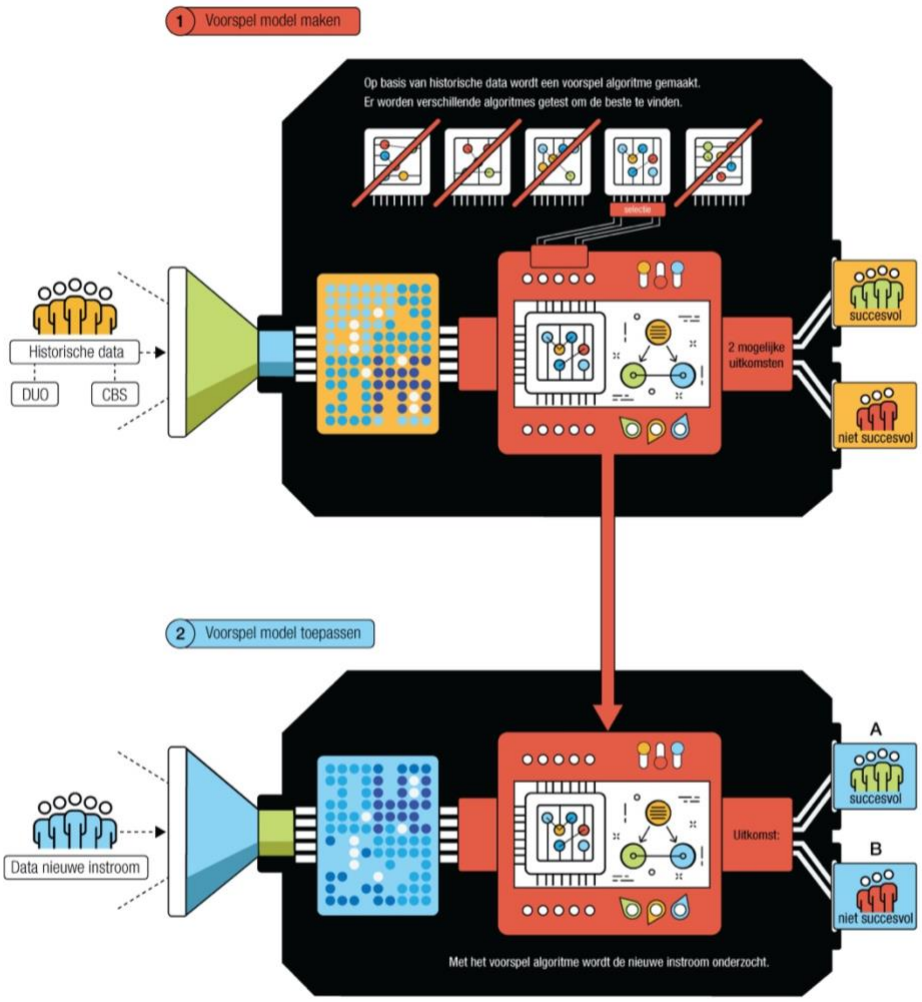
Datacoalitie Datagedreven onderzoek mbo

Communicatie en verantwoordingsdocument bij het Voorspelmodel n2-student (2021-variant)



Naar een datagedreven aanpak voor het verhogen

van het succes van niveau 2 studenten

Juni 2021

1. Inleiding

In de 'Staat van het middelbaar beroepsonderwijs 2021' schrijft de Inspectie van het Onderwijs dat "*voor studenten in een kwetsbare positie aandacht vereist blijft. Dat geldt vooral voor studenten die een niveau 2-opleiding volgen. Niet alleen omdat hun positie op de arbeidsmarkt kwetsbaar is, maar ook omdat ze, als ze opstromen naar een hoger opleidingsniveau dat meer kans biedt op een stevige positie op de arbeidsmarkt, in veel gevallen het diploma niet halen. Daarnaast treft de coronacrisis in grotere mate ongediplomeerden en de studenten op niveau 2. De kwaliteitsagenda's van instellingen die vaak speerpunten bevatten gericht op deze doelgroep, zijn door de coronapandemie onder druk komen te staan. Belangrijk is dat er zo snel mogelijk weer ruimte komt om de focus te richten op de lange termijn doelstellingen".*

Binnen het MBO zijn de niveau 2 studenten een groep waarbij er sprake is van een hoog uitvalpercentage. Dit ondanks de inspanningen van instellingen om uitval te verminderen en het succes van deze studenten te vergroten. Het is van groot belang juist deze doelgroep studenten extra te begeleiden naar een passend diploma zodat zij tenminste een startkwalificatie op zak krijgen. Het succes van een niveau 2 student zit hem niet alleen in het behalen van een diploma (startkwalificatie), maar gaat ook over opstroom naar een hoger niveau. De Datacoalitie Datagedreven Onderzoek MBO wil met datagedreven onderzoek een bijdrage leveren aan het verbeteren van de kwaliteit van de leerloopbaan van de niveau 2 student en daarmee zijn succes vergroten

Bedenken, ontwerpen en ontwikkelen van een voorspelmodel voor niveau 2 studenten

In de periode januari tot juni 2020 heeft de datacoalitie onderzoek gedaan naar "welke student-/ opleidings-/ omgevingskenmerken relevant zijn om de loopbaan van een niveau 2 student te voorspellen?" De zoektocht ging uit naar een kenmerk of combinatie van verschillende kenmerken om de kans in te schatten voor studenten om te diplomeren of opstromen enerzijds, of afstromen of uitvallen anderzijds.

Er is in dat onderzoek gekeken naar studenten die in de afgelopen studiejaren een niveau 2 opleiding aan het mbo hebben gevolgd. Van deze studenten is hun leerloopbaan inzichtelijk én zijn diverse kenmerken beschikbaar (bij DUO). Er zijn kenmerken verzameld zoals opleidingsgegevens, vooropleiding- en demografische gegevens. Een belangrijk inzicht was dat op basis van historische gegevens gedragspatronen kunnen worden ontdekt. De onderzoeksgroep heeft daar in 2020 een hele belangrijke stap in gezet. Het voorspelmodel kon op dat moment gebruikt worden om voor nieuwe niveau 2 studenten hun kans op succes, of risico op afstroom/ uitval, te voorspellen. De inzichten die het voorspelmodel oplevert kunnen door de onderwijsinstelling toegevoegd worden aan de set van informatie die onderwijsinstellingen gebruiken om te bepalen welke groepen studenten mogelijk extra hulp en begeleiding nodig hebben. Op deze manier draag het voorspelmodel bij aan gelijke kansen in het mbo, en aan meer (scherper gerichte) aandacht voor kwetsbare studenten in het mbo. Op een manier die bijdraagt aan een effectieve en efficiënte inzet van middelen in het onderwijs.

2. Opdracht: Naar een praktisch toepasbaar instrument voor onderwijsinstelling én onderwijsteam

De stuurgroep van de datacoalitie, bestaande uit bestuurlijke vertegenwoordigers van de deelnemende onderwijsinstellingen en OCW en DUO, hebben de werkgroep de opdracht gegeven om de praktische toepasbaarheid van het voorspelmodel voor de niveau 2 student nader te onderzoeken en uit te werken tot praktisch instrument voor onderwijsinstelling én onderwijsteam.

Voorliggend document is het communicatie- en verantwoordingsdocument bij het Voorspelmodel n2-student (2021-variant). In dit document wordt uitgelegd wat het voorspelmodel is, hoe het is ontwikkeld, hoe het kan worden toegepast en hoe oneigenlijk gebruik kan worden voorkomen en ethisch verantwoord gebruik wordt geborgd. Het document is bedoeld voor alle onderwijsinstellingen in de mbo-sector die gebruik willen maken van het ontwikkelde voorspelmodel.

Wie vormen de 'Datacoalitie datagedreven onderzoek mbo'?
Sinds mei 2019 werkt een datacoalitie van onderwijsinstellingen[1] nauw samen met OCW en DUO bij het onderzoeken, verkennen en exploreren van datagedreven werken in het mbo. Door pilots, experimenten en onderzoeken te initiëren worden 'hands-on' stappen gezet. Alle opgedane kennis, ervaringen, inzichten en ontwikkelde producten worden transparant gedeeld met de hele mbo sector.

De Datacoalitie DGO bestaat uit: Noorderpoort, ROC van Twente, Curio, ROC Friese Poort, ROC van Amsterdam-Flevoland, ROC Nijmegen, Gilde Opleidingen, Deltion College, Koning Willem 1 College en Zadkine

3. Het voorspelmodel

Het ontwikkelde voorspelmodel voorspelt de kans op uitval van studenten die instromen in een niveau 2 opleiding. Het ontwikkelde model is een 'ensemble' van beslisbomen dat geoptimaliseerd is door middel van een 'Gradient Boosting'-techniek. Dergelijke technieken zijn wijd verspreid en worden in een grote diversiteit aan eigentijdse applicaties gebruikt, bijvoorbeeld in klantretentiesystemen om potentieel afhakende klanten te identificeren, om advertentiecampagnes effectief te richten, of om het effect van medicatie te voorspellen.

In het geval van het ontwikkelde voorspelmodel is de data van niveau 2 studenten gelabeld met de labels 'uitval' en 'geen-uitval'. De achtergrondkenmerken van de studenten zijn gebruikt als input in het model. Vervolgens is het model getraind met historische data. Deze historische data bestaan uit alle achtergrond kenmerken en 'succes in het mbo op niveau 2' van (oud)studenten vanaf de cohorten 2015 t/m 2019. Deze data wordt gebruikt om de kans op uitval te bepalen van de studenten die 'aankomend jaar' instromen in een mbo niveau 2 opleiding.

Tijdens de training van het model is de uitkomst van elke voorspelling gebruikt om het voorspelmodel continu net wat fijner af te stellen. Dit doet het model door systematisch combinaties van achtergrondkenmerken te zoeken die samen optimaal samenhangen met de te voorspellen uitkomst.

Op dit moment kan het model de uitval van studenten die instromen in een niveau 2 opleiding voorspellen met een accuraatheid van 81%. In het algemeen is de accuraatheid groter bij een grote doelgroep. De kans op een verkeerde voorspelling is dus groter bij kleine, zeer specifieke, groepen. Bijvoorbeeld studenten met een relatief hoge leeftijd (40+ jaar) bij het starten van de opleiding. Het voorspelmodel is ook instaat om aan te geven wat de meest belangrijke indicatoren zijn bij het onderscheiden van studenten die uitvallen en niet uitvallen. Echter het model bepaald niet de mate waarin elke kenmerk uitval verklaard. Dit is belangrijk bij het interpreteren van de resultaten. Het is goed mogelijk dat een student afkomstig is uit een armoedeprobleemaccumulatiegebied (APCG), wat een belangrijke onderscheidende indicator is, maar dat er in feite andere onderliggende of samenhangende factoren zijn die bijdragen aan zijn of haar uitval.

De datacoalitie wil er specifiek zorg voor dragen dat het ontwikkelde model rechtmatig en eerlijk te werk gaat. Concreet betekent dit dat er een algoritme is gebouwd dat niet onwenselijk discrimineert op basis van geslacht, leeftijd, geaardheid, levensbeschouwelijke- of politieke overtuiging of etniciteit. Een belangrijke stap om dat te waarborgen is genomen door het gebruik van de data, het algoritme en de toepassing daarvan te toetsen met behulp van een DPIA. Concrete aandachtspunten die hierin specifiek zijn doorgelicht zijn het gebruik van geslacht, leeftijd, en migratieachtergrond als inputvariabelen voor het voorspelmodel. Met het oog op de 'positief-discriminatoire toepassing' van het algoritme en de toegevoegde waarde van deze variabelen voor het model en voor een accurate voorspelling, hebben we het gebruik van deze variabelen verantwoord.

We beseffen echter dat om de eerlijkheid en de rechtmatigheid van de ontwikkelde algoritme te kunnen blijven garanderen het algoritme periodiek zal moeten worden geëvalueerd en ook de toepassing ervan kritisch moet worden gevolgd. We benadrukken het belang van een blijvende samenwerking tussen de ontwikkelaars en de gebruikers van dit algoritme om dit te waarborgen.

4. Toepassing van het voorspelmodel

Voor het bedenken en bepalen van het gebruiksdoel van het voorspelmodel zijn bij alle deelnemende instelling twee of meer interviews afgenomen. Hiervoor zijn met name opleidingsmanagers en docenten van niveau 2 opleidingen geïnterviewd. Uit de interviews met de opleidingsmanagers bleek een voorspelling van uitval per student gewenst. Op deze manier zou het beste maatwerk kunnen worden geleverd in de begeleiding van de studenten. Aangezien het model gebruik maakt van DUO-data over de leerloopbaan van de student én om de privacy van de studenten te waarborgen is een individuele voorspelling (op dit moment) niet mogelijk. Daarvoor in de plaats geeft het model een voorspelling van uitval per totale instroom van elke opleiding. Hierdoor is het model minder geschikt voor het begeleiden van individuele studenten maar kan wel gebruikt worden om aan het begin van het schooljaar een betrouwbare inschatting te maken van de kans op uitval per opleiding en onderwijsteams. Deze schatting van de kans op uitval kan gebruikt worden bij de allocatie van (benodigde of extra) middelen voor bijvoorbeeld studieloopbaan begeleiding. Met als doel de middelen in te zetten waar ze het meeste nodig zijn

De resultaten worden via een dashboard in het portaal van DUO gepubliceerd. In dit dashboard is het mogelijk om de geschatte kans op uitval en de belangrijkste kenmerken die bijdragen aan uitval per instelling, vestiging en opleidingsteam te vinden.

Het streven is dat de informatie jaarlijks beschikbaar gesteld wordt door DUO in de periode april/mei voor het komende studiejaar. Hier zit nog een logistieke uitdaging die verband houdt met het moment waarop DUO kan beschikken over data van nieuw ingestroomde niveau 2 studenten.

4.1 Interpretatie van de resultaten

Het dashboard presenteert het geschatte percentage uitval per opleiding van de ingestroomde studenten. Daarnaast presenteert het dashboard de belangrijkste factoren die bijdragen aan uitval. Het is goed mogelijk dat andere kenmerken die niet in het model zijn opgenomen ook belangrijk zijn. Echter vanzelfsprekend worden deze niet door model gepresenteerd. Het model kent deze kenmerken namelijk niet.

Het is belangrijk bij de interpretatie van de resultaten dat de gepresenteerde kenmerken die bijdragen aan de schatting van uitval niet persé uitvallen verklaren. De kenmerken zijn alleen het meest consistent aanwezig bij studenten die uitvallen. Kenmerken zoals het verzuim van de student, schoolklimaat, psychische problematiek, leermoeilijkheden en gezinssamenstelling zijn factoren die bijdragen aan uitval[1] maar zijn niet opgenomen in het model.

[1] J. Geubels, C.E. Van der Put, M. Assink. Risk Factors for School Absenteeism and Dropout: A Meta-Analytic Review. J Youth Adolesc. 2019; 48(9): 1637–1667

4.2 Informeren van betrokkenen

Bij het informeren van de betrokken is het belangrijk om te weten dat OCW de verwerkingsverantwoordelijke is en dat OCW de persoonsgegevens van de studenten beheert. De onderwijsinstellingen hebben geen toegang tot die specifieke set persoonsgegevens, omdat het gegevens betreft over de leerloopbaan van de student in het PO en VO. De onderwijsinstellingen hebben alleen inzicht in het verwacht aandeel uitval en de score op achtergrondkenmerken die bepalend zijn voor uitval. Een voorstel is om de studenten te informeren bij inschrijving bij de opleiding (bijvoorbeeld via het privacyreglement). In deze informatie kan staan dat informatie over de leerloopbaan van de student in het PO en VO gebruikt wordt om begeleiding van student te verbeteren en de kans op studiesucces van de student te vergroten. Ook moet duidelijk zijn dat de instelling geen toegang heeft tot de persoonsgegevens en dat deze gegevens statisch zijn beveiligd zodat privacy van de student wordt gewaarborgd. Bij een eventuele “uitrol” en implementatie van het voorspelmodel moet het informeren van de betrokkenen verder worden uitgewerkt in samenspraak met privacy- en communicatieadviseurs van de onderwijsinstellingen.

5. Oneigenlijk gebruik

Oneigenlijk gebruik van data is gebruik waarvoor de data niet is bedoeld. Natuurlijk worden data voor een bepaald doel verzameld en je mag deze data (onder de gestelde voorwaarden) gebruiken voor onderzoeken. Die onderzoeken hebben bepaalde doelen en oneigenlijk gebruik is daarmee dat data voor een ander doel dan voor het onderzoeksdoel worden gebruikt. Gebruik en doelen zijn dus aan elkaar verbonden.

Onder onderzoek wordt verstaan ***het verzamelen en/of analyseren van data om uitspraken te doen over een vooraf beschreven vooronderstelling of vraag.*** *Hieronder vallen dus bijvoorbeeld ook experimentele verkenningen van bestaande persoonsgegevens en Learning Analytics.*

Oneigenlijk gebruik van data ontstaat ook als het leidt tot uitsluiting, stigmatisering en kansenongelijkheid. Dus ook al worden de verzamelde data voor het beschreven onderzoeksdoel verzameld en gebruikt, als ze deze (neven)effecten hebben is er sprake van oneigenlijk gebruik.

5a Tegengaan van oneigenlijk gebruik van het voorspelmodel

Bij dataonderzoek in het onderwijs wordt (vrijwel) altijd gebruik gemaakt van persoonsgegevens. Het is daarom belangrijk om in de eerste plaats beleid te formuleren waarin de spelregels rond het doen van onderzoek met gebruik van persoonsgegevens wordt vastgelegd. Om oneigenlijk gebruik van data tegen te gaan is het voorts noodzakelijk dat er een Code of Practice of Ethiekkompas wordt opgesteld waarin is vastgelegd welke waarden de organisatie hanteert bij het verzamelen, onderzoeken en gebruiken van data. Ethische beginselen leggen vast wat “je van je zelf mag en wat niet”. Idealiter zijn die waarden gebaseerd op de kernwaarden van de schoolorganisatie en de sector. Oneigenlijk gebruik wordt ook tegengegaan door de rollen en bijbehorende rechten en verantwoordelijkheden van iedereen die met data werkt nauwkeurig te formuleren en vast te leggen.

In de ontwikkelde Proof-of-Concept wordt oneigenlijk gebruik tegen gegaan door migratieachtergrond niet mee te nemen als kenmerk in het dashboard. En ook worden de resultaten niet op individueel niveau in de output getoond, maar in een minimale groepsgrootte van vijf studenten. Dit om zo de statistisch beveiliging te waarborgen. Het dashboard is alleen toegankelijk met de juiste autorisatie en ook dan ziet de gebruiker alleen de data van zijn of haar onderwijsinstelling.

Verder wordt oneigenlijk gebruik tegengegaan door nauwkeurig te formuleren welke bewaartermijnen gelden voor data (1 jaar), door uit te gaan van data minimalisatie bij het doen van onderzoek, door doelbinding en door informatieplicht (verantwoord in de uitgevoerde DPIA).

5b borging van ethisch gebruik van data

Op het niveau van de onderwijsinstelling wordt geadviseerd om een ethiekkompas op te stellen waarin de waarden worden genoemd die de onderwijsinstelling hanteert bij het gebruik van data. Dit is het moreel kompas, het geweten van de school. De publicatie “Waarden Wegen” van Kennisnet kan hierbij van nut zijn en vooral ook de online tool die hierbij hoort.
https://wijzer.kennisnet.nl/ethiekkompas

Voor concrete data onderzoeken kan steeds een cyclus worden doorlopen waarin het uit te voeren onderzoek eerst minutieus wordt gescreend op ethische aspecten. Een zeer handzame, Nederlandstalige, methode is DEDA (De Ethische Data Assistent).

6. Voorstel vervolg

Op basis van de resultaten uit het onderzoek, de Proof-of-Concept en de opgedane inzichten uit de interviews binnen de instellingen, ziet de onderzoeksgroep twee richtingen voor vervolgonderzoek en -aanpak:

1. Model doorontwikkelen en dashboard verbreden naar niveau 3 en 4
2. Onderzoeken praktische implementatievragen

Ad 1: Model doorontwikkelen en dashboard verbreden naar niveau 3 en 4
De (door)ontwikkeling van het voorspelmodel is een continu proces. In de volgende fase kan onderzocht worden of de kwaliteit van de voorspelling verder verbeterd kan worden door middel van uitbreiding van de kenmerkenset. Bijvoorbeeld met (micro)data van het CBS. Daarnaast kan de scope van het voorspelmodel en dashboard verbreed worden naar de niveau 3 en niveau 4 student.

Ad 2: Onderzoeken praktische implementatievragen
Met het ontwikkelen van het dashboard bij het voorspelmodel is de Proof-of-Concept-fase afgerond. De volgende fase kan gericht worden op het onderzoeken van praktische implementatie-vraagstukken die ontstaan als het instrument beschikbaargesteld wordt aan de sector. In dit onderzoek worden praktijk pilots georganiseerd met besturen en onderwijsteams. Vraagstukken die in dit onderzoek o.a. aan bod komen:

- Data: Vertaling naar en rapportage op teamniveau (indelingen aanleveren? Of gebruik RIO?)
- Logistiek: vroeg in het jaar inzetten = vroeg in het jaar data hebben bij DUO
- Dashboard testen: Welke kenmerken willen de teams zien?
- Kennis: welke praktische kennis moet bestuur en onderwijsteam in huis hebben of halen?
- Kennis: benodigde handleiding en instructies voor gebruik
- Beveiliging en autorisatie: beschikbaar stellen van het dashboard via het DUO portaal
- Privacy en ethiek: afstemming FG’s, informeren betrokkenen, handreiking ethisch kompas

Bijlage 1: Overzicht gebruikte data bij ontwikkeling voorspelmodel

De volgende persoonsgegeven zijn gebruikt bij de ontwikkeling van het voorspelmodel:

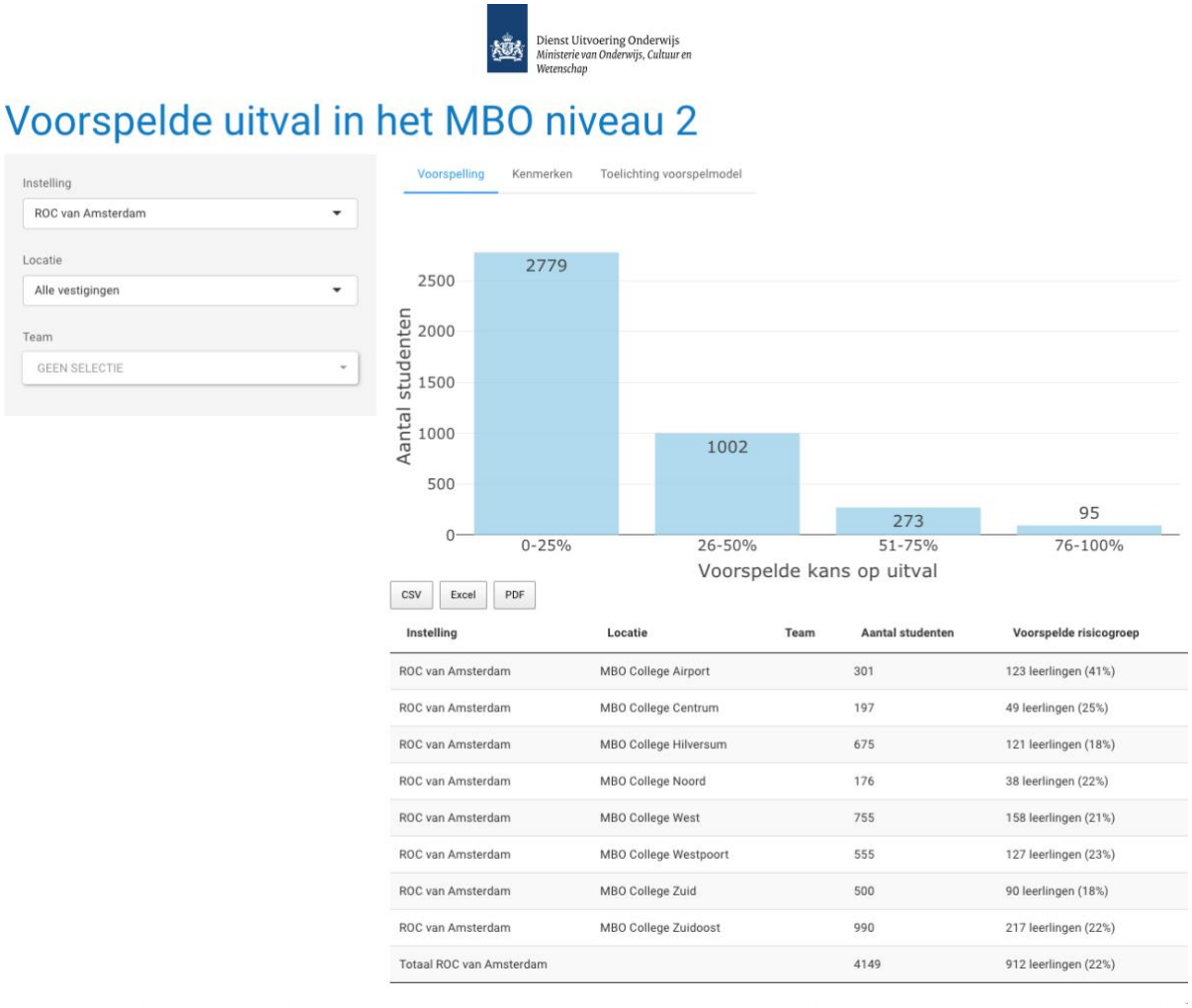
- VO-historie:
    - per inschrijvingsjaar in VO
        - Type VO (VWO, HAVO, VMBO etc.)
        - Leerjaar
        - Indicatie LWOO
    - Behaalde examencijfers Nederlands, Engels, Wiskunde
- PO-historie:
    - type PO (regulier of speciaal basisonderwijs)
    - behaalde resultaat eindtoets
- 4-cijferige postcode student
- Indicatie of student in ArmoedeProbleemCumulatie-gebied woont
- Leeftijd student
- Migratieachtergrond student (Geen MA, Westerse MA, Niet-Westerse MA, generatie)
- Inschrijvingsgegevens MBO:
    - Kwalificatiecode
    - MBO-niveau
    - Diploma behaald
    - Geldige inschrijving in MBO 1 jaar

Deze data worden verzameld over alle MBO-studenten op niveau 2 die zich voor het eerst inschrijven in het MBO. Hiervoor is geen aparte levering vanuit de instellingen aan OCW/DUO nodig. Het gaat hier om de verwerking van basisgegevens waarover DUO reeds beschikt. De data worden verzameld over de jaren 2015 tot 2020. In totaal komen we tot een databestand van ongeveer 200.000 studenten.

De bron van de data is de 1-cijfer-bestanden die door DUO-IP worden samengesteld om beleidsinformatie aan OCW en instellingen te leveren. DUO ontvangt persoonsgegevens van mbo studenten die worden ingeschreven en maakt op basis daarvan een algoritme.

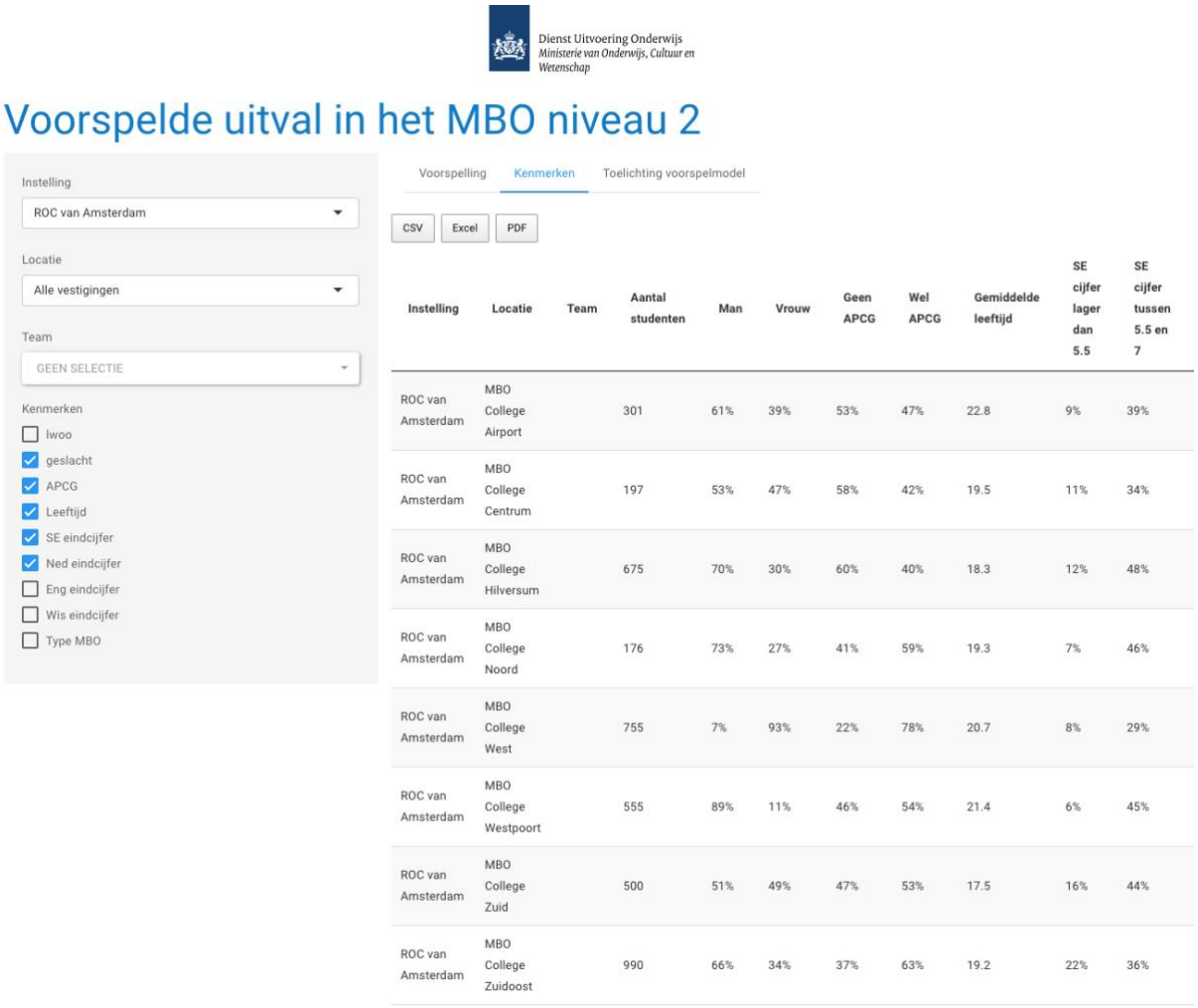
Bijlage 2: Screenshots dashboard voorspelmodel

Startscherm dashboard: toont voorspelde kans op uitval (voorbeeld ROC van Amsterdam)
Met spreiding en per MBO College het aantal ingestroomde niveau 2 studenten en de omvang van de voorspelde risicogroep.



Instelling	Locatie	Team	Aantal studenten	Voorspelde risicogroep
ROC van Amsterdam	MBO College Airport		301	123 leerlingen (41%)
ROC van Amsterdam	MBO College Centrum		197	49 leerlingen (25%)
ROC van Amsterdam	MBO College Hilversum		675	121 leerlingen (18%)
ROC van Amsterdam	MBO College Noord		176	38 leerlingen (22%)
ROC van Amsterdam	MBO College West		755	158 leerlingen (21%)
ROC van Amsterdam	MBO College Westpoort		555	127 leerlingen (23%)
ROC van Amsterdam	MBO College Zuid		500	90 leerlingen (18%)
ROC van Amsterdam	MBO College Zuidoost		990	217 leerlingen (22%)
Totaal ROC van Amsterdam			4149	912 leerlingen (22%)

Dashboard inzicht in kenmerken (voorbeeld ROC van Amsterdam)
Om de gepresenteerde voorspellingen te contextualiseren, kunt u per team zien hoe bepaalde variabelen die relevant zijn voor het onderliggende voorspelmodel verdeeld zijn over het team in kwestie.



Instelling	Locatie	Team	Aantal studenten	Man	Vrouw	Geen APCG	Wel APCG	Gemiddelde leeftijd	SE cijfer lager dan 5.5	SE cijfer tussen 5.5 en 7
ROC van Amsterdam	MBO College Airport		301	61%	39%	53%	47%	22.8	9%	39%
ROC van Amsterdam	MBO College Centrum		197	53%	47%	58%	42%	19.5	11%	34%
ROC van Amsterdam	MBO College Hilversum		675	70%	30%	60%	40%	18.3	12%	48%
ROC van Amsterdam	MBO College Noord		176	73%	27%	41%	59%	19.3	7%	46%
ROC van Amsterdam	MBO College West		755	7%	93%	22%	78%	20.7	8%	29%
ROC van Amsterdam	MBO College Westpoort		555	89%	11%	46%	54%	21.4	6%	45%
ROC van Amsterdam	MBO College Zuid		500	51%	49%	47%	53%	17.5	16%	44%
ROC van Amsterdam	MBO College Zuidoost		990	66%	34%	37%	63%	19.2	22%	36%

Bijlage 3: contactgegevens van de voor het initiatief verantwoordelijke personen

- Projectleider Datacoalitie Datagedreven onderzoek mbo, Tom Olsthoorn, tom.olsthoorn@hutspot.nl
- Datascientist DUO-INP, Martin Belder, martin.belder@duo.nl