De redenatie hierachter is dat het doel zou moeten zijn om het effect van het schenden van de onafhankelijkheidsaanname te verminderen.

Het toekennen van een gewicht aan variabelen zou beter werken dan het selecteren van variabelen [Zaidi et al., 2013]. De redenatie is dat dezelfde resultaten kunnen worden verkregen door variabelen die niet worden gebruikt op 0 te zetten; verder kunnen andere gewichten gebruikt worden om classifiers te definiëren te niet kunnen worden uitgedrukt wanneer er gebruik wordt gemaakt van een selectie methode.

5.1.2 Support Vector Machines

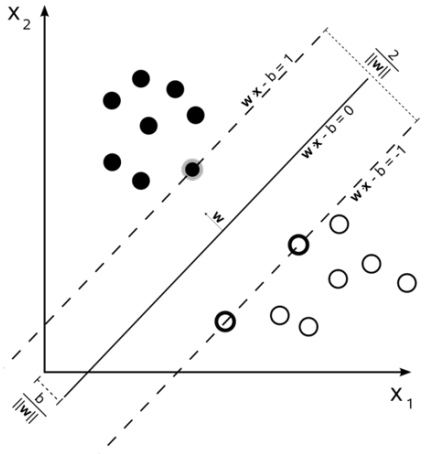
Support Vector Machines [Cristianini and Shawe-Taylor, 2000] bouwen een model dat verschillende klassen als punten in de ruimte maximaal scheidt door een 'maxed-margin hyperplane' (scheidingsvlak) in een (mogelijk) hoge-dimensionale ruimte. De naam komt van de support vectors, de datapunten die het dichtste bij de scheidingsvlak liggen en het model definiëren. Nieuwe data wordt geclassificeerd door te kijken waar zij zich ten opzichte van het scheidingsvlak bevinden.

Een SVM kan ook gebruikt worden om een niet lineair scheidingsvlak op te stellen, hiervoor maken SVM's gebruik van zogenaamde kernels. Een punt $x^{(i)}$ in de oorspronkelijke ruimte wordt door de transformatie $\phi(x^{(i)})$ naar een andere (hoger dimensionale) ruimte geprojecteerd. Dit maakt een niet lineaire scheiding in de oorspronkelijke ruimte mogelijk. Het opstellen van het scheidingsvlak wordt vervolgens gedaan door het Lagrange duale probleem op te lossen. In plaats van eerst een projectie te maken en in deze (hogere) dimensie de afstand tussen de punten te bepalen, worden kernels gebruikt om de afstand tussen twee punten impliciet te berekenen.

Een SVM probeert een scheidingsvlak te bepalen dat de klassen scheidt door een vergelijking in de vorm $w \cdot x - b = 0$ op te stellen. In het duale probleem wordt w geschreven als lineaire combinatie van de datapunten. Dit is gevisualiseerd in figuur 5.2. Het maxed-margin scheidingsvlak wordt gevonden door het optimaliseringsprobleem op te lossen:
$\min \frac{1}{2}||w||^2 \quad \text{s.t.} \quad y_i(w \cdot x_i + b) \geq 1, i = 1, \ldots m.$
Het optimaliseringsprobleem van het duale probleem is (waar $w = \sum_{i=1}^{m} \alpha_i y_i x_i$):
$\max \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j (x_i \cdot x_j) \quad \text{s.t.} \quad 0 \leq \alpha_i \leq C \text{ en } \sum_{i=1}^{m} \alpha_i y_i = 0.$
In het duale probleem hebben enkel de support vectors een $\alpha_i > 0$. Ook wordt in deze formulering de handigheid van kernels duidelijk: $(x_i \cdot x_j)$ vervangen door $(\phi(x_i) \cdot \phi(x_j))$, wat berekend wordt met een kernelfunctie $K(x_1, x_2) = \phi(x_1) \cdot \phi(x_2)$. Vaak gebruikte kernels zijn de volgenden: $K(x_1, x_2) = x_1 \cdot x_2$ (lineaire SVM), $K(x_1, x_2) = (x_1 \cdot x_2 + 1)^p$ (polynomische SVM met graad p) en $K(x_1, x_2) = e^{-||x-y||^2/2\sigma^2}$ (RBF SVM).

Het is vaak niet mogelijk om de klassen volledig te scheiden. Als dit het geval is wordt het optimialiseringsprobleem aangepast door een boete voor misclassificatie in te voeren. Hiermee wordt het probleem omgevormd tot het maken van een afweging tussen het opdelen van klassen en het maximaliseren van de afstand. Van origine wordt SVM gebruikt voor (binaire) classificatie, maar het kan ook worden gebruikt voor regressie. Dit wordt gedaan in ϵ-SV regressie door een marge in toe voeren (genoteerd met ϵ) en vervolgens punten die meer dan deze ϵ afwijken als verkeerd geclassificeerd te beschouwen.

SVMs geven in de praktijk vaak goede resultaten. Verder stelt de kernel de gebruiker in staat om niet-lineaire relaties te beschrijven en kan met regularisatie overfitting worden tegengegaan. De kernel en regularisatie parameter kunnen ook als nadeel worden gezien: kernel models kunnen



Figuur 5.2: Maximale scheiding met Support Vector Machines. Bron: Wikibooks [1]

[1] https://en.wikibooks.org/wiki/Support_Vector_Machines

vrij vatbaar zijn om te overfitten en het probleem wordt verschoven naar het optimaliseren van parameters. Een ander nadeel is dat het opgestelde model niet te interpreteren is: SVMs vallen in de categorie van black-box modellen. Verder kan de rekentijd oplopen voor grotere datasets.

5.1.3 Decision Trees

Decision Trees [Witten and Frank, 2005] kunnen zowel worden toegepast in classificatie, en zijn dan bekend onder de naam classificatie bomen, of voor regressie, en staan dan bekend als regressie bomen. Decision Trees stellen geen model op voor alle data, maar delen deze eerst op in kleinere partities. De bladeren representeren de output en de vertakkingen samenstellingen van variabelen die tot deze output leiden.

Er zijn verschillende algoritmen die kunnen worden gebruikt om deze bomen op te stellen. Twee populaire implementaties, CART en C4.5, bouwen een boom op een 'top-down' manier door recursief te partitioneren. Het bouwen van de boom begint met alle data in de hoofdnode. Hier wordt een splitsing gemaakt op de variabele die de data zo goed mogelijk opdeelt. Welke variabele hierin het meest effectief is wordt bepaald door de een 'split' function. In het geval van classificatiebomen wordt hiervoor vaak de 'information value' of entropy gebruikt en voor regressie bomen wordt vaak gebruik gemaakt van een functie die de kwadraatsom tussen groepen minimaliseert, en is gebaseerd op variantieanalyse (ANOVA) modellen die in de statistiek worden gebruikt om verschillen tussen groepsgemiddelden te analyseren. Deze splitsing kan gedaan worden door een variabele te vergelijken met een constante, maar ook door variabelen met elkaar te vergelijken, of een functie over verschillende variabelen te nemen.

Eén mogelijkheid is om dit proces van het bouwen van een boom door te laten gaan totdat het niet meer mogelijk is de data verder op te delen. Ook is het mogelijk om te stoppen wanneer het aantal voorbeelden in elke onderste node onder een vooraf ingestelde drempelwaarde zit, of wanneer een maximale diepte is bereikt. Wanneer er een boom is opgebouwd wordt deze vaak nog gesnoeid om de complexiteit te verminderen en overfitting tegen te gaan. Dit kan worden gedaan door de onderste vertakking of een sub-boom te vervangen door de meest voorkomende klasse gebaseerd op de toename in gemaakte fout en de afname in complexiteit.

De aanpak van CART en C4.5 met recursieve partitionering kent twee problemen: ze zijn geneigd te overfitten en er is een bias om variabelen met veel mogelijke splitsingen te kiezen [Hothorn et al., 2006]. Hothorn et al. [2006] stellen een algoritme voor waarin variabelen worden geselecteerd op basis van permutatie tests. De klassen worden opnieuw verdeeld onder de trainingsvariabelen en vervolgens wordt de splitsing gemaakt waar de test statistiek onder de nul hypothese van onafhankelijkheid tussen de variabelen en de response wordt verworpen met de meeste zekerheid.

Het splitsen gaat door zolang de p-waarde hoger blijft dan een vooraf bepaald niveau α, die de grootte van de boom bepaalt. Bomen opgesteld volgens dit algoritme staan beter bekend als 'conditional inference trees'.

5.1.4 (Bayes) Generalized Linear Model

Een generalized linear model (GLM) [Starkweather, 2011] is een generalisatie van het lineaire regressie model dat veel toegepast wordt in de statistiek. In vergelijking met linaire modellen hoeven bij GLMs de onafhankelijke variabelen niet de normale verdeling te volgen. Verder hoeft de afhankelijke niet lineair te variëren met de onafhankelijke variabelen: deze worden via een link functie met elkaar in verband gebracht. GLMs zijn omvatten ook linaire regressie, poisson regressie en logistische regressie.

Om de notatie te vergemakkelijken houden we hier $y \in \{0, 1\}$ aan in plaats van $y \in \{-1, 1\}$. Een Generalized Linear Model bestaat uit 3 componenten:

1. Stochastisch component: Y_i. Een random component met onafhankelijke stochastische variabelen, dit is de kansverdeling van de afhankelijke variabele.

2. Systematisch Component: $\nu_i = x_i^T \beta$. Het systematische component duidt de rol van de variabelen aan. De ν_i's zijn nog steeds lineair, vandaar de naam generalized linear model.

3. Link functie: $\nu_i = g(\mu_i)$. Het stochastische component en het systematische component worden verbonden door de functie g, deze brengen een verband aan tussen de lineaire voorspeller ν_i en de verwachte waarde van Y_i.

In het geval van classificatie ligt de verwachting van Y_i in het interval $(0, 1)$. Twee mogelijke link functies hierbij zijn de logit functie: $\log(\frac{\mu}{(1-\mu)})$, en de probit functie: $\Phi^{-1}(\mu)$ met Φ de standaard normale kansverdeling. Een GLM voor binaire classificatie kan vervolgens worden gespecificeerd:

1. Stochastisch component: $n_i Y_i \sim \text{Bin}(n_i, \mu_i)$.

2. Systematisch Component: $\nu_i = x_i^T \beta$.

3. Link functie: $\nu_i = g(\mu_i) = \log(\frac{\mu}{(1-\mu)})$ of $\nu_i = g(\mu_i) = \Phi^{-1}(\mu)$.

Het doel is gegeven de voorbeelden en de n verklarende variabelen, een $n \times 1$ coëfficiëntenvector θ te bepalen. De posterior $p(\theta \mid Y)$ wordt afgeleid van $p(Y \mid \theta)$ waarvoor de aannemelijkheidsfunctie wordt gemaximaliseerd: $L(\theta \mid Y) = \prod_{i=1}^{n} p(Y_i \mid \theta)$.

Een Bayes GLM verschilt in opzet in dat het de Bayesiaanse interpretatie gebruikt in plaats van de frequentistische. Waar bij de frequentistische interpretatie de model parameters vaststaan en de data als het ware volgt uit een steekproef, wordt bij de Bayesiaanse interpretatie aangenomen dat de model parameters willekeurig zijn en wordt de vraag gesteld: wat is de kans op een hypothese (of parameter) gegeven de data?

De Bayesiaanse methode maakt gebruik van voorafgaande informatie gecombineerd met een likelihood functie die gebaseerd is op de data. Hieruit volgt een posterior verdeling van de coëfficiënt waarden. Uit deze verdeling wordt door simulatie een empirische verdeling gemaakt voor de werkelijke parameter. Vervolgens wordt met beschrijvende statistiek de gesimuleerde empirische verdeling beschreven, zo is de meest aannemelijke schatter de modus en kan een 'credible' interval worden aangeduid. Bij een Bayes GLM kan kennis worden meegenomen in de prior, bijvoorbeeld door een gemiddelde of schaal van de coëfficiënten van te voren vast te stellen.

5.1.5 Ensemble technieken

Ensemble modellen maken een voorspelling door meerdere modellen te combineren. Hierdoor kan een ensemble model op zichzelf als supervised algoritme kunnen zien met de uitkomsten van de onderliggende modellen als variabelen.

Er zijn een aantal aanpakken mogelijk. Bij 'boostrap aggregating', beter bekend als 'bagging', wordt de uitkomst bepaald door door een ongewogen som te nemen van losse modellen. Diversiteit tussen enkele modellen wordt bevorderd door gebruik te maken van een random subsets. Bij 'boosting' wordt een model stapsgewijs opgesteld door nieuwe modellen toe te voegen. Hierbij worden de simpele modellen vaak getraind op de fout die wordt gemaakt door het tot op dat punt opgestelde model. Als laatste is er 'stacking', hierbij wordt een meta-algoritme gebruikt dat de voorspellingen van andere modellen combineert. Stacking lijkt erg op boosting, het verschil ligt in dat er een 'meta-level' model wordt gebruikt om de inputs te combineren in plaats van een gegeven empirische functie.

5.1.6 Random Forest

Random Forests [Breiman, 2001] is een ensemble methode dat gebruik maakt van bagging. Bij het Random Forest algoritme worden meerdere Decision Trees opgesteld en wordt als output de gemiddelde output gegeven. Hiermee wordt gecorrigeerd voor de neiging van Decision Trees om te overfitten. Waar Decision Trees een lage bias en een hoge variantie hebben bieden Randoms Forests een kleine toename in bias in ruil voor een afname in de variantie.

Het Random Forest algoritme verschilt op twee manieren in het opbouwen van enkele decision bomen. Ten eerste, bekend onder de term 'tree bagging', wordt niet alle data gebruikt maar wordt een steekproef met vervanging genomen. Ten tweede, bekend onder de term 'feature bagging', wordt de beslissing over op welke variabelen te splitsen gemaakt op een willekeurig deel van de variabelen.

Random Forests kunnen ze omgaan variabelen van verschillende soorten en hebben embedded variabele selectie (Zie Sectie 5.2). Verder kunnen Random Forests een indicatie geven van het belang van variabelen, waarbij ook multivariabele relaties meegenomen worden. Doordat er een steekproef wordt gebruikt om enkele bomen op te stellen is het mogelijk een schatting te krijgen van de prestaties op het 'out-of-bag' deel. Er is reeds aangetoond dat deze schatting zonder bias is [Strobl et al., 2007], waardoor er in principe geen test set nodig zou zijn.

5.1.7 AdaBoost

Een andere ensemble methode is AdaBoost [Freund and Schapire, 1999], een algoritme wat de output van 'weak learners' combineert. Als output wordt een lineaire combinatie van de output van simpele modellen gegeven. AdaBoost staat voor Adaptive Boosting en is adaptief in de zin dat de simpele modellen worden gewogen op basis van voorgaand verkeerd geclassificeerde trainingsvoorbeelden.

Boosting zorgt voor een complexer model, vandaar dat wordt aangeraden om de weak learners simpel te houden. De weak learners kunnen worden opgesteld met meerdere technieken. Veel gebruikt wordt een Decision Tree met een enkele split, een 'decision stump'. De voorkeur van weak learners gaat uit naar technieken waarbij een gewicht kan worden toegewezen aan de trainingsvoorbeelden, maar het is ook mogelijk om de trainingsvoorbeelden te kiezen aan de hand van gewichten.

In grote lijnen werkt het algoritme als volgt: Elk simpel model produceert een output $h(x_i) \in \{0, 1\}$ voor elk voorbeeld in de training set. In elke iteratie $1 \leq m \leq M$ wordt een gewicht toegekend aan elk voorbeeld in de training set die afhankelijk is van de huidige fout. Adaboost stelt iteratief een model samen waarbij in elke iteratie m een simpel model wordt toegevoegd ($C_m(x_i) = C_{m-1}(x_i) + \alpha_m k_m(x_i)$) die de exponentiële loss minimaliseert. Het gewicht wat wordt toegekend is afhankelijk van de gemaakte fout in de iteratie (ϵ_t): hoe lager de error rate, hoe hoger het gewicht ($\alpha_m = \frac{1}{2}\ln\frac{1-\epsilon_m}{\epsilon_m}$).

Tijdens het iteratieve proces wordt informatie over hoe moeilijk het is om een trainingsvoorbeeld goed te classificeren vergaard. Dit stelt AdaBoost ertoe in staat om outliers op te merken (voorbeelden die ambigu zijn of verkeerd gelabeld). Wanneer er veel outliers in een dataset zitten en er teveel nadruk op wordt gelegd, kan dit echter wel afbreuk doen aan de prestaties [Freund and Schapire, 1999]. Een ander voordeel van Adaboost is dat het mogelijk is om complexe, niet-lineaire

relaties te modelleren. Nadelen zijn dat het veel geheugen kost om een model op te slaan, en wanneer een nieuw voorbeeld wordt geëvalueerd hiervoor alle weak learners moeten worden geëvalueerd.

5.1.8 Gradient Boosting Machines

Net als andere boosting methodes combineert gradient boosting weak learners in een iteratief proces [Friedman, 2001]. Bij gradient boosting wordt in plaats van het aanpassen van gewichten van verkeerd geclassificeerde voorbeelden gebruik gemaakt van gradient descent om een kosten functie te minimaliseren.

Ook hier houden we weer $y \in \{0, 1\}$ aan in plaats van $y \in \{-1, 1\}$ om de notatie te vergemakkelijken. In elke iteratie $1 \leq m \leq M$ wordt het tot dan toe opgestelde model C_{m-1} uitgebreid door een nieuw model h te te voegen. De manier waarop h wordt geselecteerd verschilt hier in vergelijking met het Adaboost algoritme: er wordt gebruik gemaakt van gradient descent om een error functie te minimaliseren zoals de squared-error loss: $\frac{1}{2}(y - F_m(x))^2$. Het uiteindelijke model is een product van de gemaakte keuze van weak learners (waarmee wordt het model gebouwd?) en de kostenfunctie (hoe wordt het model gebouwd?).

Het is mogelijk om verschillende error functies te gebruiken. Voor de binaire classificatie zijn er twee functies die vaak worden toegepast: ten eerste de exponentiële loss functie, $L(y, F(x)) = exp(-yF(x))$, met deze kostenfunctie komt het model overeen met een Adaboost model. Ten tweede wordt de Bernoulli loss, $L(y, F(x)) = \log(1 + e^{2yF(x)})$, dit staat gelijk aan het minimaliseren van de negatieve log-likelihood. De Bernoulli loss legt minder nadruk op verkeerd geclassificeerde punten, en is daarom minder vatbaar voor outliers.

Waar Adaboost alleen toe te passen is voor classificatie, zijn GBMs ook te gebruiken voor regressie. Twee mogelijkheden hierbij zijn least-squares regressie (met de loss functie $L(y, F(x)) = \frac{(y-F)^2}{2}$) en least-absolute-deviation regressie (met de loss functie $L(y, F(x)) =\mid y - F \mid$). Hierbij komt least-squares regressie overeen met het iteratief maken van een model op de residuals en least-absolute-deviation regressie als het maken van een model met least-squares op het teken van de residuals.

5.1.9 Belang van variabelen

Voor een aantal technieken is het mogelijk om het belang van variabelen af te leiden uit de opgestelde modellen. Zo kan aan de splitsingen van een Decision Tree direct worden afgeleid welke variabelen van belang zijn. Echter kunnen variabelen die wel van invloed zijn maar niet worden gebruikt op deze manier worden 'verborgen'. Er zijn ook meer geavanceerde strategieën [Strobl et al., 2007]: voor CART kunnen de ('surrogate') variabelen die worden gebruikt bij missende waarden een indicatie geven. Voor Decision Trees kan gekeken worden naar de verbetering van de onzuiverheid in de onderste node en kan gekeken worden naar de (afname in) prestatie wanneer er ruis in variabelen wordt geïntroduceerd.

Het opstellen van een boom in een Random Forest gebeurd met een deel van de data, door het andere ('out-of-bag') deel te permuteren kan het belang van variabelen worden bepaald in een enkele boom en het belang voor het forest wordt bepaald door deze te combineren. De uitkomst hierbij is niet altijd geheel betrouwbaar: variabelen met veel categorieën kunnen de voorkeur krijgen [Strobl et al., 2007]. Wannneer Decision Trees worden opgesteld op basis van het conditionele inferentie framework, verdwijnt deze bias echter. Ook voor het Gradient Boosting Machine algoritme is het mogelijk om het belang te bepalen op basis van permutatie testen. Verder kan het effect worden bepaald dat een splitsing (op een variabele) heeft op de gemaakte fout in de onderste nodes.

5.2 Preprocessing technieken

Om een goede inschatting te krijgen van de prestatie van een model, zouden transformaties die afhankelijk zijn van andere voorbeelden binnen de validatiestap moeten gebeuren. Transformaties worden eerst toegepast op een train set, opgeslagen, en vervolgens onafhankelijk uitgevoerd op de test set.

Prepareren variabelen

Variabelen zijn van invloed op de prestaties van machine learning algoritmen, ook is het aantal variabelen van invloed op de kosten voor het trainen van een model en het genereren van output. Het is wenselijk om een beschrijving te geven van de data die simpel maar omvattend is. Guyon and Elisseeff [2003] geven een introductie en presenteren een stappenplan voor het selecteren van variabelen. Hier worden de volgende stappen benoemd: het toepassen van domeinkennis om betere variabelen te maken; het normaliseren, transformeren en combineren van variabelen; het toepassen van een ranking methode en het identificeren van outliers. De eerste stap is het toepassen van domeinkennis om betere variabelen te vormen, een proces wat beter bekend staat als 'feature engineering'. De stappen die hierin al genomen zijn, zijn toegelicht in Sectie 4.2.2.

Verder bestaan er generieke methoden om variabelen om te vormen. Clusteren kan worden gebruikt om een groep soortgelijke variabelen te vervangen door een enkele variabele. Lineaire transformaties, zoals zijn principal component analysis en linear discriminant analysis, kunnen worden gebruikt om variabelen uit te drukken in lineaire combinaties van (minder) nieuwe variabelen die variantie verklaren. Ook kunnen simpele functies worden gebruikt om variabelen te combineren. Denk hierbij aan het nemen van conjuncties en disjuncties bij binaire variabelen en het gebruiken van minima, maxima en gemiddelden bij numerieke variabelen.

Enkele machine learning algoritmes zijn enkel te gebruiken met numerieke variabelen als invoer, dit bijvoorbeeld omdat het nodig is een in-product of de afstand tussen twee voorbeelden te berekenen. Er bestaan verschillende technieken om categorische variabelen om te vormen. Een mogelijkheid is om een variabele met m categorieën te coderen als m binaire variabelen. Bij een groot aantal categorieën heeft dit een toename in dimensionaliteit tot gevolg, wat kan leiden tot slechtere prestaties. Een andere manier is om afhankelijkheid tussen categorische variabelen uit te drukken met behulp van een Bayesian Network als numerieke variabelen [Lee and Kim, 2010].

Het omgekeerde, het omvormen van numerieke variabelen naar categorische variabelen, heet discretiseren. Discretiseren kan gedaan worden aan de hand van clustering en door een indeling te maken in intervallen, ook hierin zijn weer veel verschillende mogelijkheden, Mitov et al. [2009] geven een overzicht.

Selectie van variabelen

Methodes voor het selecteren van variabelen kunnen worden onderverdeeld in filter, embedded en wrapper methodes [John et al., 1994]. Deze verschillen in de manier waarop de selectie is gecombineerd met het opstellen van een model. Bij filters worden de variabelen gekozen voordat zij in een algoritme gebruikt worden en wordt het algoritme niet direct in beschouwing genomen. Bij wrapper methodes wordt het belang van (subsets van) variabelen bepaald voor een opgesteld model en wordt het model gezien als zwarte doos. Bij embedded methodes maakt de selectie deel uit van het algoritme. Voordelen van filters is dat zij rekenkundig snel zijn en kunnen worden toegepast in de preprocessing fase. Zij geven een generieke selectie van variabelen die niet direct afgestemd is op een algoritme. Wrapper methodes worden vaak gezien als rekenkundig duur, vooral wanneer het aantal variabelen groot is. Embedded methodes zijn in rekenkundig opzicht vaak niet duur en daarnaast past de selectie bij het model.

Imputeren van missende waarden

Missende waarden kunnen ontstaan om verschillende redenen, zoals bij het invoeren of bij het samenvoegen van datasets. Bij het imputeren moet rekening worden gehouden met of de vervangende

waarde wel mogelijk is en dat relaties tussen variabelen standhouden.

Een eenvoudige manier is om missende waarden te vervangen is door het gemiddelde in het geval van numerieke variabelen of de meest voorkomende waarde voor categorische variabelen te gebruiken. Er zijn verschillende meer geavanceerde methoden zoals k Nearest Neighbour (kNN) en mice.

Mice, multivariate imputation by chained equations [Buuren and Groothuis-Oudshoorn, 2011], is een methode die kan worden toegepast wanneer meer dan één variabele een missende waarde bevat. Hierbij wordt gebruik gemaakt van zogenaamde 'chained equations'; missende waarden worden meerdere keren vervangen en de resultaten worden gecombineerd. Met mice is het mogelijk om relaties in stand te houden die komen door transformaties, combinaties of coderingen van variabelen. Daarnaast kent mice veel keuzevrijheid: bij de specificatie van een 'imputatie-model' zijn er zeven keuzes om te maken zoals de vorm van het model en de volgorde van het imputeren.

Bij Nearest Neigbour worden de k meest overeenkomende voorbeelden geselecteerd door het Nearest Neighbour algoritme toe te passen en vervolgens missende waarde vervangen door ofwel het gemiddelde (voor numerieke variabelen) ofwel de meest voorkomende waarde (voor categorische variabelen) onder deze voorbeelden. Het bepalen van de meest overeenkomende voorbeelden kan gedaan worden met verschillende technieken en is afhankelijk van de data en het soort variabelen. Voor een mix van categorische, binaire en numerieke variabelen kan de Gower gelijkheidscoëfficiënt worden gebruikt. Voor numerieke variabelen wordt er een gelijkheidsscore toegewezen tussen voorbeelden i en j: $s_{ijk} = \frac{1-|x_{ik}-x_{jk}|}{r_k}$, waar r_k het bereik is van variabele k. Voor categorische variabelen wordt er een score en gewicht van 1 toegewezen wanneer de waarden overeenkomen, voor binaire variabelen wordt een gelijkheidsscore en gewicht bepaald afhankelijk van of de klassen hetzelfde beide positief zijn en in welke combinaties de klassen voorkomen.

5.3 Validatie

De validatie is een belangrijk onderdeel van machine learning. Validatie wordt onder andere gebruikt om te testen of er sprake is van overfitting: bij overfitting leert het model patronen die niet generaliseren voor nieuwe gevallen.

Opdeling data

De beperkte beschikbare data moet worden opgedeeld om de prestatie te meten. Vaak worden hiervoor drie sets gemaakt: een train set, om de modellen te trainen, een validatie set om de parameters te kiezen en een test set om de prestaties te vergelijken.

Bij cross validatie worden meerdere sets gemaakt die worden gecombineerd tot train en validatie sets. Bij 'leave-p-out' cross validatie worden p voorbeelden overgehouden voor de validatie en wordt de rest gebruikt voor het trainen. Bij k-fold cross validatie wordt de data opgedeeld in k delen, hiervan worden vervolgens k-1 delen gebruikt bij het trainen en de overigen bij het valideren. Bij Monte Carlo cross validatie worden voorbeelden willekeurig verdeeld in een train en validatie set en is het mogelijk dat voorbeelden nooit of meerdere keren worden gekozen. Verder bij een techniek genaamd 'stratified' sampling wordt rekening gehouden met de afhankelijke variabele. Ook kunnen er meerdere rondes van cross validatie worden uitgevoerd om iets te zeggen over de variabiliteit (een voorbeeld hiervan 5x2 cv, en wordt aanbevolen door Dietterich [1998]).

5.3.1 Prestatiemaat

Classificatie

Voor classificatie kunnen prestatiematen worden onderverdeeld in twee categorieën: deterministisch en probabilistisch [Japkowicz and Shah, 2011]. Deterministische prestatiematen beschouwen enkel het voorspelde label en zijn gebaseerd op de confusion matrix (Figuur 5.3) of kunnen hiervan worden afgeleid. Probabilistische prestatiematen nemen ook de zekerheid van een voorspelling mee.

		Voorspelde class	
		Positief	Negatief
Werkelijke class	Positief (P)	True Positief (TP)	False Negatief (FN)
	Negatief (N)	False Positief (FP)	True Negatief (TN)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{P}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figuur 5.3: Confusion matrix & prestatiematen

De accuracy is gedefinieerd als het percentage dat goed geclassificeerd is. De recall is het percentage van de positieve voorbeelden dat als zodanig geclassificeerd is, de precision is het percentage van het aantal positieve voorbeelden dat als positief geclassificeerd is. De F-measure combineert de precision en recall door het harmonisch gemiddelde te nemen. Elke prestatiemaat benadrukt een ander aspect; een betere prestatie volgens de ene maat kan overeen komen met een mindere prestaties volgens een andere maat. In de afweging speelt onder andere mee of er sprake is van een onbalans in de klassen, waarbij het mogelijk interessant is om enkel de positieve voorbeelden te bekijken.

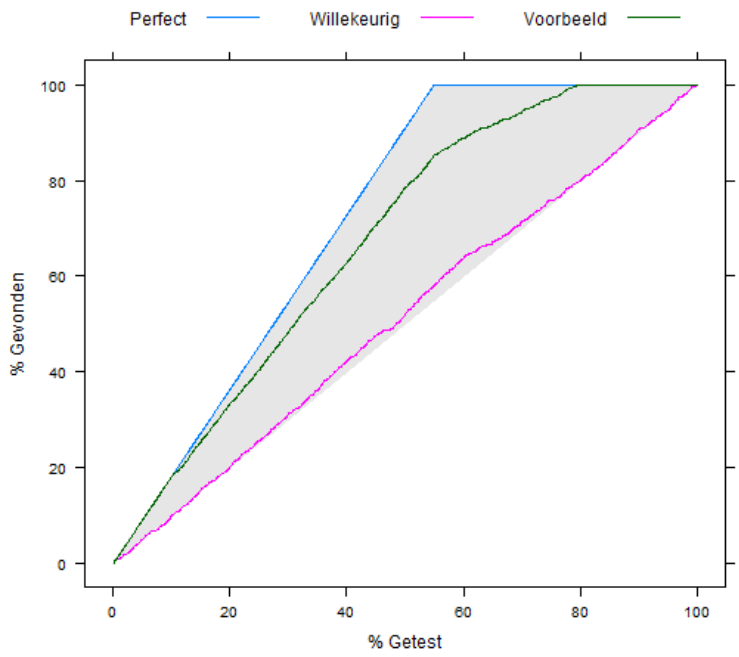
Voor classificatie is er een klasse algoritmes die naast enkel de meest waarschijnlijke klasse ook een kans produceren. Deze probabilistische classifiers geven hiermee een mate van zekerheid over de voorspelling die in de evaluatie kan worden meegenomen. Veel probabilistische prestatiematen zijn grafisch en laten zien hoe het model presteert over een bereik. Voorbeelden zijn de Precision en Recall curve, de ROC-curve en de lift grafiek. De Precision en Recall curve plot de afweging tussen het aantal true positieven ten opzichte van het aantal false negatieven. De lift grafiek plot het aantal als goed geclassificeerde uit de klasse van interesse (TP) tegen de grootte van de dataset die nodig was om dit percentage te behalen. De ROC curve lijkt erg op de lift grafiek, maar telt het aantal negatieve voorbeelden dat nodig is.

In Figuur 5.4 is een lift grafiek geplot. Hierin wordt het percentage deelnemers uit de groep van interesse geclassificeerd als zodanig, hier de groep correct geclassificeerd als ‘geen studiesucces’, uitgezet tegenover van het percentage van de trainingsvoorbeelden dat gebruikt wordt om dit percentage te vinden. De blauwe lijn geeft perfecte classificatie aan, de roze lijn geeft aan wat er gebeurd bij willekeurige classificatie: de fractie van de goed geclassificeerde instanties is in dat geval gelijk aan de fractie dat wordt gebruikt om dit percentage te vinden. Hoe hoger de lijn ligt bij een bepaald percentage, hoe beter de bijbehorende techniek presteert.

Regressie

Bij regressie ligt de voornaamste afweging in de mate waarin een grotere fout een hogere weging krijgt. Mogelijke maten zijn de mean squared error (MSE), de root mean squared error (RMSE) en de mean absolute error (MAE).

Deze maatstaven meten op verschillende manieren de gemiddelde afwijking van de voorspelde waarde van de werkelijke waarde. De MSE neemt het gemiddelde van het kwadraat van deze verschillen. De RMSE neemt hierover een wortel en heeft dezelfde eenheid als de data. De MAE neemt het gemiddelde van de absolute fouten en heeft ook dezelfde eenheid als de data en is van dezelfde orde van grote is als de RMSE. Een andere maatstaf is de determinatiecoëfficiënt R^2. De R^2 meet het deel van de variabiliteit dat wordt verklaard door een model: het is de proportionele verbetering van de voorspelling, ten opzichte van een model wat voor elke voorspelling de mean gebruikt. Deze maatstaven kunnen als volgt worden berekend, waarbij y de werkelijke waarde is, y' de voorspelde waarde en $\bar{y}$ de gemiddelde waarde van y.



Figuur 5.4: Voorbeeld lift grafiek

$$\text{MSE} = \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{n} \qquad \text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i')^2}{n}} \qquad \text{MAE} = \frac{\sum_{i=1}^{n} \mid y_i - y_i' \mid}{n}$$

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \qquad \text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad \text{SSR} = \sum_{i=1}^{n}(y_i' - \bar{y})^2 \qquad \text{SSE} = \sum_{i=1}^{n}(y_i - y_i')^2$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{n \times \text{MSE}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

De prestatiemaat kan worden gekozen afhankelijk van de loss functie die door een algoritme geoptimaliseerd wordt: er zijn loss functies die de fout minimaliseren die corresponderen met een van de benoemde maatstaven. Ook kan worden een keuze worden gemaakt op basis van de weging van de gemaakte fout: als een afwijking van 10 meer dan twee keer zo slecht is dan 5, dan ligt het voor de hand om de RMSE te kiezen over de MAE.

5.3.2 Statistische toetsen

Er zijn verschillende statistische toetsen die kunnen worden toegepast om te testen of een algoritme significant beter presteert dan een ander algoritme. Een aantal mogelijkheden worden besproken door Salzberg [1997].

Classificatie

Om te bepalen of er een significant verschil is tussen de uitkomsten van de algoritmen is het mogelijk om Cochran's Q test toe te passen. Deze test wordt toegepast om na te gaan of de proportie van de uitkomsten hetzelfde is tussen groepen en kan worden beschouwt als ANOVA voor binaire uitkomsten. Het is een zogenaamde 'omnibus' test met de nulhypothese dat de proporties van de steekproeven gelijk zijn en de alternatieve hypothese dat er ten minste één anders is. Met X_{ij} wordt de uitkomst (goed geclassificeerd of niet) op testvoorbeeld i door algoritme j uitgedrukt

en met $_\bullet$ de som over een kolom/rij. Met k het aantal algoritmen, met b het aantal gemaakte voorspellingen per algoritme en $N = b \times k$, is de test statistiek gelijk aan:

$$T = k(k-1)\frac{\sum_{j=1}^{k}(X_{\bullet j} - \frac{N}{k})^2}{\sum_{i=1}^{b} X_{i\bullet}(k - X_{i\bullet})}$$

En wordt verworpen voor significantieniveau α wanneer $T > \chi^2_{1-\alpha,k-1}$.

Het aantal voorbeelden dat goed werd geclassificeerd door algoritme A en fout door algoritme B ($A > B$) kan worden vergeleken met het aantal dat B goed heeft geclassificeerd en A fout ($B > A$). Van de n instanties noteren we met m het aantal instanties waar de algoritmes een verschillende output geven. We nemen aan dat deze onafhankelijk zijn en een Bernouilli proces weergeven. Met s (succes) noteren we het aantal $A > B$ en met f noteren we het aantal $B > A$. De verwachting van beiden is gelijk aan een half wanneer de algoritmen even goed presteren: $\mathbb{E}(s) = \mathbb{E}(f) = 0.5$. Wanneer $s > f$ willen we de kans bepalen dat A minstens zo vaak meer 'wint', wat kan worden berekend aan de hand van de binomiale verdeling:

$$P(s \geq \text{waargenomen } s \mid p(s) = 0.5) : \frac{m!}{s!(m-s)!}p^s q^{m-s}$$

Een vergelijkbare test is de McNemar's test. De McNemar's test neemt aan dat $\frac{(|f-s|-1)^2}{s+f}$ bij benadering verdeeld is volgens de χ^2 verdeling met 1 vrijheidsgraad.

Regressie

Voor regressie zijn de uitkomsten per voorbeeld een verschil in een (eventueel gekwadrateerde) numerieke fout. Een aantal testen kunnen worden toegepast op deze aparte fouten; echter wordt met deze test niet iets gezegd over de uiteindelijke maat die wordt geoptimaliseerd, en of deze maat significant verschilt. Afhankelijk van de test en de aannames wordt de nulhypothese getoetst of één 'stochastisch groter' is, hierbij wordt, afhankelijk van de test, niet altijd de grootte van de fout direct meegenomen. Om toch iets te zeggen over significante verschillen in een maatstaf zoals de RMSE moeten er meerdere RMSE beschikbaar zijn om hierop de toetsen uit te voeren. Hiervoor moet bij de onderverdeling van de data rekening worden gehouden een mogelijkheid hierin is bijvoorbeeld de 5x2cv test [Dietterich, 1998].

Voor regressie kan gebruik worden gemaakt van ANOVA om te bepalen of er een significant verschil zit tussen de gemaakte fouten of de maatstaf. Deze test wordt toegepast om na te gaan of de gemiddelden hetzelfde is tussen groepen. Het test de nulhypothese dat er geen systematische verschillen zitten in de gemiddelden en de alternatieve hypothese dat er ten minste één verschillend is. Het meet de variantie tussen groepen en binnen groepen en vergelijkt deze relatieve groottes. ANOVA is een veel toegepast en breed bestudeerde techniek binnen de statistiek. Er is veel literatuur aan de ANOVA techniek gewijd, voor een goede introductie verwijzen we naar Lane [2016].

Om ANOVA toe te passen worden een aantal aannames gemaakt, zo moet de verdeling van de residuals normaal verdeeld zijn en wordt aangenomen dat de variantie is de verschillende groepen hetzelfde is. Ondanks onenigheid over de toepasbaarheid van ANOVA wanneer deze aannames niet houden, is het ook altijd mogelijk om de niet parametrische Kruskal–Wallis H-test te gebruiken.

Wanneer een verschil bestaat kunnen gepaarde test worden gebruikt voor de combinaties van uitkomsten. Wanneer het verschil een normale verdeling volgt kan gebruik worden gemaakt van de gepaarde t-test. Wanneer het verschil geen normale verdeling volgt kan gebruik worden gemaakt van de Wilcoxon test.

Voor een gepaarde steekproef wordt aangenomen dat de verschillen Z volgen uit een aselecte steekproef van de normale verdeling, vervolgens wordt getoetst: $H_0 : \mu_Z = 0$ tegen $H_1 : \mu_Z > 0$. Voor een steekproef van omvang n en gemiddelde $\bar{Z}$ uit een normale verdeling met de verwachtingswaarde μ en standaardafwijking σ kan worden nagegaan of de verwachtingwaarde een bepaalde waarde μ_0 (in dit geval 0) heeft. Onder de nulhypothese is dit gemiddelde ook normaal verdeeld

met verwachting μ_0 en standaardafwijking $\sigma/\sqrt{n}$. De standaardafwijking is vaak niet bekend en wordt geschat door de steekproefstandaardafwijking: $T = \frac{\bar{Z}-\mu_0}{S/\sqrt{n}}$. Deze toetsingsgrootheid T is onder de nulhypothese niet meer standaardnormaal verdeeld, maar volgt een t-verdeling.

Voor de Wilcoxon test wordt met $x_{j,i}$ de observatie op voorbeeld i door algoritme j genoteerd. Voor $i = 1, \ldots, n$ wordt het absolute verschil $\mid x_{2,i} - x_{1,i} \mid$ en het teken (sign) van $(x_{2,i} - x_{1,i})$ berekend. Alle voorbeelden waarvoor het (absolute) verschil gelijk is aan nul worden vervolgens verwijderd en het nieuwe aantal observaties wordt genoteerd met n_r. De overige absolute verschillen worden geordend, en de test statistiek W wordt berekend door:

$$W = \sum_{i=1}^{n_r} [(\text{sign}(x_{2,i} - x_{1,i}) \times R_i]$$

Waarbij R_i de rang is. Deze waarde wordt vergeleken met de statistiek waarvan de verdeling onder de nulhypothese bekend is.

Meerdere significantietoetsen

Omdat de prestaties van meerdere algoritmes worden beschouwd moet rekening gehouden worden met het feit dat de kans dat de test uitwijst dat een algoritme significant beter presteert, terwijl dit niet zo is, groter is: de kans op een uitzonderlijk geval neemt toe met het aantal hypotheses dat getest wordt, en daarmee de waarschijnlijkheid dat een nulhypothese incorrect wordt verworpen [Bland and Altman, 1995].

Er zijn verschillende technieken om dit tegen te gaan. Deze technieken compenseren voor het aantal inferenties dat gedaan wordt. Eén van deze manieren is de Bonferroni correctie. Het idee is dat wanneer m hypotheses worden getest, bij de individuele tests als significantieniveau $1/m$ keer het algemene niveau wordt aangehouden. Deze correctie wordt als conservatief beschouwt, voornamelijk wanneer veel hypotheses worden getoetst en wanneer test statistieken gecorreleerd zijn. Een tweede manier is de Holm–Bonferroni correctie en werkt als volgt: De hypotheses $H_1, ..., H_m$ met bijbehorende p-waarden $p_1, ..., p_m$ worden van laag naar hoog geordend: $p_{(1)} \ldots p_{(m)}$ en de bijbehorende hypotheses worden aangeduid met $H_{(1)} \ldots H_{(m)}$. Voor significantieniveau α, is k de minimale index zodat $p_{(k)} > \frac{\alpha}{m+1-k}$. Vervolgens worden de nulhypothesen $H_{(1)} \ldots H_{(k-1)}$ verworpen en $H_{(k)} \ldots H_{(m)}$ aangenomen.

6. Opzet

In dit hoofdstuk zal de werkwijze worden beschreven. Hierbij komt de gebruikte software aan bod en wordt de manier van valideren van de prestaties besproken.

6.1 Opzet

Onderverdeling data

Er is een opdeling gemaakt zoals weergeven in Figuur 6.1. Eerst wordt de data op basis van de klasse gebalanceerd verdeelt in twee delen, één van 70% en één van 30%. Met het deel met 70% van de trainingsvoorbeelden worden de beste parameters bepaald per model met 5-fold cross validatie. Met de parameters van het beste model wordt een uiteindelijk model opgesteld op de volledige 70%. Dit model wordt vervolgens vergeleken met de modellen die zijn opgesteld met andere technieken op de testset die 30% van de data bevat.

Figuur 6.1: Verdeling data in train, validatie en test sets

In de opdeling van de sets is gebruik gemaakt van dezelfde random seed. Dit betekent dat de opdeling van de data voor elk model op dezelfde manier gebeurd. Door deze opzet te gebruiken kunnen de resultaten met elkaar vergeleken worden met behulp van gepaarde tests.

Prestatiematen

Voor classificatie zijn er hoofdzakelijk twee prestatiematen aangehouden om direct de prestaties vergelijken: de accuracy en de lift grafiek. De keuze voor de accuracy is gemaakt omdat er geen sprake is van een sterke onbalans bij de verdeling van de afhankelijke variabele, er is geen noodzaak om verschillende fouten anders te wegen en ook in de besproken literatuur wordt voornamelijk de accuracy aangehouden. Verder is de interpretatie van accuracy duidelijk.

Voor regressie is de keuze is gemaakt bij de keuze voor het maken van de beste parameters voor een model gebruik te maken van de Root Mean Squared Error. Met deze keuze wordt een grotere fout zwaarder gewogen waardoor de langstudeerders met een hoge ratio meer worden benadrukt.

Software

Alle modellen zijn opgesteld in de programmeertaal R [R Core Team, 2014]. De gebruikte implementaties van de gebruikte algoritmes zijn de volgende: de rpart package [Therneau et al., 2015] is gebruikt om regressie bomen te bouwen en snoeien, de randomForest package [Liaw and Wiener, 2002] is gebruikt om Random Forests aan te leggen, voor het Naïve Bayes model is gebruik gemaakt van de klaR package [Weihs et al., 2005] en voor Tree Augmented en Weighted Naïve Bayes is gebruik gemaakt van het bnclassify package [Bojan et al., 2014]. Voor Support Vector Machines is de kernlab package [Karatzoglou et al., 2004] gebruikt, verder is het Bayes GLM model opgesteld met het arm package [Gelman and Su, 2015], en is gebruik gemaakt van het gbm package [with contributions from others, 2015] voor zowel het Adaboost als het GBM model. Voor het bepalen van het belang van verschillende variabelen, het toepassen van verschillende pre-processing technieken en het trainen van modellen (parameter tuning) is gebruik gemaakt van de caret package [Kuhn et al., 2015], voor veel van het worstelen van de data is de dplyr package [Wickham and Francois, 2015] gebruikt. Voor het toepassen van verschillende discretisatie algoritmen is gebruik gemaakt van de discretization package [Kim, 2012].

Instellingen

Voor een aantal technieken en instellingen zijn proefnemingen gedaan om te onderzoeken wat op het eerste gezicht goed werkt en of het het waard is om meer tijd te investeren om bepaalde technieken verder uit te werken. Deze proefnemingen zijn gedaan met de leerlinggebonden en opleidingsgegevens die beschikbaar zijn aan het begin van de studie. We kunnen niet met zekerheid zeggen dat deze eerste resultaten met meer inspanning later beter zouden presteren of wellicht beter zouden presteren op andere data.

Algoritmen verschillen in het aantal parameters dat zijn in te stellen. Het kiezen van deze parameters, ook bekend als het 'tunen', heeft als doel de parameters zo te kiezen dat gegeven zaken als tijd en hardware, prestaties worden geoptimaliseerd. Per techniek verschilt de tijd die nodig is om één model op te stellen en is deze daarnaast afhankelijk van de manier van tunen. Zo profiteren enkele algoritmen van parallel processing, zal het bouwen van een Decision Tree in vergelijking met bijvoorbeeld boosting technieken een stuk minder tijd kosten, en kan de invloed van de complexiteitsparameter bij een Decision Tree snel worden nagegaan door slechts een enkele boom op te stellen en de prestaties te vergelijken door te stoppen met snoeien voor verschillende parameterwaarden.

Sommige technieken hebben meer parameters. Hieruit volgt de kwestie of modellen met veel parameters meer rekentijd moeten krijgen om deze te optimaliseren. Er is aangehouden dat met de huidige setup het tunen van de modellen niet langer duurt dan drie uur, al zit het merendeel van de technieken ruim onder deze grens.

Voor Decision Tree en Random Forest zijn naast de technieken gebaseerd op CART proeven gedaan met bomen gebaseerd op 'conditional inference' aan de hand van de party package [Hothorn et al., 2006]. Eerste vergelijkingen lieten echter geen betere resultaten zien bij bomen opgesteld gebaseerd op dit framework. Ook Random Forest bestaande uit conditional trees zijn niet verder verkent, deze leverde bij het opnieuw inladen geheugen problemen op in R.

Voor de SVMs zijn een aantal proefnemingen gedaan om de prestaties van verschillende kernels te verkennen. De radial basis (Guassian) kernel presteerde hierbij het best binnen beperkte rekentijd en is verder gebruikt bij het bij het opstellen van modellen. Voor benadering van een goede waarde voor sigma is gebruik gemaakt van de ingebouwde hyperparameter schatter sigest, de kostenparameter is gevarieerd. Voor regressie is gebruik gemaakt van epsilon regressie met een epsilon van 0.1.

Voor Gradient Boosting zijn Decision Trees als weak learners gebruikt. Bij het parameter tunen is de shrinkage parameter vastgesteld op 0.1 en is het minimum aantal voorbeelden om een splitsing te maken in een node constant gehouden op 10. Het aantal splitsingen per boom is gevarieerd van één tot zes, en het aantal bomen van 50–300, met stappen van 50. Dezelfde opties zijn gebruikt voor het Adaboost algoritme. De fractie van de data dat wordt gebruikt om deze modellen op te

stellen is 50%. Het belang van variabelen is afgeleid door het effect te meten van de verschillende splitsingen op de fout in de onderste node.

Voor de Decision Trees zijn de volgende parameters gebruikt: het minimum aantal voorbeelden om een split te maken is vastgesteld op 20, het minimum aantal voorbeelden in een eindnode is gelijk gesteld aan zes. Met deze parameters wordt een enkele boom opgesteld, vervolgens wordt de beste complexiteitsparameter vastgesteld door te stoppen met snoeien voor verschillende parameterwaarden.

Het Random Forest bestaat uit 500 bomen en er wordt gevarieerd bij het aantal variabelen dat wordt gebruikt bij een splitsing. Voor het baggen wordt gebruik gemaakt van een steekproef met vervanging ten grootte van het aantal voorbeelden.

Voor Naïve Bayes is het mogelijk wel of niet een Laplace schatter te gebruiken, beide opties zijn geprobeerd. Ook is het mogelijk om een kernel te gebruiken om de verdeling van numerieke variabelen om te zetten naar numerieke waarden. Deze optie is niet onderzocht omdat deze niet beschikbaar zijn bij de uitbreidingen. Voor Tree Augmented Naïve Bayes is het mogelijk een score functie toe te kennen die wordt gebruikt voor het opstellen van het netwerk, hierbij zijn alle opties geprobeerd (de loglikelihood, bic en aic). Voor het bayes GLM model is voor classificatie gebruik gemaakt van de binomiale familie en een logit link functie, voor regressie is gebruik gemaakt van de gaussian familie en de identity link functie.

Preprocessing

Voor modellen waarbij niet is onderzocht in welke mate variabelen belangrijk zijn is gebruik gemaakt van de zogenaamde 'formule interface' van R. Hiermee worden categorische variabelen gecodeerd als m binaire variabelen. Het omzetten van categorische variabelen naar numerieke variabelen met behulp van een Bayesian Network is niet verder onderzocht omdat het veel tijd kostte om te implementeren omdat geen package bestaat waar deze preprocessing stap al in was geïmplementeerd en lieten eerste proefnemingen geen significante verbeteringen lieten zien.

Een aantal modellen hebben interne mechanismen die ervoor zorgen dat ze kunnen omgaan met missende waarden, hieronder vallen Decision Trees en de boosting modellen. Voor deze algoritmen zijn missende waarden doorgegeven aan het model, voor anderen zijn missende numerieke variabelen vervangen door de mediaan en missende categorische waarden zijn gecodeerd als nieuwe categorie.

Voor missende waarden die vaak missen zoals de toetsresultaten is geprobeerd deze in te vullen met behulp van kNN imputatie en mice. Dit leverde echter geen betere resultaten op: de prestaties op de groep waar de gegevens beschikbaar zijn gaan omlaag en bij de groep waar de gegevens worden ingevuld wordt ook geen winst behaald. Dit is mogelijk te wijten aan dat het moeilijk is om de toetsresultaten te voorspellen (in te vullen). Verder wordt het als het ware 'ruis' toegevoegd waardoor patronen in de variabelen die niet zijn ingevuld mogelijk niet meer worden opgepikt.

Verder zijn numerieke waarden zijn geschaald naar het interval $[0, 1]$ voor modellen die hier niet invariant onder zijn. Voor het Bayes GLM zijn variabelen geselecteerd door recursief variabelen te verwijderen zolang dit een verbetering opleverde.

Voor de Naïve Bayes methodes zijn numerieke variabelen gediscretiseerd. Hiervoor zijn verschillende manieren van discretiseren onderzocht. Er is gebruik gemaakt van de Fayyad-Irani MDL methode, de Liu-Setiono Chi2 methode met $\alpha = 0.05$ en het ('naïef') onderverdelen gebaseerd op enkel de verdeling van de variabele zelf. Omdat de discretisatie gebeurd binnen de validatie stap voor elke training set had dit een grote invloed op de rekentijd. Met de gebruikte implementaties met een factor hoger dan 100: waar het opstellen van zowel TAN als AWNB modellen gebeurd in minder dan drie seconden, duurt het discretiseren meer dan vijf minuten voor de Fayyad-Irani MDL methode en nog langer voor de Liu-Setiono Chi2 methode. Hoewel de meer geavanceerde manieren van discretiseren soms iets betere resultaten lieten zien ten opzichte van de 'naïve' methode van discretiseren (bij Naïve Bayes een verbetering van 0.5%, bij de TAN een verbetering van 0.1% en bij het gewogen model bleef de accuracy hetzelfde), is omdat een van de voordelen van de Naïve Bayes methoden juist de korte rekentijd is, ervoor gekozen de beste 'naïve' methode aan te houden: numerieke variabelen worden onderverdeeld in vier groepen. Dit onderverdelen wordt zo gedaan zodat in de train data de groepen ongeveer even veel trainingsvoorbeelden bevatten.

7. Resultaten

In dit hoofdstuk worden de resultaten besproken. We beginnen met de resultaten behaald met de gegevens die beschikbaar zijn bij de aanvang van de opleiding: dit zijn gegevens over de opleiding zelf, de gemiddelde prestaties op de opleiding en informatie over de leerling. Hierbij zal eerst een antwoord worden gezocht op de vraag die beantwoord wordt aan de hand van classificatie: zal een leerling een diploma halen? Vervolgens wordt de regressietaak behandeld waarbij de vraag beantwoord wordt: hoe lang zal een leerling doen over de studie? Dezelfde vragen zullen vervolgens worden beantwoord met gegevens over de vooropleidingen en de gegevens die beschikbaar zijn bij de momentopname na een half-jaar. Tot slot wordt het belang van variabelen in de opgestelde modellen besproken om de toegevoegde waarde van bijvoorbeeld open data bronnen te beoordelen.

7.1 Resultaten gegevens start opleiding

7.1.1 Classificatie

De eerste resultaten gaan over de voorspelling of een leerling succesvol zijn studie zal afronden op basis van gegevens die aan het begin van de opleiding beschikbaar zijn. De accuracy behaald op de testset is weergeven in Tabel 7.1.

In deze tabel geeft de eerste rij de resultaten weer die zijn behaald met enkel opleidingsvariabelen. Hierbij reflecteert de laatste kolom een classificatie die gemaakt wordt aan de hand van de regel om binnen een intern team de grootste groep (succes of geen succes) te voorspellen, wat een accuracy oplevert van 65.2%. Verrassend genoeg presteren de modellen opgesteld met (Average Weighted) Naïve Bayes, SVM en Bayes-GLM minder dan deze eenvoudige regel. Dat de Naïve Bayes methode minder presteert is misschien niet verrassend omdat de onafhankelijksheidsaanname niet op gaat, maar ook het AWNB model die deze aanname relaxeert presteert minder.

Het Random Forest model presteert het beste met een accuracy van 68%. De prestaties van beide boosting algoritmen komen vrijwel overeen, eveneens als met de prestaties van de Decision Tree. Met behulp van een Cochran's Q test is nagegaan of er een significant verschil zit tussen de prestaties: met een Cochran's Q van 11579 en 8 vrijheidsgraden (het aantal geteste algoritmen minus één) is de p-waarde $< 2.2 \times 10^{-16}$. Een gepaarde vergelijking met McNemar's tests met Holm–Bonferroni correctie laat vervolgens zien dat het GBM model significant beter presteert dan het AWNB model, het Naïve Bayes model, het Bayes-GLM model en het SVM model en sluit verder niet uit dat de overige behaalde resultaten significant verschillen met de resultaten van het GBM model.

De accuracy van de modellen die enkel de persoonlijke variabelen gebruiken zijn weergeven in de tweede rij. Uit het literatuuronderzoek bleek dat het lastig is een voorspelling te doen uitgaande van enkel achtergrondgegevens. Dit zien we ook (deels) terug in de resultaten: het percentage dat correct wordt voorspelt is lager dan wanneer opleidingsgegevens wordt gebruikt en het beste model voorspelt net iets meer dan 62% correct. Beter dan een simpele regel (het altijd voorspellen van de grootste klasse, succes) is het voor het merendeel van de modellen wel: McNemar's tests ($\alpha = 0.05$, Holm–Bonferroni correctie) verwerpen de nulhypothese dat er geen verschil bestaat tussen de verkeerd geclassificeerde voorbeelden ten opzichte van die van de simpele regel voor alle resultaten behalve die van het Random Forest en de SVM.

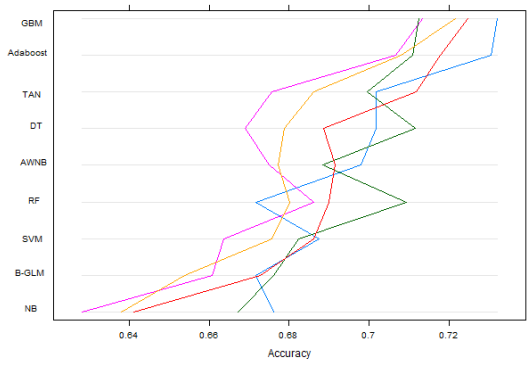
De boosting algoritmen presteren wederom goed en behalen een accuracy van 62%. De Decision Tree (61.5%) en AWNB (61.8%) modellen zitten hier dicht bij. Deze modellen, net als het TAN

model presteren volgens een McNemar's tests ($\alpha = 0.05$, Holm–Bonferroni correctie) niet significant minder dan het beste model. Waar het Random Forest met enkel opleidingsgegevens het best presteert, scoort deze beduidend minder met enkel leerlinggegevens.

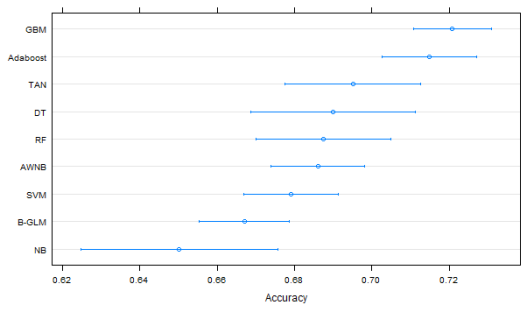
De laatste rij in Tabel 7.1 zijn de resultaten bij de combinatie van opleiding en persoonlijke variabelen en variabelen die iets zeggen over de gemiddelde prestaties bij een opleiding. Deze resultaten geven hiermee als het ware een beeld van hoe goed een opleiding past bij een leerling. Ten opzichte van de losse gegevens presteren alle technieken behalve Random Forest beter. Dat het Random Forest model met meer variabelen en embedded variabele selectie minder presteert is tegen verwachting. Waar de overige methodes dus winst halen bij de combinatie van variabelen, is dit voor het Random Forest niet het geval. De mogelijke oorzaak hiervan is onbekend. De boosting methodes presteren wederom het best en de overige volgorde is vergelijkbaar met wanneer leerlinggegevens worden gebruikt. McNemar's tests ($\alpha = 0.05$, Holm–Bonferroni correctie) wijzen uit dat de resultaten van alle modellen significant verschillen ten opzichte van de simpele regel. Verder presteert alleen het Adaboost model niet significant minder dan het best presterende GBM model.

Gegevens	SVM	Adaboost	GBM	DT	RF	B-GLM	NB	TAN	AWNB	Regel
Opleiding	64.9*	67.6†	67.6†	66.6†	**68.0**†	63.4*	61.9*	65.9†	62.6*	65.2
Leerling	55.6*	**62.2**†	61.9†	61.5†	56.3*	57.1	59.7	60.5†	61.8†	55.4
Combinatie	66.5	71.8†	**72.0**†	69.3	67.3	68.0	65.6	68.9	69.1	65.2

Tabel 7.1: Behaalde accuracy (percentage goed voorspelt) op de data beschikbaar aan het begin van de opleiding. De kolom 'regel' kan worden geïnterpreteerd als baseline en is voor de opleiding en combinatie het percentage behaalt door de grootste groep binnen een intern team te voorspellen en bij de leerling het voorspellen van de grootste groep. Met McNemar's tests met Holm–Bonferroni correctie en significantieniveau $\alpha = 0.05$ is aangegeven welke methoden *niet* significant beter presteren dan de simpele regel (*) en welke methoden niet significant minder presteren het beste model in de rij (†)



(a) Prestaties per cross-validatie fold



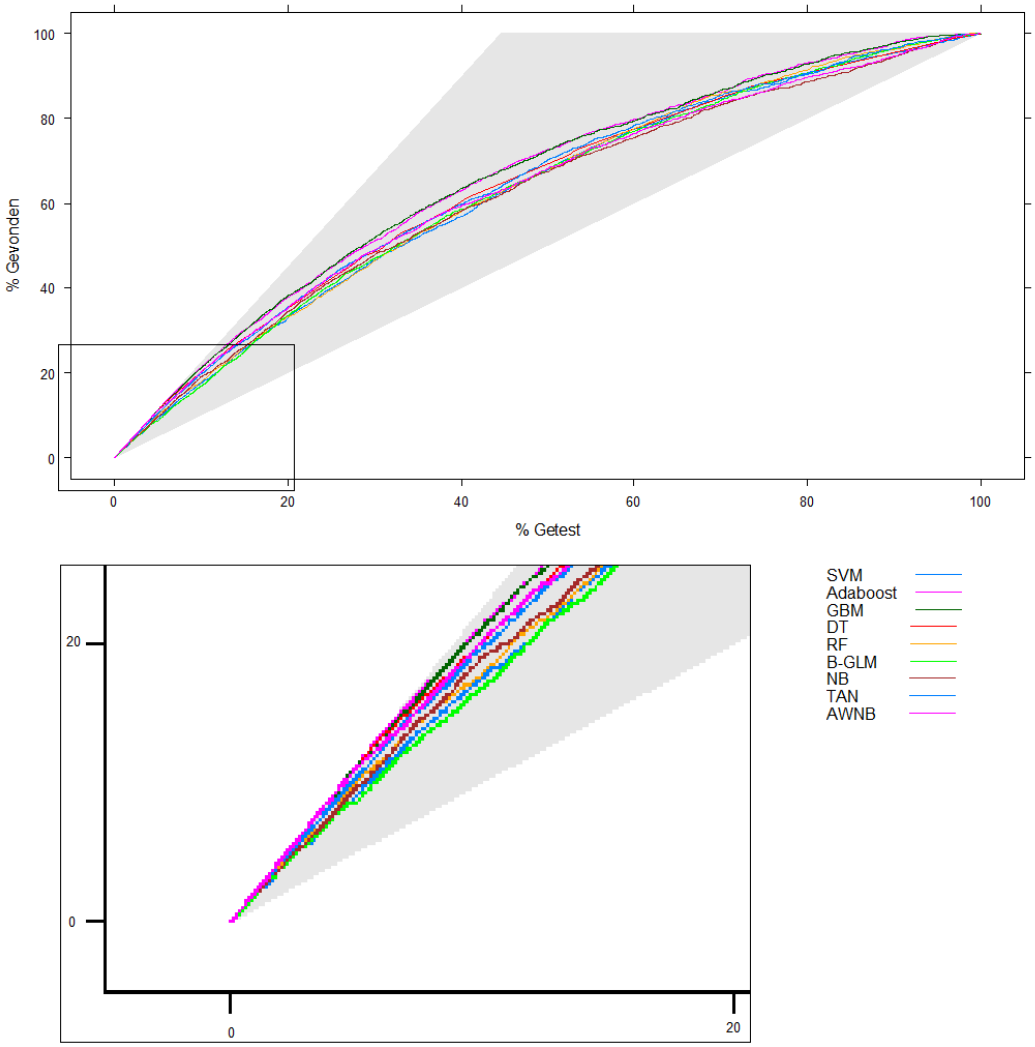
(b) Betrouwbaarheidsintervallen gebaseerd op een tweezijdige t-test op de cross validatie prestaties, met significantieniveau $\alpha = 0.05$ en bonferroni correctie

Figuur 7.1: Prestaties bij de classificatietaak over de cross-validatie folds

De prestaties behaald door de verschillende algoritmes in de trainingsfase op de verschillende folds is weergeven in figuur 7.1a. Dit zijn de resultaten behaald met de beste parameters, waarbij gebruik gemaakt wordt van alle variabelen die beschikbaar zijn aan het begin van de opleiding. Hieruit is af te leiden welke folds moeilijker zijn en wat de variantie in prestaties is. De validatie-folds die corresponderen met de gele en paarse lijken moeilijker: op deze wordt door vrijwel alle algoritmes het minst gepresteerd. Verder valt op dat de resultaten vergelijkbaar zijn over de verschillende folds: er is er geen waarbij een enkele techniek veel beter presteert dan alle anderen.

De betrouwbaarheidsintervallen die zijn bepaald op basis van de prestaties in cross validatie zijn weergeven in 7.1b. Deze zijn gebaseerd op een tweezijdige t-test, met significantieniveau $\alpha = 0.05$ en bonferroni correctie. Zoals verwacht vertoont de Decision Tree veel variantie, eveneens heeft ook de minst presterende Naïve Bayes methode een groot betrouwbaarheidsinterval. De boosting algoritmen die het best presteren vertonen ook niet veel variantie, wat gewenst is: de mate waarin een nieuwe voorspelling te vertrouwen is ligt hoger bij een lager variantie.

In Figuur 7.2 is de lift grafiek voor de resultaten op de test set bij de combinatie van opleiding variabelen en persoonlijke variabelen weergeven. Hieruit blijkt dat goed wordt gepresteerd op de eerste 20% van de leerlingen waarvoor de hoogste kans op uitval wordt voorspelt. Dit betekent dat er voor een groep effectief voorspelt kan worden dat zij hoge kans hebben om uit te stromen zonder diploma. Na ongeveer 20% zwakken de prestaties af. Dit kan worden geïnterpreteerd als dat het voor de rest van de groep lastig is om te voorspellen of zij zullen gaan slagen of niet. Verder valt op dat de beste technieken het goed over het gehele bereik.



Figuur 7.2: Lift grafiek van de classificatie waarbij alle gegevens worden gebruikt die beschikbaar zijn aan het begin van de opleiding

De groep van leerlingen die door het best presterende GBM model worden voorspelt als behorende tot de 10% met de hoogste kans om uit te stromen zonder diploma kenmerkt zich als volgt. Het gemiddelde niveau ligt met 2.78 ten opzichte van het gemiddelde van 2.52 van

de hele leerlingenpopulatie een stukje hoger. Met 26.8% ligt het percentage bbl lager dan voor de hele populatie (32.2%) verder is het percentage dat een opleiding met een verzwaarde intake volgt een stuk hoger met 10.2% ten opzichte van 2.9%. Verder is het percentage met Nederlandse Nationaliteit in de gehele leerlingenpopulatie hoger (72.7% ten opzichte van 52.8%). Geen van de extranei zitten in de groep met hoge uitvalkans en verder bevat de groep minder deeltijdleerlingen (26.1% ten opzichte van 32.7%).

Op dit moment is op hoger niveau bij de onderwijsinstelling nog geen manier om in te schatten of een leerling succesvol zal zijn nauwkeuriger dan uitgaande van het succespercentage van een opleiding op zich. Uit de behaalde resultaten blijkt dat het mogelijk is met een voorspelmodel een groep te onderscheiden waarin de kans dat wordt uitgestroomd zonder diploma hoog is. Het is mogelijk deze groep effectief begeleiding te bieden. De groep waarin de voorspelde kans op uitval hoog is vertoont wel andere kenmerken dan de gehele leerlingenpopulatie, maar er zijn geen allesomvattende kenmerken binnen deze groep. De combinatie van deze resultaten benadrukken de meerwaarde van een voorspelmodel des te meer.

7.1.2 Regressie

Naast te voorspellen of een leerling succesvol de opleiding zal afronden is ook voorspelt hoe lang de leerling hierover zal doen. Hiertoe is de ratio, de werkelijke duur gedeeld door de verwachte duur, voorspelt. Als maatstaf wordt de Root Mean Squared Error (RMSE) aangehouden. De resultaten behaalt op de testset zijn te vinden in Tabel 7.2, hierbij is weer dezelfde indeling van variabelen aangehouden. De 'Regel' kolom bevat de RMSE die wordt behaald door altijd de gemiddelde ratio te voorspellen.

De eerste rij in Tabel 7.2 geven de resultaten behaalt met enkel opleidingsvariabelen weer. De ensemble modellen presteren wederom het beste. De Decision Tree en SVM modellen vormen de middenmoot, en het Bayes-GLM model presteert het minst. De RMSE behaalt met enkel de persoonlijke variabelen zijn weergeven in de tweede rij. Ook hier blijkt weer dat het lastig is een voorspelling te doen enkel op basis van achtergrondgegevens van een leerling. Beter dan de simpele regel is het voor al de opgestelde modellen wel, maar de verbetering is in vergelijking met de andere combinaties van variabelen minimaal. De resultaten in de laatste rij laten net als bij classificatie zien dat alle technieken behalve Random Forest beter presteren ten opzichte van de losse gegevens. De verschillen zijn echter vrij klein maar het beste model evenaart enkel de beste prestatie met enkel de opleidingsgegevens. Ondanks dat de gegevens moeilijk te interpreteren zijn, kan wel worden geconcludeerd dat de combinatie van leerling- en opleidingsgegevens geen grootse verbetering opleveren in de voorspelling.

In vergelijking met de maatstaf die gebruikt wordt bij classificatie is het moeilijk een directe interpretatie van de absolute waarden bij de RMSE. Er wordt met de resultaten wel een relatief beeld geschetst, maar de mate van verbetering is moeilijk in te schatten: Wat is het verschil tussen een RMSE van 0.390 en 0.396 en 0.456?

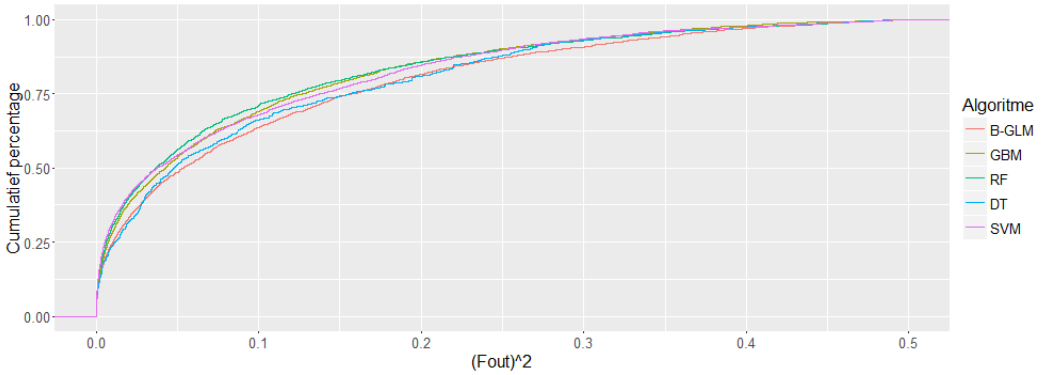
Gegevens	SVM	GBM	DT	RF	B-GLM	Regel
Opleiding	0.415	0.396	0.414	0.390	0.438	0.474
Leerling	0.464	0.456	0.459	0.461	0.466	0.474
Combinatie	0.411	0.390	0.413	0.391	0.426	0.474

Tabel 7.2: Behaalde RMSE op de data beschikbaar aan het begin van de opleiding. De kolom 'regel' kan worden geïnterpreteerd als baseline en is de RMSE die behaald wordt wanneer altijd de gemiddelde ratio wordt voorspeld.

De individuele gemaakte kwadratische fouten op de combinatie van leerling- en opleidingsgegevens onder de 0.5 zijn weergeven in Figuur 7.3. Met behulp van de Kruskal-Wallis H-test is gekeken of er verschil zit in de verdeling van deze individuele fouten. Voor alle drie de combinaties van variabelen wordt de nulhypothese dat er geen verschil is verworpen, de hoogste p-waarde is p =

0.027 voor de fouten gemaakt door de modellen op enkel de leerlinggegevens. Met behulp van een Wilcoxon test met Holm–Bonferroni correctie is gekeken of er een significant verschil bestaat tussen de resultaten en die behaald door de simpele regel en tussen het beste model in vergelijking met de andere modellen. De gevonden p-waarden zijn weergeven in Tabel 7.3. Bij de vergelijking met de simpele regel zijn alle p-waarden kleiner dan $\alpha = 0.05$ en geeft de Bayes-GLM met leerlinggegevens de kleinste p-waarde van 0.033. Wanneer er een vergelijking wordt gemaakt met het beste model wordt de nulhypothese niet verworpen voor de SVM, voor de overige algoritmen wel. Dit lijkt vreemd omdat de SVM dichtbij de DT zit op basis van de RMSE. Dit geeft dan ook aan dat deze niet als test voor prestaties op basis van RMSE moet worden beoordeeld.

Omdat enkel de individuele fouten worden vergeleken, en niet de RMSE die als maatstaf is aangehouden, is ervoor gekozen om bij de overige modellen deze toets niet toe te passen. Vanwege de opzet zijn er niet meerdere RMSE beschikbaar, wel kan gekeken worden naar de RMSE in de validatiesets om een beoordeling te doen over de mate waarin een model hier beter presteert dan een ander.



Figuur 7.3: Cumulatieve verdeling van de individuele kwadratische fouten gemaakt door de verschillende algoritmen op de combinatie van leerling- en opleidingsgegevens.

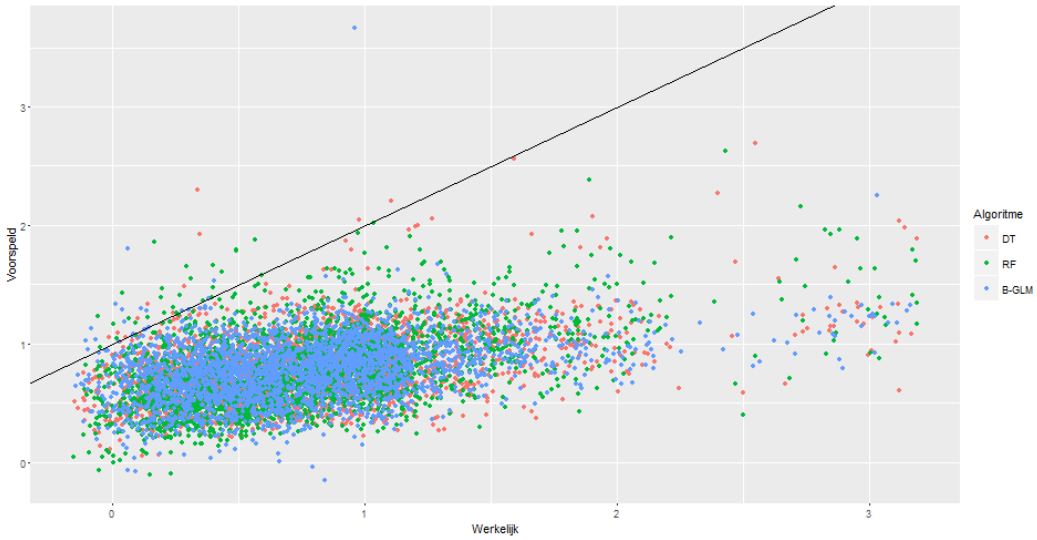
	Beste					**Regel**				
Gegevens	DT	B-GLM	SVM	GBM	RF	DT	B-GLM	SVM	GBM	RF
Opleiding	1.527E-09	0.000	0.115	0.000	-	1.720E-24	0.000	0.000	6.718E-38	0.000
Leerling	1.000E+00	0.037	1.000	-	1.000E+00	2.081E-04	0.033	0.000	7.345E-05	0.001
Combinatie	1.655E-04	0.000	0.607	-	4.362E-01	1.552E-20	0.000	0.000	1.054E-39	0.000

Tabel 7.3: P-waarden Wilcoxon test met Holm–Bonferroni correctie op de kwadratische fouten. Links wordt een vergelijking gemaakt met het beste model (zonder p-waarde) en rechts met de simpele regel van het voorspellen van de gemiddelde ratio.

Om inzichtelijk te maken hoe goed modellen kunnen inschatten of een leerling daadwerkelijk lang of kort bezig is, in Figuur 7.4 een zogenaamde 'jitter' plot gemaakt waarbij de voorspelde waarde van de ratio wordt afgezet tegen de daadwerkelijke ratio. Om het overzichtelijk te houden zijn de uitkomsten van drie algoritmen gebruikt, voor de missende algoritmen geldt dat de uitkomsten van de RF model erg lijken op die van de GBM en de uitkomsten van de SVM vertonen een soortgelijk patroon als de B-GLM. In een jitter grafiek wordt 'overplotten' tegengegaan en blijft de verdeling van de punten behouden. Wel wordt hierbij gebruik gemaakt van extra ruis waardoor punten niet direct meer hoeven te corresponderen met de originele datapunten.

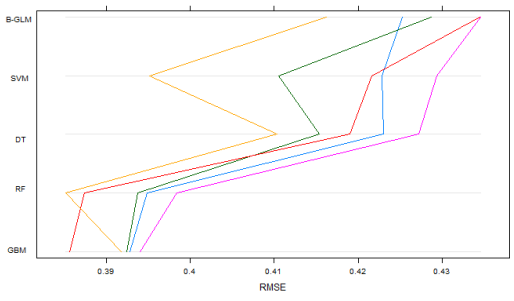
Bij perfecte regressie zouden alle punten op de zwarte lijn liggen, hier is de voorspelde waarde gelijk aan de werkelijke waarde. Bij alle technieken is de dichtheid hoog in het gebied tussen de 0 en 1.5 van de werkelijke waarde en onder de lijn van perfecte regressie. In de grafiek is te zien dat de voorspellingen systematisch te laag zijn, de modellen zijn terughoudend in hun voorspellingen.

Wel is te zien dat naarmate de werkelijke waarde toeneemt ook de voorspelde waarden over het algemeen hoger liggen, dit is in het speciaal goed te zien bij de punten die corresponderen met de uitkomsten van het RF model.

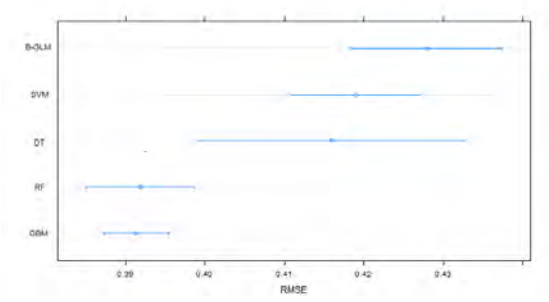


Figuur 7.4: Jitter plot van de voorspelde waarde van de ratio ten opzichte van de daadwerkelijke ratio. De voorspelling is gemaakt op de combinatie van leerling- en opleidingsgegevens.

Net als bij de classificatietaak is ook hier weer gekeken naar de prestaties in de cross-validatie om onder andere iets te zeggen over de variantie in de uitkomsten. De prestaties behaald door de verschillende algoritmes in de trainingsfase op de verschillende folds is weergeven in figuur 7.5a. Net als bij de resultaten bij de classificatie is te zien dat de resultaten vergelijkbaar zijn over de verschillende folds: een 'moeilijke fold' voor een enkele techniek is ook moeilijk voor de andere technieken. De betrouwbaarheidsintervallen die kunnen worden bepaald op basis van de prestaties in cross validatie is weergeven in 7.5b. Deze zijn gebaseerd op een tweezijdige t-test, met significantieniveau $\alpha = 0.05$ en bonferroni correctie. De best presterende (ensemble) technieken hebben (zoals gewenst) een klein betrouwbaarheidsinterval.



(a) Prestaties per cross-validatie fold



(b) Betrouwbaarheidsintervallen gebaseerd op een tweezijdige t-test op de cross validatie prestaties, met significantieniveau $\alpha = 0.05$ en bonferroni correctie

Figuur 7.5: Prestaties bij de regressietaak over de cross-validatie folds

7.2 Meerwaarde extra variabelen

Om te bepalen wat de meerwaarde is van variabelen die voor bepaalde groepen aanwezig zijn of niet beschikbaar zijn bij de aanvang van de opleiding zijn aparte modellen opgesteld. Hierbij verschilt de groep waarop de voorspelling wordt gemaakt voor ieder model en hiermee ook de 'baseline' waarmee wordt vergeleken. Omdat de samenstelling altijd overeen komt met de gehele groep van leerlingen is de generaliseerbaarheid in twijfel is te trekken, zoals besproken in Sectie 4.2.3.

7.2.1 Vooropleiding

Modellen zijn opgesteld voor de groep van 4,721 leerlingen waar variabelen die gaan over vooropleiding beschikbaar zijn. De resultaten hiervan zijn te vinden in Tabel 7.4.

	RMSE					**Accuracy**				
	DT	RF	SVM	B-GLM	GBM	DT	RF	SVM	B-GLM	GBM
Zonder	0.303	0.290	0.316	0.323	0.280	69.6	71.5	69.2	68.8	73.1
Met	0.309	0.291	0.316	0.319	0.274	71.2	73.4	70.0	69.7	72.2

Tabel 7.4: Verschil in prestaties wanneer de vooropleiding variabelen worden meegenomen; McNemar's test bij de classificatie resultaten vertellen dat inclusief Holm–Bonferroni correctie geen van de p-waarden significant zijn.

Voor de regressietaak hebben de Bayes-GLM en de best presterende GBM modellen baat bij de informatie over de vooropleiding. De overige modellen laten geen verbetering zien: de RF en DT modellen presteren zelfs wat minder. Bij de accuracy is daarnaast wel voor het meerendeel van de technieken een verbetering te zien. Waar het GBM model zonder de vooropleiding nog het best presteerde, presteert deze echter met deze gegevens minder. Het RF model presteert met deze gegevens beter dan alle overige modellen. Wederom is met McNemar's getest of het verschil tussen de prestaties significant is: inclusief Holm–Bonferroni correctie en $\alpha = 0.05$ wordt geen van de nulhypotheses dat er geen verschil bestaat verworpen met als laagste p-waarde 0.143 bij de resultaten van het RF model.

7.2.2 Gegevens momentopname half-jaar

Na een half jaar wordt een momentopname gedaan waarbij elke leerling die binnen een half jaar klaar is verwijderd. Ook zijn variabelen die iets zeggen over de opleiding geüpdatet.

Wisselen opleiding

Bij de momentopname is bekend hoe vaak een leerling van opleiding is gewisseld. Na een half jaar zijn van de 6,630 leerlingen er slechts 229 minstens één keer van opleiding gewisseld. Een nieuwe voorspelling is gemaakt waarbij variabelen zijn toegevoegd die aangeven of een leerling van opleiding is gewisseld, hiervan zijn de resultaten te vinden in Tabel 7.5. Voor classificatie wordt de nulhypothese dat er een verschil is McNemar's tests voor geen van de modellen verworpen: zelfs zonder Holm–Bonferroni correctie correctie liggen de p-waarden ver van het kritieke gebied, de laagste zijnde $p = 0.35$ voor de resultaten van het GBM model. Waar uit de literatuur en bij de data analyse het dus leek dat het wisselen mogelijk een goede voorspeller zou zijn, blijkt dat de modellen opgesteld die gebruik maken van deze informatie geen significante verbetering laten zien, wat waarschijnlijk komt door de beperkte grootte van de groep die na een half jaar is gewisseld.

Toetsresultaten

Na een half jaar zijn ook de eerste toetsresultaten beschikbaar. Zoals besproken in Sectie 4.2.3 zijn er slechts beperkt resultaten beschikbaar, en bestaat de groep waarvoor deze beschikbaar zijn

	RMSE					Accuracy				
	SVM	GBM	DT	RF	B-GLM	SVM	GBM	DT	RF	B-GLM
Zonder	0.408	0.381	0.404	0.383	0.417	68.2	71.9	69.1	68.1	67.5
Met	0.406	0.382	0.404	0.383	0.417	68.3	71.5	69.1	68.1	67.4

Tabel 7.5: Verschil in prestaties wanneer de 'ommezwaai' variabelen worden meegenomen; McNemar's test bij de classificatie resultaten vertellen dat geen van de verschillen significant zijn.

voornamelijk uit leerlingen uit de recentere cohorten.

Naast de toetsresultaten is voor iedere leerling het aantal ommezwaaien bekend. Wanneer modellen voor deze groep worden opgesteld met deze gegevens is net als bij de vorige groep geen significante verbetering (Tabel 7.6). De resultaten behaald met wanneer variabelen worden meegenomen die iets zeggen over de resultaten staan in de onderste rij in Tabel 7.6. Er is nagegaan of er een significant verschil bestaat tussen de uitkomsten per algoritme met een Cochran's Q test. Geen van de p-waarden zijn significant: de laagste p-waarde van het Random Forest komt zelfs zonder p-waarden correctie, met 0.439, niet in de buurt van het kritieke gebied. Hoewel de verschillen in percentages in Tabel 7.6 groot lijken, komt dit doordat het slechts de testset voor een klein groep van 210 leerlingen beslaat.

	RMSE					Accuracy				
Gegevens	DT	RF	SVM	B-GLM	GBM	DT	RF	SVM	B-GLM	GBM
Begin	0.292	0.278	0.329	0.317	0.286	81.4	81.9	79.5	80.5	81.0
Ommezwaaien	0.292	0.278	0.330	0.317	0.285	81.4	81.9	79.5	80.5	81.0
Resultaten	0.305	0.266	0.304	0.305	0.282	81.4	83.3	80.5	78.1	83.3

Tabel 7.6: Verschil in prestaties met de toetsresultaat variabelen; Cochran's Q test bij de classificatie resultaten wijst uit dat er per kolom geen significante verschillen zijn.

Belang van variabelen

In deze sectie zullen we de rol die variabelen spelen in de opgestelde modellen onderzoeken. In Hoofdstuk 5, Technieken, is voor verschillende technieken besproken hoe het belang van variabelen in opgestelde modellen kan worden afgeleid. Voor de gebruikte technieken is dit mogelijk bij de RF, GBM, Adaboost en DT algoritmes. Er is gekozen om enkel het belang van de GBM modellen te weergeven om het overzicht te bewaren, het mogelijk maken om een directe vergelijking te maken en verder presteert GBM consistent goed.

Bij de belangrijke variabelen in het model opgesteld met de data die voor iedereen aan de start van een opleiding beschikbaar is (Tabel 7.7) valt op dat de stad en beroepsopleiding bij zowel regressie als classificatie de belangrijkste variabelen zijn. Vooral dat de stad zo belangrijk is, is opvallend. Verder is een interessante observaties dat voor classificatie het gemiddelde succespercentage van belang en bij regressie de gemiddelde duur. Verder is ook de geplande duur van belang in het regressiemodel. Verder valt op dat de variabelen uit open data bronnen niet terugkomen, alleen de studieintensiteit die deels gebaseerd is op de studiebelasting die uit open data gegevens afkomstig is, maar ook op de geplande duur, wat waarschijnlijk de reden is dat deze variabele van belang is. Wanneer vooropleidingsgegevens worden meegenomen (Tabel 7.8) zijn de variabelen die hierover gaan van groot belang. De vooropleiding school, code en tussentijd komen zowel bij de regressie als de classificatietaak terug in de top tien. Verder valt op dat dezelfde variabelen als in het vorige model weer een grote rol spelen. Wanneer gebruik wordt gemaakt van variabelen die gaan over toetsresultaten (Tabel 7.9), zien we deze variabelen slechts beperkt terugkomen. Bij classificatie komt de eerste pas terug op plek acht en bij regressie op plek tien. Verder valt op dat er geen bepaalde variabele is uit de toetsresultaten die van belang is.

Classificatie		Regressie	
Variabele	**Relatief belang**	**Variabele**	**Relatief belang**
Stad	100.00	Stad	100.00
Beroepsopleidingnaam	59.20	Beroepsopleidingnaam	82.31
Gemsucces	38.61	Geplandeduur	33.40
VorigNiveau	23.21	DeelnemerGeboorteland	24.44
Gemeente	13.51	Opleidingscluster	24.17
Opleidingscluster	12.19	Leeftijdopdatuminstroom	14.80
DeelnemerGeboorteland	7.08	Grootteteam	13.55
Studieintensiteit	6.02	Intensiteit	12.54
Postcode	5.06	Gemduur	11.94
Intensiteit	3.78	CrebokoppelSubgroep	11.08

Tabel 7.7: Belangrijkste tien variabelen in de GBM modellen die gebruik maken van een combinatie van opleiding en persoonlijke gegevens

Classificatie		Regressie	
Variabele	**Relatief belang**	**Variabele**	**Relatief belang**
Stad	100.00	Geplandeduur	100.00
Voorropleidingschool	89.49	Voorropleidingschool	93.74
Beroepsopleidingnaam	67.82	Stad	71.81
Gemsucces	34.31	Beroepsopleidingnaam	36.45
Opleidingscluster	22.25	Vooropleidingtussentijd	35.30
Vooropleidingcode	18.63	Vooropleidingcode	27.16
Vooropleidingtussentijd	14.51	Opleidingscluster	23.21
VorigNiveau	11.33	Studieintensiteit	20.36
Geplandeduur	7.24	CrebokoppelSubgroep	19.44
Gemeente	4.28	Grootteteam	17.04

Tabel 7.8: Belangrijkste tien variabelen in de GBM modellen die gebruik maken van een combinatie van opleiding, persoonlijke gegevens en vooropleidingsgegevens

Classificatie		Regressie	
Variabele	**Relatief belang**	**Variabele**	**Relatief belang**
Beroepsopleidingnaam	100.00	Geplandeduur	100.00
Stad	77.54	Beroepsopleidingnaam	39.66
Opleidingscluster	64.60	Stad	28.67
Gemeente	26.18	Gemeente	23.43
Geplandeduur	12.35	Opleidingscluster	20.59
VorigNiveau	9.39	DeelnemerGeboorteland	12.05
Studieintensiteit	3.26	Grootteteam	11.66
Gemiddeldbovengemhalf	2.95	Studieintensiteit	9.55
Grootteopleiding	2.80	Geplandeduur	6.99
DeelnemerGeboorteland	2.60	Minresultaathalf	6.19
Grootteteam	1.82	Gemiddeldbovengemhalf	5.11

Tabel 7.9: Belangrijkste elf variabelen in de GBM modellen die gebruik maken van een combinatie van opleiding, persoonlijke gegevens en toestresultaten na een half jaar

8. Conclusie

In dit onderzoek is een antwoord gezocht op de onderzoeksvraag:

"Wat is de toegevoegde waarde van interne-, externe data en voorspellende modellen op beleid voor studiesucces?" Deze hoofdvraag is opgedeeld in vier deelvragen. In dit hoofdstuk zullen we deze deelvragen bespreken en hiermee een antwoord geven op de hoofdvraag.

8.1 Factoren van invloed op studiesucces

De eerste deelvraag luidt als volgt: Welke factoren worden in de literatuur gedefinieerd als het hebben van invloed op studiesucces en hoe corresponderen deze met beschikbare variabelen?

In vergelijking met sommige andere onderzoeken zijn de gegevens die voor dit onderzoek beschikbaar zijn vrij beperkt. Twee factoren die veel naar voren komen maar helaas missen voor dit onderzoek, zijn delinquent gedrag en gegevens over ouders van leerlingen. Een grootschalig onderzoek naar de invloed van loopbaan en de kenmerken van opleidingen geven aan dat er vrij significante verschillen zitten tussen de leerwegen en niveau en blijkt ook dat de loopbaan een grote rol speelt. Vooropleidingsgegevens zijn voor dit onderzoek echter slechts beperkt (en in beperkte vorm) beschikbaar zijn. Studieuitval ligt vaak aan persoonlijke omstandigheden en uitval is daarnaast te danken aan een samenspel van variabelen.

Het onderzoek naar toepassingen van Machine learning gaf het algemene beeld dat wanneer er alleen gebruik wordt gemaakt van leerlinggegevens, er vaak geen significant betere voorspelling gedaan kan worden ten opzichte van een simpele classifier. Het gebruik van toetsresultaten zorgde in een groot deel van de toepassingen voor een betere voorspelling dan een referentiemodel. Hierbij bleek dat uitschieters vaak meer zeggen dan het 'standaardresultaat' van een leerling. Hiermee is rekening gehouden bij het afleiden van variabelen van de toetsresultaten.

8.2 Toegevoegde waarde voorspelmodel

De tweede deelvraag luidt als volgt: Hoe kunnen de uitkomsten van een voorspelmodel worden ingezet om toegevoegde waarde te creëren? Deze deelvraag is aan bod gekomen in de achtergrond waar behandeld is hoe analytics reeds wordt toegepast in het onderwijs en verder met het literatuuronderzoek naar toepassingen van Machine Learning.

In de achtergrond zijn twee verschillende toepassingen kort aan bod gekomen: één op microniveau, het onderscheiden van kansrijke en kansarme leerlingen, en één op mesoniveau, het sturen op basis van de uitkomsten van en analyses op een voorspelmodel.

Verder komt de toegevoegde waarde terug in de analyse van de uitkomsten van de modellen. Hierbij is om inzicht te geven in welke mate het mogelijk is om een onderscheid te maken tussen leerlingen die wel of geen diploma halen gekeken naar de lift grafiek. De resultaten schetsen het beeld dat op basis van de voorspelling of een leerling een diploma haalt (en of de leerling hier lang over gaat doen) begeleiding kan worden gestuurd. De voorspellingen zijn niet zo accuraat om op mesoniveau een veel accurater beeld te geven over toekomstige leerlingaantallen. Uit de resultaten kunnen een aantal inzichten gehaald worden, bijvoorbeeld over het verschil in kenmerken tussen de groep waarin de uitval hoog is en de gehele leerlingenpopulatie. De data en de uitkomsten zijn echter niet zo toereikend om direct aanbevelingen te doen over het aanpassen van schoolprocessen anders dan het sturen van begeleiden.

8.3 Externe data

De derde deelvraag gaat over externe data. Welke externe data kan worden ontsloten en hoe kan deze worden gekoppeld aan de beschikbare interne data?

De gebruikte open data bronnen zijn besproken in Hoofdstuk 4, Data. De eerste open data bron zijn gegevens van www.postcodedata.nl/, die gekoppeld zijn aan de postcode die van een leerling bekend is. De tweede open data bron zijn de gegevens van SCP, de SES Statusscores, die de sociaal-economische status in een wijk weerspiegelen. Ook deze gegevens zijn gekoppeld aan de (in dit geval 4-cijferige) postcode. De laatst gebruikte open databron is de DUO Crebokoppeltabel, een tabel die informatie bevat per opleiding in het mbo en aan de hand van het Crebonummer gekoppeld is aan de beschikbare data.

Verder zijn nog twee mogelijke uitbreidingen besproken. De eerste is Studie in Cijfers, welke een beeld geeft over een opleiding en de kansen van die opleidingen op de arbeidsmarkt. De tweede is de keuzegids, welke een overzicht van mbo-studies per vakgebied en per niveau ingedeeld in regio's met informatie over mogelijke beroepen, baankansen en het gemiddelde startsalaris.

Om de meerwaarde van de open data bronnen te bepalen is gekeken naar het belang van variabelen gevormd uit de open data bronnen in de opgestelde modellen met het GBM algoritme. Hierbij kwamen wel een aantal variabelen terug zoals de afstand tot de instelling, de statusscore en de prijsfactor behorende bij een opleiding, maar vielen deze voor alle modellen buiten de top tien. Of zonder deze variabelen de modellen minder gaan presteren is niet getest, vanwege het al grote aantal verschillende modellen dat wordt opgesteld. Er kan worden geconcludeerd dat deze variabelen mogelijk een kleine bijdrage leveren maar geen belangrijk deel van het model uitmaken.

8.4 Prestaties verschillende technieken en data bronnen

De laatste deelvraag wordt beantwoord in Hoofdstuk 7, Resultaten, en luidt: Hoe goed zijn de voorspellingen gemaakt door voorspelmodellen opgesteld met verschillende technieken en gebruik makend van verschillende data bronnen?

Voor de classificatie kan al een redelijk accurate voorspelling worden gedaan enkel op basis van opleidingsgegevens en een voorspelling enkel op basis van leerlinggegevens blijkt lastig. Wanneer deze gegevens worden gecombineerd verbeteren over het algemeen de resultaten, wat mogelijk kan worden geïnterpreteerd dat het mogelijk is om een inschatting te maken hoe goed een opleiding bij een leerling past. Voor de modellen opgesteld op deze gegevens apart geldt dat in alle gevallen significant beter wordt gepresteerd dan een simpele regel. Verder laat een lift grafiek zien dat het goed mogelijk is een groep te voorspellen waarin de kans op uitval hoog is, wat mogelijkheden biedt voor het beter sturen van begeleiding.

De interpretatie van de resultaten in de regressietaak is lastig. Wel kan worden geconcludeerd dat er een verbetering kan worden gerealiseerd in vergelijking met de simpele regel; dit wijst erop dat de modellen een klein vermogen hebben een inschatting te maken van wie langer over een opleiding zal doen. Wanneer echter de voorspelde waarde wordt afgezet tegen de werkelijke waarde blijkt dat deze verbetering minimaal is.

Bij zowel de classificatie- als de regressietaak presteren zogenaamde ensemble methodes het best. Vooral boosting algoritmes, algoritmes die gebaseerd zijn op het iteratief opstellen van kleine modellen om tot een uiteindelijk model te komen, presteren consistent goed.

Er zijn meerdere modellen opgesteld om de toegevoegde waarde van verschillende variabelen te onderzoeken. Hieruit blijkt dat het wisselgedrag en toetsresultaten nauwelijks een verbetering opleveren. Bij het gebruiken van de vooropleiding is een kleine, maar niet significante, verbetering op te merken. Verder is uit de opgestelde modellen het belang van variabelen afgeleid. Hieruit blijkt dat de vooropleiding variabelen van belang zijn in het model dat hiervan gebruik maakt. Verder zijn zowel opleidings- als leerlinggegevens van belang in de opgestelde modellen en blijkt dat de variabelen afgeleid van de toetsresultaten en ook de variabelen uit open data bronnen vrij weinig bijdragen leveren in de opgestelde modellen.

9. Discussie & verder onderzoek

In dit hoofdstuk wordt een discussie gegeven over het verrichte werk en worden een aantal aanbevelingen gedaan voor verder onderzoek.

Data ligt aan de kern van ieder Machine Learning onderzoek. In het hoofdstuk dat hieraan gewijd wordt, worden enkele problemen aangekaart. Machine Learning heeft (vrijwel) altijd baat bij meer gegevens. Eventueel kunnen gegevens waarover literatuuronderzoek uitwijst dat zij mogelijk voorspellende waarde bevatten worden toegevoegd, zoals het aantal woonplaatsen, informatie over delinquent gedrag, gegevens over de ouders en de uitslag van een persoonlijkheid test. Verder zouden volledigere gegevens over begeleiding (en wanneer deze aangeboden wordt) en aanwezigheid de mogelijkheid bieden hier meer onderzoek over te doen; deze twee soorten gegeven zijn in het huidige onderzoek niet gebruikt in de opgestelde modellen, vanwege de beperkte (vorm van) beschikbaarheid. Een dataset met de aanwezigheid per leerling per week is aanwezig, een vorm waar mee gewerkt kan worden, echter beslaan de gegevens enkel het afgelopen jaar waardoor het merendeel nog bezig is en deze data niet kan worden gebruikt in de voorspelmodellen.

Verder wordt met het toevoegen van vooropleidingsgegevens een kleine verbetering gerealiseerd; wellicht dat hier meer en gedetailleerdere informatie een verder positief effect heeft op de resultaten. Uit het literatuuronderzoek blijkt dat resultaten over de tijd bij voorspellingen op een lager niveau (zoals een vak) goede voorspellers kunnen zijn waarbij een trend maar ook uitschieters van belang zijn. De toegevoegde waarde van de resultaten in dit onderzoek reflecteren dit echter niet: het toevoegen van de toetsresultaten laat slechts een kleine, niet significante, verbetering zien.

Bij een momentopname is een patroon in toetsresultaten lastig vast te leggen. Dit komt deels door de beperkte aanwezigheid en verder doordat een verschillend aantal resultaten per leerling beschikbaar is. Het is moeilijk te zeggen hoe een variabele gevormd kan worden dat juist op het moment van een opname een patroon vastlegt dat een indicator kan zijn voor uitval, wat vervolgens nog geleerd moet worden door een algemeen voorspelmodel. Overigens is het ook goed mogelijk dat meer toetsresultaten met de huidige variabelen betere resultaten opleveren.

Ook kan het mogelijk zijn dat een gerichtere aanpak nodig is. Zo zal het bij een aantal opleidingen zo zijn dat er bepaalde toetsen of kerntaken een cruciaal punt in de opleiding zijn en zal het verder van belang zijn om te weten waar een leerling zich in het leertraject bevindt. Het is lastig in te zien hoe dit soort informatie goed tot zijn recht zou kunnen komen in een algemeen voorspelmodel.

In dit onderzoek is de nadruk geldt op de classificatietaak van het voorspellen of een leerling zal uitstromen met of zonder diploma. Hiervoor is gekozen omdat hierbij een duidelijke interpretatie van de resultaten mogelijk is en biedt deze voorspelling meer mogelijke toepassingen. Verder vertoonde de data die gaat over de duur een aantal eigenaardigheden waar de regressiemodellen moeite mee hebben, zoals het afstuderen binnen een maand van een voltijdleerling. Wellicht is het mogelijk om de regressietaak op een andere manier vorm te geven zoals het voorspellen of een leerling doorstroomt naar een volgend jaar, of door de regressie om te zetten naar een classificatietaak. Een wellicht nog andere interessante invalshoek is het meenemen van de ene voorspelling (wordt een diploma behaald) in de andere voorspelling (duur).

Bibliografie

Annika Wolff and Zdenek Zdrahal. Improving Retention by Identifying and Supporting "At-Risk" Students, July 2012. URL http://er.educause.edu/articles/2012/7/improving-retention-by-identifying-and-supporting-atrisk-students. 12

Mariska de Baat. Wat werkt bij het voorkomen en verminderen van schoolverzuim?, 2010. 9

J Martin Bland and Douglas G Altman. Multiple significance tests: the bonferroni method. *Bmj*, 310(6973):170, 1995. 37

Mihaljevic Bojan, Bielza Concha, and Larranaga Pedro. *bnclassify: Learning Discrete Bayesian Network Classifiers from Data*, 2014. URL http://github.com/bmihaljevic/bnclassify. R package version 0.3.2. 39

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A:1010933404324. 30

Stef Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(3), 2011. 33

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006. 26

A.D. Chapman. *Principles of Data Quality*. GBIF Secretariat, 2005. ISBN 978-87-92020-03-1. 21

Jie Cheng and Russell Greiner. Comparing bayesian network classifiers. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 101–108. Morgan Kaufmann Publishers Inc., 1999. 26

Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. *Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*, pages 5–12, 2008. 11

Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000. 27

Gerben W Dekker, Mykola Pechenizkiy, and Jan M Vleeshouwers. Predicting students drop out: A case study. *International Working Group on Educational Data Mining*, 2009. 11

Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998. 33, 36

Andrea Frankowski, Martijn van der Steen, Albert Meijer, and Mark van Twist. *De Publieke Waarde(n) van Open Data*. nsob, January 2015. ISBN 978-90-75297-50-8. URL http://www.nsob.nl/wp-content/uploads/NSOB-15-17-DT_Publieke-Waarde-web-DEF.pdf. 14

Yoav Freund and Robert Schapire. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999. 30

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001. 31

Andrew Gelman and Yu-Sung Su. *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*, 2015. URL https://CRAN.R-project.org/package=arm. R package version 1.8-6. 39

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003. 32

Peter G.M. van der Heijden, David J. Hessen, and Theo Wubbels. Studiesucces of -falen van eerstejaarsstudenten voorspellen: een nieuwe aanpak, 2010. 10

Serge Herzog. Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New Directions for Institutional Research*, 2006(131):17–33, 2006. 8, 11

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006. 28, 39

Inspectie van het onderwijs. De staat van het onderwijs Onderwijsverslag 2013/2014, 2013. 6

Nathalie Japkowicz and Mohak Shah. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011. 33

George H John, Ron Kohavi, Karl Pfleger, et al. Irrelevant features and the subset selection problem. In *Machine learning: proceedings of the eleventh international conference*, pages 121–129, 1994. 32

Floris Rutger Kappe. *Determinants of success: a longitudinal study in higher professional education*. Vrije Universiteit, 2011. 7

Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004. URL http://www.jstatsoft.org/v11/i09/. 39

HyunJi Kim. *discretization: Data preprocessing, discretization for classification.*, 2012. URL https://CRAN.R-project.org/package=discretization. R package version 1.0-1. 39

S. Kotsiantis, C. Pierrakeas, and P. Pintelas. Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, 18(5):411–426, 2004. doi: 10.1080/08839510490442058. URL http://dx.doi.org/10.1080/08839510490442058. 12

Max Kuhn, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, and Luca Scrucca. *caret: Classification and Regression Training*, 2015. URL http://CRAN.R-project.org/package=caret. R package version 6.0-41. 39

David M Lane. Online statistics - analysis of variance, 2016. URL http://onlinestatbook.com/2/analysis_of_variance/anova.pdf. Geraadpleegd op 2 juni 2016. 36

Namgil Lee and Jong-Min Kim. Conversion of categorical variables into numerical variables via bayesian network classifiers for binary classifications. *Computational Statistics & Data Analysis*, 54(5):1247–1265, 2010. URL http://EconPapers.repec.org/RePEc:eee:csdana:v:54:y:2010:i:5:p:1247-1265. 32

David D. Lewis. *Naive (Bayes) at forty: The independence assumption in information retrieval*, pages 4–15. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998. ISBN 978-3-540-69781-7. doi: 10.1007/BFb0026666. URL http://dx.doi.org/10.1007/BFb0026666. 25

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3): 18–22, 2002. URL http://CRAN.R-project.org/doc/Rnews/. 39

Ilia Mitov, Krassimira Ivanova, Krassimir Markov, Vitalii Velychko, Peter Stanchev, and Koen Vanhoof. Comparison of discretization methods for preprocessing data for pyramidal growing network classification method. *New trends in intelligent technologies, sofia*, pages 31–39, 2009. 32

Noorderpoort, Kennisnet, saMBO ICT, and MBO15. Big Data, van hype naar actie: Op zoek naar waardevolle inzichten voor het vergroten van studiesucces, 2013. 7, 8

The Open Knowledge Foundation. The Open Data Handbook, 2016. URL http://opendatahandbook.org/. Geraadpleegd op 4 maart 2016. 14

Frank Pijpers. Wat beïnvloedt het risico op voortijdig schoolverlaten? - Een multivariate analyse, 2010. 8

Leo L Pipino, Yang W Lee, and Richard Y Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002. 21

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/. 39

Els van Rooij, Jessica Pass, and Anja van den Broek. Geruisloos uit het onderwijs - Het verschil tussen klassieke en geruisloze risicofactoren van voortijdig schoolverlaten, 2010. 9

Steven L Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data mining and knowledge discovery*, 1(3):317–328, 1997. 35

George Siemens and Ryan SJ d Baker. Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 252–254. ACM, 2012. URL http://dl.acm.org/citation.cfm?id=2330661. 10

Jon Starkweather. Bayesian generalized linear models in r. *Benchmarks RSS Matters*, 2011. 29

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1): 1–21, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-25. URL http://dx.doi.org/10.1186/1471-2105-8-25. 30, 31

Diane M. Strong, Yang W. Lee, and Richard Y. Wang. Data Quality in Context. *Commun. ACM*, 40(5):103–110, May 1997. ISSN 0001-0782. doi: 10.1145/253769.253804. URL http://doi.acm.org/10.1145/253769.253804. 21

Terry Therneau, Beth Atkinson, and Brian Ripley. *rpart: Recursive Partitioning and Regression Trees*, 2015. URL http://CRAN.R-project.org/package=rpart. R package version 4.1-9. 39

Tanja Traag, Olivier Marie, and Rofl van der Velden. Risicofactoren voor voortijdig schoolverlaten en jeugdcriminaliteit, 2010. 8

Claus Weihs, Uwe Ligges, Karsten Luebke, and Nils Raabe. klar analyzing german business cycles. In D. Baier, R. Decker, and L. Schmidt-Thieme, editors, *Data Analysis and Decision Support*, pages 335–343, Berlin, 2005. Springer-Verlag. 39

Hadley Wickham and Romain Francois. *dplyr: A Grammar of Data Manipulation*, 2015. URL https://CRAN.R-project.org/package=dplyr. R package version 0.4.3. 39

Barbara van Wijk, Erik Fleur, and Sandra van den Dungen. Over reguliere wegen, hobbelige sporen en hinkelpaden, 2012. 9

Greg Ridgeway with contributions from others. *gbm: Generalized Boosted Regression Models*, 2015. URL `https://CRAN.R-project.org/package=gbm`. R package version 2.1.1. 39

Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005. 28

Rick Wolff, Sara Rezai, and Sabine Severiens. Het is maar in wat voor gezin je geboren bent... Een onderzoek naar de effecten van het opleidingsniveau ouders en etnische afkomst op studiesucces in het hoger onderwijs. In opdracht van het Ministerie OCW, 2010. 8

Nayyar A Zaidi, Jesus Cerquides, Mark J Carman, and Geoffrey I Webb. Alleviating naive bayes attribute independence assumption by attribute weighting. *The Journal of Machine Learning Research*, 14(1):1947–1988, 2013. 26, 27

Appendices

A. Data

Deze appendix geeft een overzicht van de beschikbare variabelen in de verschillende databestanden die zijn gebruikt voor dit onderzoek.

A.1 Interne Data

A.1.1 Leerlinggegevens export 20 Juli 2015

Deelnemernummer	Code reden uitschrijving {eerste/laatste} inschrijving
Geslacht	Omschrijving reden uitschrijving {eerste/laatste} inschrijving
Postcode	Laatste inschrijving cursusjaar verlaten instelling
Gemeente	Leeftijd op verlaten instelling
Deelnemer Nationaliteit	Inschrijvingen. aantal verschillende crebo codes
Deelnemer Geboorteland	Hoogste niveau diploma crebo
Vooropleiding code	Hoogste niveau diploma opleiding naam
Vooropleiding naam	Hoogste niveau diploma
Vooropleiding diploma behaald	Startkwalificatie behaald bij de instelling
Vooropleiding categorie nummer	VSV-er in convenantjaar
Vooropleiding categorie naam	VSV-er voor crebo
Vooropleiding startkwalificatie behaald	VSV-er voor opleiding naam
Voorropleiding toeleverende school	VSV-er voor niveau
Vooropleiding datum uitschrijving	Presentie. periode start
Cursusjaar instroom	Presentie. periode einde
Kalendermaand instroom	Presentie. percentage aanwezig
Leeftijd op datum instroom	Presentie. percentage geoorloofd afwezig
Leeftijd op versiedatum	Presentie. aantal lessen
Eerste inschrijving is gelijk aan laatste inschrijving	Presentie. aantal niet geregistreerde lessen
Opleiding met verzwaarde intake	Presentie. percentage lesuitval
{Eerste/Laatste} inschrijving. verbintenis ID	Bekostiging. eerste bekostigingsjaar
{Eerste/Laatste} inschrijving. opleiding code	Bekostiging. laatste bekostigingsjaar
{Eerste/Laatste} inschrijving crebo	Bekostiging. aantal keren bekostigd
{Eerste/Laatste} inschrijving naam opleiding	Bekostiging. onderbreking in bekostigingsjaren
{Eerste/Laatste} inschrijving niveau	Begeleiding. aantal gesprekken mbt verzuim
{Eerste/Laatste} inschrijving leerweg	Begeleiding. aantal gesprekken mbt studievoortgang
{Eerste/Laatste} inschrijving. geplande einddatum	Begeleiding. traject bij Servicecentrum
{Eerste/Laatste} inschrijving. datum uitschrijving	Begeleiding. traject bij Zorgteam
{Eerste/Laatste} inschrijving. opleiding duur in maanden	Deelnemer. aantal bijzonderheden (zoals
{Eerste/Laatste} inschrijving. SBU	handicap. ADHD etc.)

Tabel A.1: Velden in de export van 20 Juli 2015

A.1.2 Verbintenissen export 17 September 2015

Deelnemer nr.	Verbintenis Einddatum
Crebo	Reden uitschr. Code
Opleiding Naam	Reden uitschr. Naam
Opleiding Leerweg	Interne Organisatie - Sector
Niveau	Interne Organisatie - Team
Intensiteit	Interne Organisatie - Opleidingscluster
Opleiding Duur in maanden	Crebokoppel Hoofdgroep
Verbintenis cohort	Crebokoppel Subgroep
Verbintenis Begindatum	Crebokoppel Beroepsopleiding
Verbintenis Geplande einddatum	

Tabel A.2: Velden in de export van 17 September 2015

A.1.3 Aanwezigheid export 18 Februari 2015

Deelnemernummer	Vak
Schooljaar	# min. aanwezig
Kalenderjaar	# min. afwezig
Week	# min. geoorloofd afwezig
Sector	# min. lesaanbod
Team	

Tabel A.3: Velden in de export van 18 Februari 2016

A.1.4 Testresultaten export 17 September 2015

Variabelen beschikbaar voor alle testresultaten			
Deelnemernummer	Locatie	Sector	Klas
Datum	Label	Gereed	
Niveau, percentage beschikbaar voor onderdelen per enkele test			
Reken	**Taal**	**Taal vooraf**	**LLA**
Getallen	Lezen	Woordbeelden	Luisteren
Verhoudingen	Spelling	Woordvorming	Lezen
Meten & meetkunde	Stijl	Werkwoordspelling	Woorden
Verbanden	Woordkennis	Woordbetekenis	Structuur
		Leesvaardigheid	

Tabel A.4: Velden in de export van 17 September 2015

A.1.5 Testresultaten export 13 April 2016

Deelnemernummer	Resultaat ID
Resultaat	Pogingsnr.
Aanmaakdatum	Wijzigingsdatum
Toetscode	Toetsnaam
Toets vrijstelling	Onderwijsproductcode
Toets notitie	Mederwerker naam
Type	Vergeven pogingen
Verbintenis Status	Verbintenis Id Opleiding Naam

Tabel A.5: Velden in de export van 13 April 2016

A.2 Externe/Open Data

A.2.1 Postcodedata.nl

ID	Postcode
Pnum	Pchar
Minnumber (huisnummer)	Maxnumber (huisnummer)
Numbertype (huisnummer)	Straat
Stad	Stad ID
Gemeente	Gemeente ID
Provincie	Provincie Code
Lengtegraad (LAT)	Breedtegraad (LON)
RD x	RD y
Locatie detail	Veranderdatum

Tabel A.6: Velden in het `www.postcodedata.nl/` bestand

A.2.2 DUO Crebokoppeltabel

Crebo Code	Volgnummer SSB
Domein mbo	Startdatum kwalificatie
Uiterste instroomdatum	Einddatum kwalificatie
Hoofdgroep SSB	Subgroep SSB
Beroepsopleiding SSB	Beroep SSB
Kwalificatiedossier	Kwalificatie
Kwalificatieniveau	Leerweg
Studiebelastingsuren	Prijsfactor
BRIN nummer kenniscentrum	Mbo sector

Tabel A.7: Velden in de DUO Crebokoppeltabel

A.2.3 SES Statusscores

Variabele	Beschikbaar jaargangen
Gemeente nummer cbs	98/02/05/10/14
Gemeente naam cbs	98/02/06/10/14
Factorscore sociale status	98/02/06/10/14
Rangorde sociale status	98/02/06/10/14
Aantal huishouden	98/02/06/10/14
Bevolking totaal	97/01/05/10/14
Indicatie naam postcodegebied	02/06/10/14
Postcode-4	

Tabel A.8: Velden in het SES Statusscores bestand

A.3 Voorspelmodel data

	Naam	Open	Groep	Type	Range/Mode/% True
1	Niveau	Nee	opleiding	numeriek	1-4
2	SBU	Ja	opleiding	numeriek	1600-6400
3	Prijsfactor	Ja	opleiding	numeriek	1-1.8
4	OpleidingLeerweg	Nee	opleiding	categorisch	BOL
5	Verzwaardeintake	Nee	opleiding	logisch	3.3%
6	Beroepsopleidingnaam	Nee	opleiding	categorisch	Entree
7	Intensiteit	Nee	opleiding	categorisch	Voltijd
8	OpleidingDuurinmaanden	Nee	opleiding	numeriek	12-48
9	InterneOrganisatieSector	Nee	opleiding	categorisch	Economie & Ondernemerschap
10	InterneOrganisatieTeam	Nee	opleiding	categorisch	Retail & Logistiek
11	InterneOrganisatieOpleidingscluster	Nee	opleiding	categorisch	Entree (open aanbod)
12	CrebokoppelHoofdgroep	Nee	opleiding	categorisch	Zorg en welzijn
13	CrebokoppelSubgroep	Nee	opleiding	categorisch	AKA
14	Kenniscentrum	Ja	opleiding	categorisch	ECABO
15	Sector	Ja	opleiding	categorisch	economie
16	Studieintensiteit	Ja	opleiding	numeriek	1.5-152.381
17	Grootteopleiding	Nee	opleiding	numeriek	1-376
18	Grootteteam	Nee	opleiding	numeriek	8-496
19	Percentageongeoorloofd	Nee	opleiding	numeriek	0.369-0.762
20	GemAanwezigheidpercentageTeam	Nee	opleiding	numeriek	0.299-0.705
21	BegroteLoonkostenFTETeam	Nee	opleiding	numeriek	67711.59-78755.5
22	Geplandeduur	Nee	opleiding	numeriek	42-1634
23	Lat	Ja	leerlinggebonden	numeriek	50.823-53.389
24	Lon	Ja	leerlinggebonden	numeriek	3.488-7.015
25	Statusscore	Ja	leerlinggebonden	numeriek	-5.24-2.532
26	Rangordestatus	Ja	leerlinggebonden	numeriek	26-3538
27	AfstandDVC	Ja	leerlinggebonden	numeriek	1-228243.249
28	Geslacht	Nee	leerlinggebonden	categorisch	Man
29	Gemeente	Nee	leerlinggebonden	categorisch	Dordrecht
30	DeelnemerNationaliteit	Nee	leerlinggebonden	categorisch	Nederlandse
31	DeelnemerGeboorteland	Nee	leerlinggebonden	categorisch	Nederland
32	Postcode	Nee	leerlinggebonden	categorisch	Overig
33	VorigNiveau	Nee	leerlinggebonden	numeriek	0-4
34	Deelnemerbijzonderheden	Nee	leerlinggebonden	numeriek	0-4
35	Stad	Nee	leerlinggebonden	categorisch	Dordrecht
36	Leeftijdopdatuminstroom	Nee	leerlinggebonden	numeriek	15-62
37	Gemsucces	Nee	gemiddeld succes	numeriek	0-1
38	Gemduur	Nee	gemiddeld succes	numeriek	222.057-1140.429
39	Crebos	Nee	1/2j emperisch	numeriek	1-3
40	Internesectors	Nee	1/2j emperisch	numeriek	1-2
41	Interneteams	Nee	1/2j emperisch	numeriek	1-3
42	Interneclusters	Nee	1/2j emperisch	numeriek	1-3
43	Crebohoofd	Nee	1/2j emperisch	numeriek	1-3
44	Crebosub	Nee	1/2j emperisch	numeriek	1-3
45	Creboberoeps	Nee	1/2j emperisch	numeriek	1-3
46	Overeenkomsten	Nee	1/2j emperisch	numeriek	1-4
47	Minresultaat	Nee	1/2j resultaten	numeriek	0-10
48	Maxresultaat	Nee	1/2j resultaten	numeriek	0-10
49	Sdresultaat	Nee	1/2j resultaten	numeriek	0-4.677
50	Gemiddeldresultaat	Nee	1/2j resultaten	numeriek	0-10
51	Gemiddeldbovengem	Nee	1/2j resultaten	numeriek	-6.453-3.75
52	Voldoendepercentage	Nee	1/2j resultaten	numeriek	0-1
53	Toetsen	Nee	1/2j resultaten	numeriek	1-51
54	Extrapogingen	Nee	1/2j resultaten	numeriek	0-6
55	Resultaatfreq	Nee	1/2j resultaten	numeriek	0.001-0.071
56	Extrapogingenpercentage	Nee	1/2j resultaten	numeriek	0-1
57	Percentageaanwezig	Nee	1/2j presentie	numeriek	0-0.947
58	Percentagelesuitval	Nee	1/2j presentie	numeriek	0-1
59	Percentagegeoorloofdafwezig	Nee	1/2j presentie	numeriek	0-1
60	Gesprekkenmbtstudievoortgang	Nee	begeleiding	numeriek	0-36
61	Gesprekkenmbtverzuim	Nee	begeleiding	numeriek	0-54
62	TrajectbijServicecentrum	Nee	begeleiding	categorisch	Nee
63	TrajectbijZorgteam	Nee	begeleiding	categorisch	Nee
64	Studievoortgangintensiteit	Nee	begeleiding	numeriek	0-0.172
65	Verzuimintensiteit	Nee	begeleiding	numeriek	0-0.074
66	Vooropleidingcode	Nee	vooropleiding	categorisch	VMBOTLEC
67	Vooropleidingdiplomabehaald	Nee	vooropleiding	logisch	61.5%
68	Vooropleidingcategorienummer	Nee	vooropleiding	categorisch	4
69	Vooropleidingstartkwalificatie	Nee	vooropleiding	logisch	8.9%
70	Voorropleidingschool	Nee	vooropleiding	categorisch	Wellantcollege
71	Vooropleidingtussentijd	Nee	vooropleiding	numeriek	0-16865

Tabel A.9: Data zoals gebruikt in voorspelmodel

B. Bekostiging mbo

In deze appendix zal is uitgewerkt hoe de bekostiging voor de mbo instellingen in elkaar steekt. Mbo instelling ontvangen van het rijk een bijdrage (lumpsum) voor de exploitatiekosten en de huisvestigingskosten voor een kalender jaar. De rijksbijdrage bestaat uit twee delen: voor de entreeopleiding en voor de overige opleidingen, de totale rijksbijdrage is de som.

Rijksbijdrage entreeopleiding

Het rijksbijdragedeel voor de entreeopleiding wordt berekend volgens de formule:

$$\frac{\text{IDW}}{\text{LDW}} \cdot \text{LB}, \text{ waarbij:}$$

IDW: de deelnemerswaarde voor de entreeopleiding, afgerond op 2 decimalen
LDW: de landelijke deelnemerswaarde voor de entreeopleiding, deze is gelijk aan de som van de deelnemerswaarden voor de entreeopleiding van alle instellingen
LB: het landelijk beschikbare budget voor de entreeopleiding.
De deelnemerswaarde wordt berekend volgens de formule:

$$\text{IDW} = \sum[(\text{Dbbl} \cdot 0.5 \cdot \text{PF} \cdot \text{Vf}) + (\text{Dbol} \cdot \text{PF} \cdot \text{Vf})] \cdot \text{Cf}$$

Dbbl: elke deelnemer die op 1 oktober van het tweede kalenderjaar voorafgaand aan het bekostigingsjaar is ingeschreven voor de entreeopleiding in de beroepsbegeleide leerweg.
Dbol: elke deelnemer die op 1 oktober van het tweede kalenderjaar voorafgaand aan het bekostigingsjaar is ingeschreven voor de entreeopleiding in de beroepsopleidende leerweg.
PF: de prijsfactor voor de opleiding waarvoor de deelnemer is ingeschreven, deze is per opleiding vastgesteld.
Vf: de toegekende factor voor het verblijfsjaar in de entreeopleiding. Deze bedraagt voor het eerste verblijfsjaar 1.2, het tweede 0.6 en het derde en volgende verblijfsjaren 0.
Cf: de correctiefactor tweede teldatum voor de entreeopleiding. Met deze factor wordt de bekostiging gecorrigeerd voor de gevolgen van het verschil tussen het aantal deelnemers op 1 oktober t-2 en op 1 februari t-1.

Bij de vaststelling van de correctiefactor wordt geen rekening gehouden met de prijsfactoren van de opleidingen en wordt in de weging alleen onderscheid gemaakt tussen voltijd en deeltijd.

$$\text{Cf} = \frac{[\text{Dbbl1} \cdot 0.5 + \text{Dbol1}] + [\text{Dbbl2} \cdot 0.5 + \text{Dbol2}]}{2 \cdot [\text{Dbbl1} \cdot 0.5 + \text{Dbol1}]}$$

Rijksbijdrage overige opleidingen

Het rijksbijdragedeel voor voor de basisberoepsopleiding, vakopleiding, middenkaderopleiding en specialistenopleiding (hierna bvms) wordt berekend volgens de formule:

$$\frac{\text{IDW} + \text{IDiW}}{\text{LDW} + \text{LDiW}} \cdot \text{LB}, \text{ waarbij:}$$

IDW: de deelnemerswaarde voor de bvms van de instelling, afgerond op 2 decimalen
IDiW: de diplomawaarde, afgerond op 2 decimalen
LDW: de landelijke deelnemerswaarde voor de bvms, deze is gelijk aan de som van de deelnemerswaarden voor de bvms van alle instellingen
LDiW: de landelijke diplomawaarde, deze is gelijk aan de som van de diplomawaarde van alle instellingen
LB: het landelijk beschikbare budget voor de basisberoepsopleiding, vakopleiding, middenkaderopleiding en specialistenopleiding.

$$\text{IDW} = \sum[(\text{Dbbl} \cdot 0.4 \cdot \text{PF} \cdot \text{Vf}) + (\text{Dbol} \cdot \text{PF} \cdot \text{Vf})] \cdot 0.8 \cdot \text{Cf}$$

Dbbl/Dbol/PF: zie entreeopleiding
Vf: de aan de desbetreffende deelnemer toegekend factor voor het verblijfsjaar. Deze bedraagt:

- Het eerste jaar: 1.2
- Het tweede t/m vierde jaar: 1.0
- Het vijfde en zesde jaar: 0.5
- Het zevende en opvolgende jaren: 0

Cf: de correctiefactor tweede teldatum voor de entreeopleiding.

$$\text{Cf} = \frac{[\text{Dbbl1} \cdot 0.4 + \text{Dbol1}] + [\text{Dbbl2} \cdot 0.4 + \text{Dbol2}]}{2 \cdot [\text{Dbbl1} \cdot 0.4 + \text{Dbol1}]}$$

$$\text{IDiW} = \sum[(\text{D} \cdot \text{DiW} - \text{DiE}) + \text{DS}] \cdot 0.2$$

D: elke deelnemer die in het tweede kalenderjaar voorafgaand aan het bekostigingsjaar een dimploma van een basisberoepsopleiding, een vakopleiding of een middenkaderopleiding heeft behaald
DiW: de diplomawaarde, bedraagt voor een basisberoepsopleiding 1, een vakopleiding 3 en een middenkaderopleiding 5.
DiE: DiW van het hoogste door de deelnemer eerder behaalde diploma van een basisberoepsopleiding, een vakopleiding of een middenkaderopleiding
DS: de diplomawaarde voor een specialistenopleiding bedraagt 2 voor elke deelnemer die een diploma van een specialistenopleiding heeft behaald in het tweede kalenderjaar voorafgaand aan het bekostigingsjaar, en niet eerder een diploma van een specialistenopleiding heeft behaald.