Praktische producten voor inzet binnen jouw instelling

De Uitnodigingsregel: DPIA Pilotproject

HOE ZET JE DIT PRODUCT IN?

Een DPIA (data protection impact assessment) is een instrument dat kan worden gebruikt om vooraf de privacyrisico's van gegevensverwerking in kaart te brengen. Je kan als organisatie dan afwegen of de opbrengst in balans is met de vastgestelde risico's en waar gewenst extra maatregelen nemen om deze risico's te verkleinen.

Deze producten zijn ontwikkeld voor de praktijkpilot van de datacoalitie DGO. Het is aan de instelling om ze aan te vullen en te verrijken met de eigen context.

Andere beschikbare producten:

- Interventie overzicht
- Kwaliteit van het model
- Procesplaat
- Ethische hulpmiddelen
- Plan van aanpak
- Spelregels
- Kick-off presentatie

DATACOALITIE
DGO
MBO

Inhoud

DATACOALITIE
DGO
MBO

Versiebeheer

Versie	Datum	Auteur(s)	Opmerking
0.1	1 oktober 2024	Datacoalitie DGO	Generieke DPIA gebaseerd op een on-premise analyse-omgeving. Te gebruiken door andere mbo instellingen die aan de slag willen met de uitnodigingsregel. Het is aan de gebruiker om verwijzingen naar de interne technologie, processen, mensen en beleid toe te voegen en de DPIA voor te leggen aan FG of soortgelijke functies binnen de instelling.

A. Beschrijving gegevenskenmerken

1. Voorstel

Studentuitval is, zoals gepresenteerd in de Staat van het onderwijs 2023, in tien jaar niet zo hoog geweest. Bij meerdere mbo-instellingen is het voorkomen en/of terugdringen van uitval een prioriteit die terug te vinden is in de strategische doelen. Pilotproject “de uitnodigingsregel” implementeert een signaleringsmethode waarmee uitvalpreventie en -interventie kan worden verbeterd.

Instellingen worstelen al jaren om meer grip op uitval te krijgen. Daarbij wordt steeds meer gebruik gemaakt van data over de ontwikkeling van de studieloopbaan van de student. In haar promotieonderzoek heeft Irene Eegdeman (in die tijd werkzaam als docent-onderzoeker en SLB’er bij TOP) een methode geïntroduceerd waarin studenten met de hoogste kans op uitval in het eerste jaar "gesignaleerd" kunnen worden. Met gebruik van studiedata en met behulp van machine learning modellen is de zogenaamde 'Uitnodigingsregel' ontwikkeld. De methode geeft SLB'ers en mentoren een signaleringssysteem waarmee uitvalpreventie en -interventie kan worden verbeterd. Zes onderwijsinstellingen, hebben in een replicatieonderzoek, met historische data, aangetoond dat de ontwikkelde methodiek werkt en in staat is om studenten, met verhoogde kans op uitval, vroegtijdig te identificeren in verschillende opleidingen.

Om signalering zo accuraat mogelijk te maken wordt voor elke opleiding een eigen model ontwikkeld. Na signalering kunnen studenten -indien nodig bevonden door de SLB’er- tijdig worden uitgenodigd voor een gesprek waarin stappen kunnen worden besproken om potentiële uitval tegen te gaan. Op dit moment gebruiken SLB’ers data zoals cijfers en verzuim (bv. doormiddel van Socius) om hun gesprekken te kunnen plannen en datagedreven de inhoud vorm te kunnen geven. Het model van “de uitnodigingsregel” maakt van dezelfde data gebruik om accurater studenten met een grote kans op uitval te signaleren.

De gegevens (variabelen) die het model voeden zijn bepaald op basis van de methodiek van Eegdeman (2023), waarvan aantoonbaar is gemaakt dat deze informatie noodzakelijk is om accuraat genoeg uitval te kunnen voorspellen. Een bijkomend voordeel is dat volgens deze bestaande methodiek al bekend is welke (persoons)gegevens als in- en output dienen voor de modellen, en dus geen additionele data verwerkt hoeft te worden. Daarnaast biedt het model een alternatief voor de intuïtieve selectie van de SLB’er die mogelijk tot stand komt uit (impliciete) biases. Het product is een geordende lijst (op kans van uitval) van studenten (op basis van hun naam en klas/groep) waarvan een SLB’er gebruik kan maken om gesprekken te plannen om zo tijdig te kunnen interveniëren om uitval tegen te gaan.

Specifiek maken we gebruik van leeftijd, geslacht, leerjaar, vooropleidingniveau, richting van de vooropleiding, verzuim, aanmelddatum, en hoeveelheid aanmeldingen om uitval in het eerste jaar te voorspellen. Vanuit het oogpunt van dataminimalisatie maken we geen gebruik van vooropleidingcijfers, en intaketoetsen, die wel in de methodiek van Eegdeman (2023) worden gebruikt. Daarnaast maken we geen gebruik van voorspellende, maar gevoelige factoren zoals sociaaleconomische status of armoedeprobleemcumulatiegebieden. Verdere minimalisatie van de data is ethisch onverantwoord, aangezien dit de nauwkeurigheid van de voorspellingen van studentenuitval zou verminderen richting willekeur.

Nu bekend is dat de techniek werkt, hebben het MT, CvB en scholen aangegeven dat deze methode zorgvuldig naar de onderwijspraktijk toe moet worden gebracht.
<<Formulering aanpassen naar de situatie van de instelling>>
Er is daarom een pilot gestart voor implementatie in het schooljaar 2024/2025. [x] scholen hebben zich aangemeld voor deze praktijkpilot.

De verwerkingsstappen van persoonsgegevens worden in deze DPIA in meer detail omschreven. Hoog-over geldt dat:

- ROC [x] vanuit de afspraken t.a.v. de strategische doelstellingen en bijbehorende wetgeving verplicht is om maatregelen te treffen tegen het reduceren van voortijdige uitval. Deze pilot is onderdeel van deze maatregelen.

- De modellen (algoritmes) bij de uitnodigingsregel de minimale hoeveelheid gegevens gebruiken om tot een adequaat accuraat model te komen. Deze gegevens zijn bepaald o.b.v. de gevalideerde methodiek.

- De gegevens die worden verwerkt in principe reeds zichtbaar zijn voor (en dus verwerkt kunnen worden door) SLB'ers (de eindgebruikers). Deze informatie wordt enkel verrijkt met voorspellingen op studentniveau (afgeleid persoonsgegeven) om te helpen prioriteren.

2. Persoonsgegevens

De AVG onderscheidt drie typen van persoonsgegevens – (1) gewone, (2) bijzondere en (3) strafrechtelijke persoonsgegevens – en stelt verschillende eisen aan een rechtmatige verwerking daarvan. Bij de implementatie van dit project 'pilot Uitnodigingsregel'', verwerken wij geen data typen (2) bijzondere en of data typen (3) strafrechtelijke persoonsgegevens.

Het machine learning model wordt getraind op basis van de onderstaande (persoons)gegevens. Omdat het een pilot betreft op basis van een bestaande, gevalideerde methodiek die specifiek is ontwikkeld voor het MBO (Eegdeman, 2023), maken de modellen die bij ROC [x] worden ontwikkeld ook voor zoveel mogelijk gebruik van dezelfde gegevens (afhankelijk van beschikbaarheid en data-kwaliteit). Specifiek betreft dit de lijst van variabelen in Eegdeman (2023), Appendix 5B, p. 134. Om het gebruik van persoonsgegevens te minimaliseren maken we geen gebruik van vooropleidingcijfers en intake toetsen die wel vermeld staan in de lijst van Eegdeman (2023), Appendix 5B, p. 134. Daarnaast maken we geen gebruik van voorspellende, maar gevoelige factoren zoals sociaaleconomische status of armoedeprobleemcumulatiegebieden, die wel frequent worden gebruikt in bestaande uitval gerelateerde projecten en informatieproducten. In haar geüpdatete methodiek heeft Eegdeman vooropleidingniveau, richting van de vooropleiding, aanmelddatum, en hoeveelheid aanmeldingen toegevoegd. Op aanraden van Eegdeman en na testen zijn deze toegevoegd omdat het anders onmogelijk is om nauwkeurig studenten uitval in het eerste jaar te voorspellen.

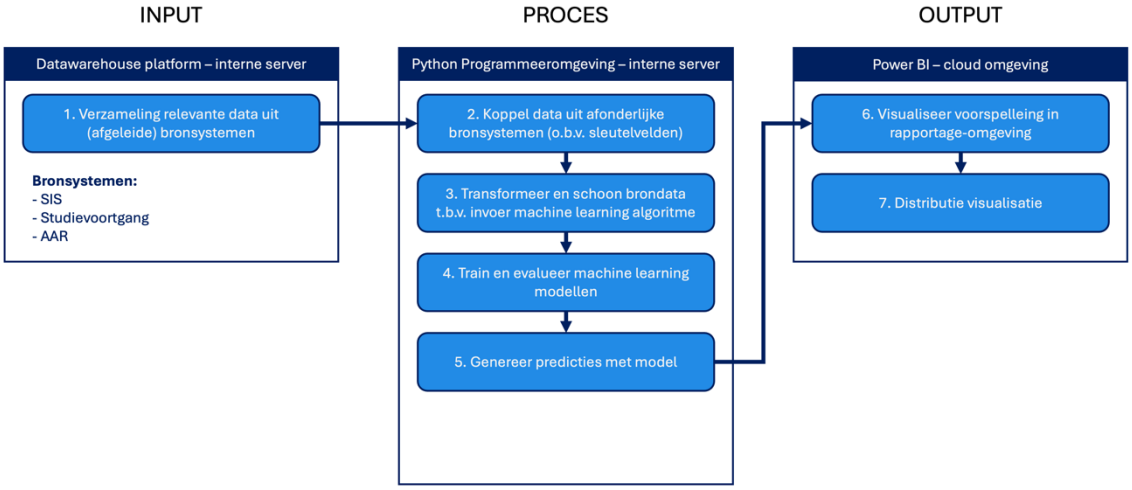
Te verwerken persoonsgegevens t.b.v. de uitnodgingsregel							
	Persoonsgegeven typen			Bron systeem			
Attribuut (Één systeem veld dat een persoonsgegeven indicator weergeeft)	**Gewone**	**Bijzonder**	**Strafrechtelijke**	**SIS**	**PVerzuim-systeem**	**Aanmeldsysteem**	**Reden/Toelichting**
Stamnummer	X	-	-	X			• Stamnummer en Student ID zijn een bovenliggend unieke sleutel (ankerpunt) om 'dynamisch object' (student) over verschillende bron systemen heen correleer baar te houden.
Student ID	X	-	-	X	X	X	
Naam	X	-	-	X			• De studenten zijn voor ons in de datamodellen letterlijk alleen één nummer. • Om de informatie uit deze modellen daarna functioneel te maken wordt dit nummer op de server buiten zicht van de personen die die modellen ontwikkelen weer aan de student gekoppeld en in een lijstvorm via een dashboard toegankelijk voor de SLB'er gemaakt.
Huidige leeftijd	X	-	-	X			Voorspeller (model input) zoals gedefinieerd in methodiek Eegdeman (2023).
Geslacht	X	-	-	X			Voorspeller (model input) zoals gedefinieerd in methodiek Eegdeman (2023).
Opleiding	X	-	-	X			Uitsluitend gebruikt om opleiding specifieke modellen te creëren. Elke opleiding heeft eigen model.
Leerweg	X	-	-	X			• BOL en BBL-leerwegen hebben aparte modellen nodig. • Momenteel worden alleen BOL-leerwegen voorspeld.
Leerjaar	X	-	-	X			• Voorspeller (model input) zoals gedefinieerd in methodiek Eegdeman (2023) ("Cohort"). • T.b.v. afwijkende trend jaren (b.v. Coronajaren)
Klas, Groep	X	-	-	X			Om de dashboard schermen relevant te kunnen differentiëren naar rol(en) van de gebruiker, opleiding, vakgroep & klas.
Vooropleiding (niveau en richting)	X	-	-	X			Voorspeller (model input) geüpdatet methodiek Eegdeman (2023).
Aanmeldingen: hoeveel aanmeldingen + hoeveel dagen voor studiestart		-	-			X	Voorspeller (model input) geüpdatet methodiek Eegdeman (2023).
Opleiding beëindigd tijdens/na eerste jaar	X	-	-	X			Uitkomst om te voorspellen.
Verzuim: afwezigheid uren per dag	X	-	-		X		Voorspeller (model input) zoals gedefinieerd in methodiek Eegdeman (2023).

*data uit de groen gearceerde attributen worden uitsluitend verwerkt bij de selectie van data

*data uit de blauw gearceerde attributen is niet zichtbaar tijdens dataverwerking en worden niet gebruikt bij de uitvalvoorspelling.

3. Gegevensverwerking

De onderstaande Figuur 1 en Tabel 2 beschrijven de gegevensverwerkingsstappen die benodigd zijn om het product te kunnen leveren. Tabel 2 beschrijft per stap uit de figuur welke betrokkenen en persoonsgegevens dit betreft.



Figuur 1: Stroomdiagram gegevensverwerkingen

Tabel 2: Overzicht verwerkingsstappen				
Verwerkings-stap	Toelichting	Categorie betrokkenen	Categorie persoonsgegevens	Persoonsgegevens
1.	Ophalen van brongegevens die benodigd zijn om tot de input en output van model te komen, en dit te kunnen weergeven in de output	Studenten	Naam, demografisch, studiegericht	Alle gegevens Tabel 1
2.	Koppelen van gegevens uit afz. bronnen tot een geïntegreerde dataset	Studenten	Naam, demografisch, studiegericht	Alle gegevens Tabel 1
3.	Gegevens opschonen en verwerken zodat deze in een geldig/optimaal format voor de modellen staan.	Studenten	Demografisch, studiegericht	In- en output variabelen (huidige leeftijd, geslacht, leerjaar, vooropleiding, aanmeldingen, opleiding beëindigd, verzuim)
4.	Trainen van daadwerkelijk ML-model met verwerkte inputvariabelen	Studenten	Demografisch, studiegericht	Zie stap 3.
5.	Getraind model gebruiken om voorspellingen te genereren	Studenten	Demografisch, studiegericht	Alle geprepareerde inputvariabelen (output van stap 3.)
6.	Gegenereerde voorspellingen per student naar de visualisatietool sturen en rapport publiceren	Studenten	Naam, studiegericht	Model-predicties (output stap 5), Naam en Klas/Groep
7.	Embedding met Socius tool, toegang tot de inzichten voor de SLB' ers	Studenten	Naam, studiegericht	Zie stap 6.

Gegevensverwerking blijft volledig bij het eigen ROC.

- In week 0 tm 10 van het nieuwe schooljaar wordt er wekelijks een model voor elk van de 5 opleidingen gedraaid op de applicatieserver, met data van het Datawarehouse. Met deze modellen wordt studentuitval voorspeld. Dit is volledig ‘on premise.’
- De model uitkomsten worden opgeslagen in het Datawarehouse en via Student ID gekoppeld op naam. Dit is volledig on premise.
- Vervolgens wordt de informatie op de (on premise) productieserver via een gateway naar de cloud gebracht om in een dasboard in de vorm van een lijst gevisualiseerd te worden voor de eindgebruiker. We hebben het hier (en in de rest van de DPIA) over de standaard ROC cloud producten die we binnen de organisatie veelvuldig voor (gevoelige) data gebruiken, geen nieuwe externe cloud partijen.*
- De informatie wordt gepubliceerd in een dashboard via het product Microsoft Power BI.
- Business information communiceert vervolgens naar de deelnemende scholen/opleidingen dat de dashboards gevuld zijn met actuele informatie.

*Alle gegevens en informatie tussen de on-premises datagateway en Power BI zijn versleuteld. De in Power BI ingevoerde inloggegevens voor de gegevensbron worden versleuteld en vervolgens in de cloud opgeslagen. Alleen de gateway kan de inloggegevens ontsleutelen.

Er worden er geen inkomende poorten door de gateway gebruikt. Dit zorgt voor een veilig communicatieprotocol tussen uw on-premises databaseserver en de Power BI-service. Deze beveiliging wordt voor alle interne Power BI dashboards gebruikt.

4. Verwerkingsdoeleinden

Het doel van de verwerkingen is tweezijdig, afhankelijk van de verwerkingsstap (Tabel 2):

1. Om de machine learning algoritmes te kunnen trainen, en predicties af te geven voor studenten. De gegevens die hiervoor verwerkt worden staan benoemd in Tabel 2, stap 1 t/m 5. Deze gegevens zijn benodigd om een model te trainen wat adequaat accuraat uitval kan voorspellen, zonder overbodige variabelen mee te nemen.
2. Om SLB’ers van de relevante informatie te voorzien (d.m.v. geordende lijsten in PowerBI) zodat zij de inzichten van de modellen kunnen benutten in het begeleiden van studenten. De gegevens die hiervoor verwerkt worden staan benoemd in Tabel 2, stap 6 & 7. Deze gegevens zijn benodigd (naam, klas/groep) zijn minimaal nodig om de voorspellingen te koppelen aan studenten en deze bruikbaar te maken voor SLB’ers, dat wil zeggen dat SLB’ers hiermee de studenten kunnen identificeren en uitnodigen voor een gesprek en mogelijke vervolgacties.

Het doel van alle persoonsgegevens staat concreet beschreven in Tabel 1.

5. Betrokken partijen

Dit zijnde de betrokken partijen in kader van gegevensverwerking.

Rol	Werkzaamheden	Instelling	Team / Persoon
Verwerkings-verantwoordelijke	• Coördinatie dashboard publicatie	x	Afd. Business information Steven Ramondt
Verstrekker	• Systeem data	x	Afd. Business information Techniek
Verwerker	• Data schonen • Data transformeren • Model ontwikkelen	x	Afd. Business information Steven Ramondt
Verwerker	• Dashboard optimalisatie	x	Afd. Business information Techniek
Verwerker	• Link opnemen Socius	x	Technische applicatie beheerders Peter Groen en Danny Oosenbrug, Functionele applicatie beheerder Bastiaan Wanders.

6. Belangen bij de gegevensverwerking

De belanghebbenden zijn de student, SLB'ers en ROC [x]. De gegevensverwerking is nodig ten behoeve van het voorkomen van eerstejaars studenten uitval. Onderstaand overzicht weergeeft hun belangen.

Instelling	Belangen
Student	Optimaal ondersteund worden voor studiesucces. Tijdig (extra) hulp en ondersteuning krijgen indien noodzakelijk en/of gewenst
SLB'er	Krijgt toegang tot instrument om studenten tijdig te kunnen uitnodigen en hoeft zich niet meer alleen op intuïtie te baseren die mogelijk leidt tot (impliciete) biases.
x	De gegevensverwerking is nodig ten behoeve van het voorkomen van eerstejaars studenten uitval. Dit is een strategisch doel van ROC x, waar afspraken over zijn gemaakt met ministerie OCW. Vergroot de mogelijkheid om de begeleiding van studenten tijdens de studie te verbeteren. ROC x heeft hier een wettelijke taak in, en bindende doelstellingen binnen haar strategische kader op gesteld.

7. Verwerkingslocaties

De verwerking vindt geheel bij x in Europa plaats. Tot het wordt gevisualiseerd via de x cloud in Microsoft Power BI vind dit zich geheel in Nederland plaats. Microsoft Power BI is een standaard x

cloud product. De visualisatie wordt getoond in de interne BI-omgeving. De verwerking van de data gebeurt op een on-premise server, de servers bevinden zich bij x in Nederland.

Hieronder de locatie waar data zich bevindt per technologie en of technologie houder.

Technologie	**Locatie**
Microsoft Power BI	Europa, Ierland
DWH-omgeving (x)	Europa, Nederland

8. Technieken en methoden van gegevensverwerking

De voornaamste methodiek voor gegevensverwerking betreft die van kunstmatige intelligentie en algoritmen. Specifiek gaat het om een supervised learning model. De betrokken gegevens worden gebruikt als in- en/of output om een reeks aan modellen te trainen die trachten zo goed mogelijk de kans op uitval per opleiding te voorspellen. Wederom geldt dat om de relevante in- en output te bepalen, en de prestaties van de modellen te evalueren (bijv. de accuratesse van voorspellingen te beoordelen) de ontwikkelde methodiek van Eegdeman (2023) als leidraad wordt genomen. De verwerkingsstappen zijn beschreven in figuur 1 en tabel 2. Daarnaast is er voor deze algoritmes gekozen omdat het mogelijk is om de voorspelling van deze algoritmes inzichtelijk te maken. Het is hierdoor te verklaren hoe individuele voorspellingen tot stand zijn gekomen.

Elk uiteindelijk model is gebaseerd op een van drie welkbekende machine learning algoritmes: lasso regression, random forest regression, of support vector machines. De replicatie studie heeft aangetoond dat deze algoritmes adequaat uitval kunnen voorspellen. Bij al deze algoritmes geldt dat het bekende, veelgebruikte algoritmes zijn binnen het vakgebied en voor hetzelfde doel ingezet kunnen worden (gegeven dezelfde in- en outputs). Voor elk algoritme is het doel om voorspellingen van uitval te genereren die zo zicht mogelijk bij de historische geobserveerde uitval liggen. Het algoritme ‘leert’ door iteratief de afwijkingen tussen de voorspelde uitkomst en de werkelijke (historische) uitkomst te minimaliseren, totdat het een optimum bereikt. Het resulterende model wordt vervolgens gebruikt om voor toekomstige gevallen te voorspellen.

Het hoofdzakelijke verschil zit in de mathematische uitwerking van deze algoritmes, ook al is het globale optimalisatieproces hetzelfde:

- Het random forest algoritme maakt gebruik van een groot aantal beslisbomen (welke aparte machine learning modellen zijn), die per variabele (bijv. heeft de student meer dan 10% verzuim?) studenten in steeds kleinere groepen met bepaalde kans op uitval indeelt. Door veel beslisbomen te combineren, ontstaat er een woud (forest) van mogelijke uitkomsten, waaruit het algoritme de gemiddelde voorspelling kan berekenen. Het voordeel van een random forest algoritme is dat het rekening kan houden met complexe patronen en interacties tussen de kenmerken, en dat het minder gevoelig is voor ruis en overfitting dan een enkele beslisboom.
- Het lasso regressie algoritme is een manier om uitval te voorspellen op basis van een lineaire combinatie van de kenmerken van de studenten, zoals leeftijd, geslacht, opleiding, cijfers en verzuim. Het algoritme zoekt naar de optimale gewichten die elke kenmerk moet krijgen om de afhankelijke variabele zo goed mogelijk te benaderen. Het voordeel van een lasso

regressie algoritme is dat het eenvoudig te interpreteren is en dat het overfitting kan voorkomen door minder belangrijke kenmerken te verwijderen.

- Het support vector machines algoritme zoekt naar de optimale scheidingslijn (of hypervlak) die groepen zo goed mogelijk van elkaar kan scheiden. Bijvoorbeeld: Als de student meer dan 25 jaar oud is en meer dan 15% verzuim heeft, dan wordt hij of zij geclassificeerd als een uitvaller. Het algoritme maakt gebruik van zogenaamde support vectors, die de punten zijn die het dichtst bij de scheidingslijn liggen. Het voordeel van een support vector machine algoritme is dat het goed kan omgaan met niet-lineaire scheidingslijnen en dat het robuust is tegen ruis en outliers.

Voor een verdere wiskundige toelichting van deze algoritmes kan bijvoorbeeld worden verwezen naar de uitwerking in Eegdeman (2023, Appendix 5A – Machine Learning algorithms, p. 133) of de eerder verwezen documentatie van scikit-learn.

Het doel is om per dataset (dus per school) het best presterende model te trainen (door alle bovengenoemde algoritmes te gebruiken), gegeven dezelfde in- en output variabelen (zoals omschreven in Tabel 1). Omdat de kenmerken van de data variëren (per school/studie) is het bepalen van het best presterende model een empirische kwestie; dit is lastig a-priori te bepalen. Dat wil zeggen dat het beste model (per school) bepaald wordt door alle drie de algoritmes te trainen op deze data, en hieruit op basis van gebruikelijke evaluatie metrieken (zie Eegdeman, 2023) het beste model te selecteren. Het volgen van de bestaande methodiek en replicatiestudie heeft als bijkomend voordeel dat deze algoritmes (uit een veel bredere set aan mogelijke opties) reeds geïdentificeerd zijn als adequaat, en dus niet overmatig veel algoritmes getraind dienen te worden.

9. Juridisch en beleidsmatig kader

De activiteiten die ondernomen worden voor de uitnodigingsregel dienen het Strategie 2024-2027 doel om VSV én algemeen schoolverlaten terug te dringen. Dit blijkt uit de volgende beleidsmatige kaders:

- De overheid stelt normatieve kaders met betrekking tot Voortijdig School Verlaten. Mbo-instellingen maken afspraken met de minister van Onderwijs, Cultuur en Wetenschap (OCW) en leggen deze vast in de Strategie 2024-2027 van de mbo-instelling
- <<Verwijzing opnemen naar de interne kwaliteitsafspraken, missie en visie van de instelling waar mogelijk en gepast.>>De Strategie 2024-2027 van ROC x heeft onder speerpunt 1: Kansengelijkheid de doelstelling om begeleiding te versterken en vsv te verminderen (Doel 1.3). ROC x is verantwoordelijk voor het behalen van dit doel, welke geëvalueerd zal worden door een onafhankelijke commissie.
- De relevante wet op basis waarvan de gegevens worden verwerkt/er een grondslag bestaat om deze gegevens te verwerken is Art 8.3.4: Wet educatie en beroepsonderwijs, welke verwijst naar programma's met maatregelen ter voorkoming en bestrijding van voortijdig schoolverlaten van jongeren tussen twaalf en drieëntwintig jaar.

10. Bewaartermijn

De uitkomsten van de modellen worden 24 maanden na gebruik bewaard, en alleen ten behoeve van de evaluatie van de gebruikte modellen en evaluatie van de praktijkpilot gebruikt.

Het overgrote gedeelte van de data moet 2 jaar na uitschrijving bewaard worden. Nadat de student de school heeft verlaten is de school volgens de archiefwet verplicht om persoonsgegevens nog 2 jaar te bewaren. Ook resultaten moeten minimaal 2 jaar worden bewaard wanneer de student wordt uitgeschreven van de opleiding. Aanwezigheidsdata dient 7 jaar na uitschrijving bewaard te worden. Dit staat opgenomen in het Documentair Structuurplan (DSP) van het MBO.

B. Beoordeling rechtmatigheid gegevensverwerkingen

11. Rechtsgrond

De AVG geef als beginsel dat persoonsgegevens moeten worden verwerkt op een wijze die ten aanzien van de betrokkene rechtmatig, behoorlijk en transparant is. Als uitwerking van dit beginsel is geregeld dat een gegevensverwerking alleen rechtmatig is als deze gebaseerd kan worden op ten minste een van de volgende zes rechtsgronden: 1. De betrokkene heef toestemming gegeven voor de verwerking van zijn persoonsgegevens voor een of meer specifieke doeleinden. 2. De verwerking is noodzakelijk voor de uitvoering van een overeenkomst waarbij de betrokkene partij is, of om op verzoek van de betrokkene vóór de sluiting van een overeenkomst maatregelen te nemen. 3. De verwerking is noodzakelijk om te voldoen aan een wettelijke verplichting die op de verwerkingsverantwoordelijke rust. 4. De verwerking is noodzakelijk om de vitale belangen van de betrokkene of van een andere natuurlijke persoon te beschermen. 5. De verwerking is noodzakelijk voor de vervulling van een taak van algemeen belang of van een taak in het kader van de uitoefening van het openbaar gezag dat aan de verwerkingsverantwoordelijke is opgedragen.6. De verwerking is noodzakelijk voor de behartiging van de gerechtvaardigde belangen van de verwerkingsverantwoordelijke of van een derde, behalve wanneer de belangen of de grondrechten en de fundamentele vrijheden van de betrokkene die tot bescherming van persoonsgegevens nopen, zwaarder wegen dan die belangen, met name wanneer de betrokkene een kind is.

De AVG geef als beginsel dat persoonsgegevens moeten worden verwerkt op een wijze die ten aanzien van de betrokkene rechtmatig, behoorlijk en transparant is.

<u>Rechtmatig</u>

De AVG geef als beginsel dat persoonsgegevens moeten worden verwerkt op een wijze die ten aanzien van de betrokkene rechtmatig, behoorlijk en transparant is. Als uitwerking van dit beginsel is geregeld dat een gegevensverwerking alleen rechtmatig is als deze gebaseerd kan worden op ten minste een van zes mogelijke rechtsgronden (aldus het Model DPIA Rijksdienst). Voor deze pilot gelden de volgende rechtsgronden:

1. De verwerking is noodzakelijk om te voldoen aan een wettelijke verplichting die op de verwerkingsverantwoordelijke rust
2. De verwerking is noodzakelijk voor de vervulling van een taak van algemeen belang of van een taak in het kader van de uitoefening van het openbaar gezag dat aan de verwerkingsverantwoordelijke is opgedragen.

Deze rechtsgronden worden geldig geacht voor alle gegevensverwerkingen, aangezien zij noodzakelijk zijn om het eindproduct te kunnen leveren.
De huidige gegevensverwerkingen moeten worden uitgevoerd als onderdeel van een maatregel t.a.v. een van de strategische doelen van ROC x, waarvan een wettelijke verplichting hangt (Doel 1.3 ; WEB Art. 8.3.4.).

Voor details zie de onderbouwing juridisch en beleidsmatig kader.

12. Bijzondere persoonsgegevens

Er worden geen bijzondere persoonsgegevens verwerkt.

13. Doelbinding

De gegevens zijn oorspronkelijk verwerkt op basis van de Wet educatie en beroepsonderwijs. De verwerkingen beschreven in deze DPIA vallen zoals toelicht onder het juridische kader onder dezelfde wet. Specifiek zijn deze gegevensverwerkingen noodzakelijk om een maatregel te kunnen inzetten met als doel (vroegtijdig) schoolverlaten te voorkomen. Op basis daarvan concluderen we dat de gegevensverwerking in dit project verenigbaar zijn met de oorspronkelijke verwerking.

De te verwerken gegevens worden in een leeromgeving ook verzameld en ingezien door een SLB'er om de student adequaat te kunnen begeleiden en om de studievoortgang vast te stellen. De uitnodigingsregel maakt gebruik van dezelfde gegevens om ditzelfde doel accurater en met minder bias vast te stellen. De uitkomsten komen in ditzelfde studievoortgang inzichtsysteem en kunnen naar inzicht (na training) door de SLB'er worden gebruikt. Oorspronkelijk doel en nieuw doel komen dus overeen en zijn dus verenigbaar.

Systeem	Doel
Onderwijsapplicties opnemen	Student informatiesysteem
Onderwijsapplicties opnemen	Aan -en afwezigheidsregistratie
Onderwijsapplicties opnemen	Aanmeld en intake systeem
Onderwijsapplicties opnemen	Studievoortgang inzichtsysteem

14. Noodzaak en evenredigheid

Proportionaliteit: De gegevensverwerking zijn nodig ten behoeve van betere begeleiding van de student en het voorkomen van eerstejaars studenten uitval, een kerntaak van de instelling. Daarnaast is de noodzaak hoog. De studentuitval is, zoals gepresenteerd in de Staat van het onderwijs 2023, in tien jaar niet zo hoog geweest. OCW in hun kamerbrief stelt uitvalvoorspelling als maatregel voor.

De uitnodigingsregel maakt gebruik van gegevens waartoe de SLB'er nu ook al veelal toegang toe heeft (o.a. om begeleidingsgesprekken te kunnen voeren met studenten), het betreft hier enkel een verrijking van deze gegevens met een (op machine-learning) gebaseerde voorspelling (uitvalkans) die helpt de SLB'ers hun begeleiding (beter) te prioriteren. Omdat een verbetering in het werkproces (m.b.t. begeleiding door SLB'ers) bereikt kan worden beschouwen we deze verwerking als proportioneel.

Subsidiariteit: Zonder de voorgestelde gegevens is het onmogelijk om uitval accuraat te voorspellen. Dit is met wetenschappelijk onderzoek in de vorm van een replicatieonderzoek aangetoond. Daarnaast heeft DUO geprobeerd dit met hun gegevens te verwezenlijken (ROD-data, geen instelling data), maar DUO kreeg geen accuraat beeld. Dezelfde prestaties van de modellen kunnen dus niet op een andere manier (met minder gegevens) bereikt worden. Het is nogmaals waard om te vermelden dat de gegevens voor de uitnodigingsregel nu al voor precies hetzelfde doeleinde wordt gebruikt

15. Rechten van betrokkenen

Mensen hebben een aantal rechten als organisaties hun (persoons)gegevens verwerken. Dit noemen we privacy rechten, ook wel de rechten van betrokkenen. x faciliteert deze rechten. In relatie tot de pilot met de Uitnodigingsregel is het volgende relevant:

Rechten van betrokkene	**Procedure ter uitvoering**	**Beperking op grond van wettelijke uitzondering**
Recht van inzage	De student heeft te allen tijde recht op inzage in zijn studievoortganggegevens. Deze zijn reeds beschikbaar voor studenten, onafhankelijk van de verwerkingen benodigd voor dit project.	
Recht op rectificatie en aanvulling	Studenten en of medewerkers van de school kunnen altijd verzoeken om een aanpassing of verduidelijking van gegevens via afd. Kwaliteitszorg.	Privacy policy van x voorziet hierin.
Recht op vergetelheid	De afgesproken bewaartermijn voor gebruikte gegevens is 24 maanden. Bronsysteem data maakt het onmogelijk om input data te verwijderen. Model data zoals de uitval kans kan verwijderd worden.	Archivering afgestemd om evaluatie en verbetering van de resultaten vast te stellen.
Recht op dataportabiliteit	De student kan een verzoek indienen om zijn/haar persoonsgegevens (opgeslagen in de bronsystemen) op te vragen of over te laten dragen. De model-outputs (voorspellingen) zijn afgeleide persoonsgegevens, deze is ROC x niet verplicht te verstrekken.	
Recht niet onderworpen te worden aan geautomatiseerde besluitvorming	n.v.t., er is geen geautomatiseerde besluitvorming.	
Recht om bezwaar te maken	Er is een klachtenregeling van x.	

C. Beschrijving en beoordeling risico's voor de betrokkenen

16. Risico's

Hieronder volgt een uiteenzetting van de ingeschatte risico's, onder hoofdstuk 16. Maatregelen wordt beschreven hoe hier mee wordt omgegaan en een eventueel resterend risico ingeschat.

Risico	Kans	Impact	Toelichting
Datalek van de gegevens van de leerlingen/studenten	Beperkt	Groot	Indien er sprake is van een inbreuk op de beveiliging van de persoonsgegevens; toegang hiertoe zonder dat dit mag of de bedoeling is.
Ten onrechte inzage in de persoonsgegevens	Beperkt	Middel	Betreft het risico dat onbevoegden inzage krijgen in de persoonsgegevens die worden verwerkt.
Ongeautoriseerde toegang Studievoortgangomgeving	Beperkt	Middel	Betreft het risico dat onbevoegden inzage krijgen in de persoonsgegevens die worden verwerkt.
Ongeautoriseerde toegang Power BI	Beperkt	Middel	Betreft het risico dat onbevoegden inzage krijgen in de persoonsgegevens die worden verwerkt.
Fout in voorspelling	Groot	Beperkt	Model is niet accuraat in een voorspelling met als gevolg dat een student later aan de beurt is voor een SLB-gesprek dan idealiter het geval is.
Onjuist gebruik van model-voorspellingen	Middel	Beperkt	Er bestaat een risico dat uitkomsten van de modellen foutief worden geïnterpreteerd of als absolute waarheid worden aangehouden zonder rekening te houden met (de beperkingen van) de methodiek.
Discriminatie/benadeling van leerlingen/studenten	Middel	Beperkt	Risico dat de (inzet van) modeluitkomsten binnen het begeleidingsproces leiden tot indirecte discriminatie of oneerlijke behandeling. Bijvoorbeeld indien SLB'ers enkel de hoogst-geschatte studenten (vaker) spreken ten nadele van overige studenten. Valt deels samen onder voorgaand risico omtrent onjuist gebruik modeluitkomsten.

D. Beschrijving voorgenomen maatregelen

17. Maatregelen

Zie H16 Risico's

Risico	Maatregelen	Resterend risico en risico-inschatting	Beheerder van maatregelen
Datalek van de gegevens van de leerlingen/studenten	Dataverwerking vindt volledig plaats in beveiligde omgeving. Datawarehouse op de applicatie en productieserver (on-premises). Een deel van de gegevens wordt gehost via Microsoft (vanaf de verwerkingsstap de gegevens naar PowerBI te versturen.) Hierbij geldt dat de Azure cloud omgeving minimaal voldoet aan de security en veiligheidseisen gesteld door ROC x. Autorisatie is nodig om in bovengenoemde omgevingen te kunnen komen. Aan de gebruikerskant wordt de optie om de rapportage te exporteren (bijvoorbeeld naar een lokale laptop) uitgezet. Daarnaast worden de gegevens naar PowerBI/SLB'ers geminimaliseerd tot enkel het strikt noodzakelijke benodigd voor een nuttig inzicht. Datelekken kunnen gemeld worden bij de FG om o.a. de impact hiervan waar mogelijk te	Naast algemene (grootschalige) datalekken bijv. door inbreuk in de systemen en/of mogelijk opzettelijk omzeilen van de geplaatste maatregelen (bijvoorbeeld niet houden aan beleid), wordt dit resterende risico beperkt geschat. Wederom geldt dat eindgebruikers in huidige proces al toegang hebben tot veel van de persoonsgegevens die leidend zijn voor de voorspelling.	Data-eigenaren, business information.

	minimaliseren en zo vroeg mogelijk te identificeren.		
Fout in voorspelling	Om met dit pilotproject mee te doen zijn de opleidingen verplicht om alle studenten te spreken. Ook is het model niet de enige bron om de volgorde te bepalen en maakt de SLB'er ook gebruik van ervaring en de data in platte vorm (b.v. verzuim data).	Door triangulatie waarbij naast het project ook ervaring en de data in platte vorm gebruikt wordt om volgorde te bepalen wordt dit resterende risico beperkt geschat. Daarnaast zijn consequenties beperkt aangezien de enige consequentie een later gesprek is.	SLB'ers, Business information
Ten onrechte inzage in de persoonsgegevens	Voor ontwikkelaars en gebruikers gelden standaard privacy en dataveiligheid regels. Eindgebruikers (SLB'ers) ontvangen instructies over verantwoord omgaan met deze (persoons)gegevens. Toegang tot de persoonsgegevens wordt aan de ontwikkelkant beheerd d.m.v. rolautorisaties. Aan de gebruikerskant wordt een row-level security toegepast zodat SLB'ers enkel bij relevante studenten gegevens kunnen bekijken. Specifieke systemen worden hieronder behandeld.	Hier berust een verantwoordelijkheid op individuen om zorgvuldig met deze informatie om te gaan, dit is lastig met 'harde' maatregelen te beheren. De informatie die wordt blootgesteld is al zo veel mogelijk beperkt en niet uniek tot uitvoer van huidige project. Daarom schatten we het resterend risico als klein in.	Business information
Ongeautoriseerde toegang Studievoortgangsomgeving	De omgeving bevindt zich op IIS server on-premises. Er vind geen externe data uitwisseling plaats. Aan de gebruikerskant wordt een row-level security toegepast zodat SLB'ers enkel bij relevante studenten gegevens kunnen bekijken.	Idem berust hier een verantwoordelijkheid op individuen om zorgvuldig met deze informatie om te gaan, dit is lastig met 'harde' maatregelen te beheren. De informatie die wordt blootgesteld is al zo veel mogelijk beperkt en niet uniek tot uitvoer van huidige project. Socius is al buiten dit project is beschikbaar voor alle SLB'ers. Daarom	Business information

		schatten we het resterend risico als klein in.	
Ongeautoriseerde toegang Power BI	Alle gegevens en informatie tussen de on-premises datagateway en Power BI zijn versleuteld. De in Power BI ingevoerde inloggegevens voor de gegevensbron worden versleuteld en vervolgens in de cloud opgeslagen. Alleen de gateway kan de inloggegevens ontsleutelen. Er worden er geen inkomende poorten door de gateway gebruikt. Dit zorgt voor een veilig communicatieprotocol tussen uw on-premises databaseserver en de Power BI-service. Deze beveiliging wordt voor alle x Power BI dashboards gebruikt. Ingelogde SLB'er heeft alleen toegang tot studenten van eigen klas. Dit beperkt impact van ongeautoriseerde toegang.	Idem berust hier een verantwoordelijkheid op individuen om zorgvuldig met deze informatie om te gaan, dit is lastig met 'harde' maatregelen te beheren. De informatie die wordt blootgesteld is al zo veel mogelijk beperkt en niet uniek tot uitvoer van huidige project. Daarom schatten we het resterend risico als klein in.	Business information
Onjuist gebruik van model-voorspellingen	SLB'ers ontvangen gebruiksinstructies en awareness-training over hoe om te gaan met de outputs van het model. Deze instructies omvatten inzicht in de werking van het model, de beperkingen (zoals onzekerheid en foutmarges), en hoe de verkregen inzichten effectief kunnen worden	Het resterende risico wordt ook hier als klein ingeschat. De outputs van de modellen vervullen een ondersteunende rol in het bredere begeleidingsproces. SLB'ers worden niet verplicht om de volgorde te volgen, wel verplicht tot triangulatie (zie fout in voorspelling) en maken de uiteindelijke beslissing over (de volgorde van) gesprekken.	Business information

	ingezet in het begeleidingsproces.		
Discriminatie/benadeling van leerlingen/studenten	In de gebruikerstraining wordt ook omgegaan met dit topic om verantwoord gebruik te borgen en hier bewustzijn op te creëren. Daarnaast geldt dat het model wordt getoetst volgens beschikbare ethische kaders. Ook geldt dat SLB'ers niet verplicht zijn om de volgorde te volgen en maken de uiteindelijke beslissing over (de volgorde van) gesprekken. Na de initiële inzet van de pilot wordt feedback opgevraagd om het effect van het product op het begeleidingsproces te analyseren, en eventuele correcties te maken om mogelijke bias/discriminatie tegen te gaan. De opzet van het project is dusdanig dat deze ook huidige vooringenomenheid in de behandelingen van bestrijden door een meer data-gedreven aanpak.	Er bestaat een restrisico op gebied van potentieel discrimineren op factoren waarop praktisch niet te meten is omdat ROC x niet beschikt over deze data (bijvoorbeeld etniciteit).	Business information, SLB'ers

Referenties

Eegdeman, I. M. (2023). Enhancing Study Success in Dutch Vocational Education.