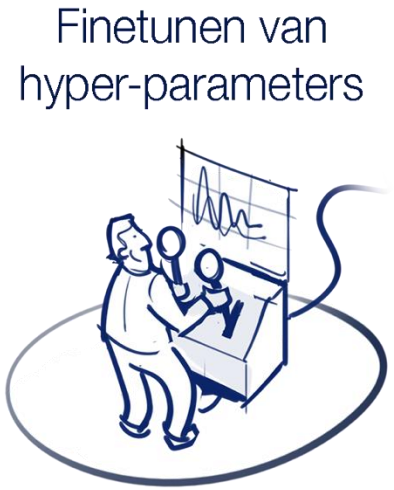
De Uitnodigingsregel: de methodiek zorgvuldig naar de praktijk brengen

Praktische producten voor inzet binnen jouw instelling

HOE ZET JE DIT PRODUCT IN?

Gebruik deze presentatie om de kwaliteit van het model te bepalen in twee fases:

- Dataverzameling en datakwaliteit
- Performance van het model



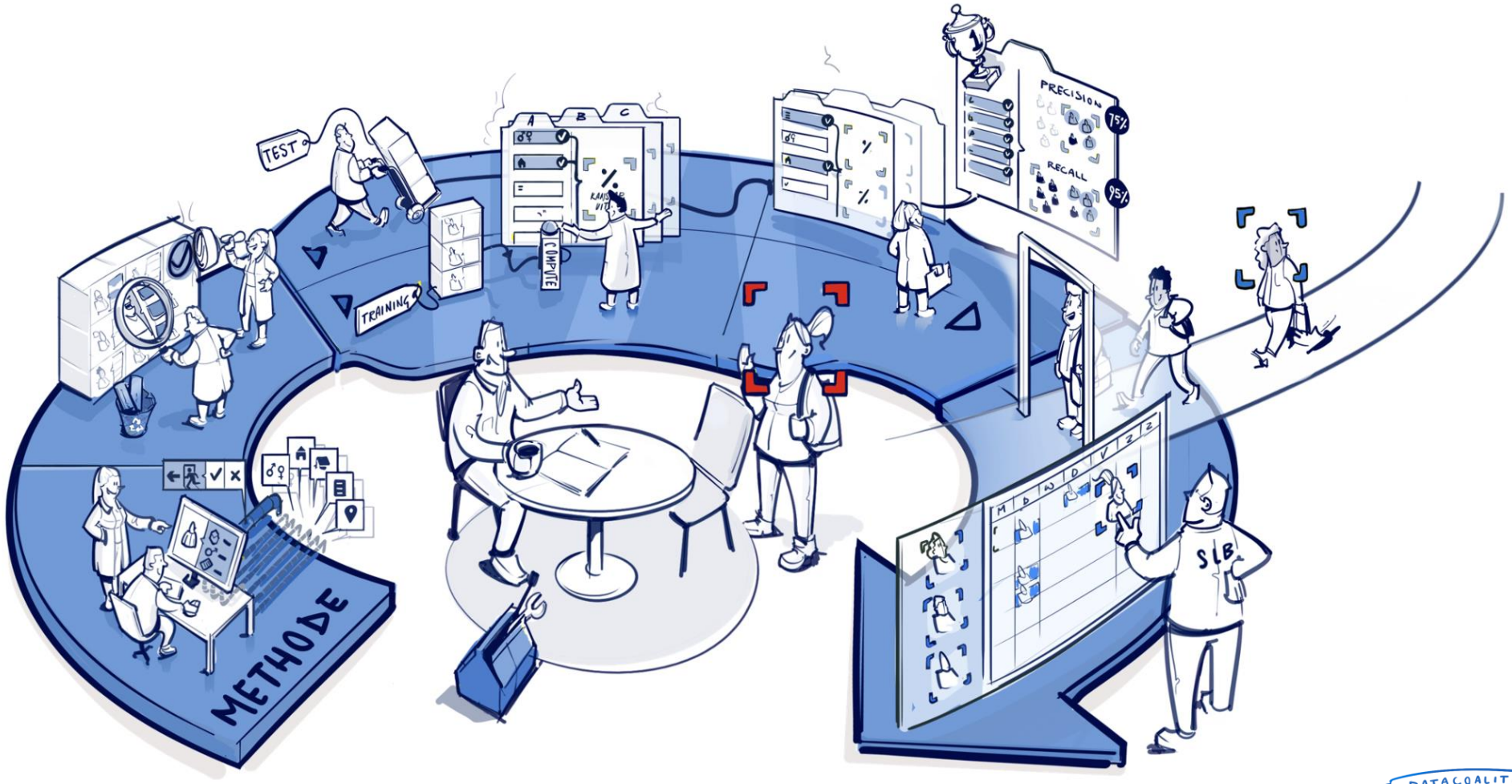
Andere beschikbare producten:

- Interventie overzicht
- DPIA
- Procesplaat
- Ethische hulpmiddelen
- Plan van aanpak
- Spelregels
- Kick-off presentatie

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.

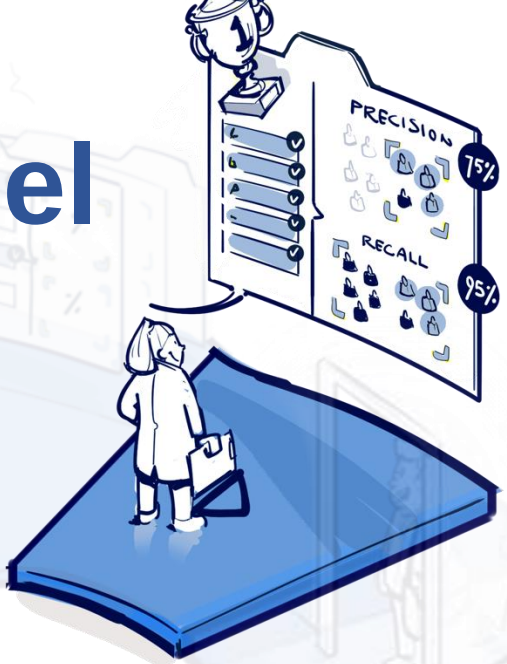
DATACOALITIE DGO MBO

De uitnodigingsregel-methodiek zorgvuldig naar de onderwijsteams brengen



DATACOALITIE
DGO
MBO

De kwaliteit van het model



De uitnodigingsregel is een voorspelmodel dat kijkt naar kans op uitval op basis van data die beschikbaar is over de student en zijn/haar leertraject.

Om een dergelijke methode in te kunnen zetten is het van belang om zeker te zijn van de kwaliteit van de voorspelling.

Dit doen we aan de hand van drie vragen:

- 1: Hebben we de benodigde data beschikbaar?
- 2: Is de kwaliteit van die data op orde?
- 3: Werkt het model goed genoeg voor ons doel?

DATACOALITIE DGO MBO

Dataverzameling en kwaliteit

Welke data wordt gebruikt?

Basisset (noodzakelijk):

- Studentkenmerken (leeftijd en geslacht)
- Basis vooropleidingsdata (historie en diploma's)
- Opleidingsdata
- Opleidingsnaam
- Niveau en leerweg
- Presentie / verzuimmeldingen op hoog niveau
- Behaalde resultaten (formatief en/of summa

Aanvullende data maakt het model kansrijker, bijvoorbeeld:

- Intake data
- Presentie / verzuimmeldingen op laag niveau
- Gedetailleerde vooropleidingsdata inclusief diploma & resultaten alle vakken
- Voortgang in de ELO

Extra data toevoegen

DATACOALITIE DGO MBO

Datakwaliteit - checks

Om de kwaliteit van de data te onderzoeken hanteren we de volgende stappen:

1. Datakwaliteitscheck uitvoeren

Grondige analyse van de beschikbare data op

- Volledigheid: Controleer of alle benodigde variabelen aanwezig zijn en of er geen belangrijke gegevens ontbreken (
- Accuraatheid: Onderzoek of de waarden logisch en correct zijn, bijvoorbeeld of de juiste eenheden worden gebruikt
- Consistentie: Controleer op dubbelingen en tegenstrijdigheden in de dataset

2. Dataprofiling en -visualisatie

- Gebruik dataprofiling tools om inzicht te krijgen in de structuur en kwaliteit van de databron
- Maak visualisaties van de data om patronen, uitschieters en mogelijke problemen te identificeren

3. Data cleaning en preprocessing

- Verwijder of corrigeer foutieve gegevens
- Handel ontbrekende waarden af op een geschikte manier (verwijderen, imputeren, etc.)
- Normaliseer of standaardiseer variabelen indien nodig

Voorbeeldcode is of komt beschikbaar via GitHub.

DATACOALITIE DGO MBO

Datakwaliteit - voorbereiding

Om de kwaliteit van de data te onderzoeken hanteren we de volgende stappen:

4. Feature engineering en -selectie

- Creëer nieuwe relevante features op basis van de bestaande data.
- Selecteer de meest informatieve features voor je predictiemodel.

5. Dataverdeling analyseren

- Onderzoek de verdeling van de doelvariabele en de predictoren.
- Controleer op class imbalance bij classificatieproblemen

6. Train-test split en cross-validatie

- Verdeel de data in train- en testsets om de voorspellingskracht te evalueren.
- Pas cross-validatie toe om de robuustheid van het model te testen

7. Baseline model bouwen

- Ontwikkel een eenvoudig baseline model om de prestaties te benchmarken
- Gebruik dit als uitgangspunt om te bepalen of de data voldoende voorspellende kracht heeft

Voorbeeldcode is of komt beschikbaar via GitHub.

DATACOALITIE DGO MBO

Fase 2: Kwaliteit van het model

De kwaliteit van het voorspelmodel meten we aan de hand van 2 concepten.

Dit wordt gedaan aan de hand van 2 vragen:

Blijken de voorspelde uitvallers ook werkelijk uit te vallen?

Precision (Precisie)

Hoe nauwkeurig is het model in het correct identificeren van uitvallers. Het is het percentage werkelijke uitvallers van alle voorpelde uitvallers. Hoge precision betekent dat als het systeem zegt "dit is een uitvaller", het meestal gelijk heeft.

Formule: Precision = (Correct geïdentificeerde uitvallers) / (Alle als uitvaller gelabelde studenten)

Hoeveel uitvallers hebben we "gemist"?

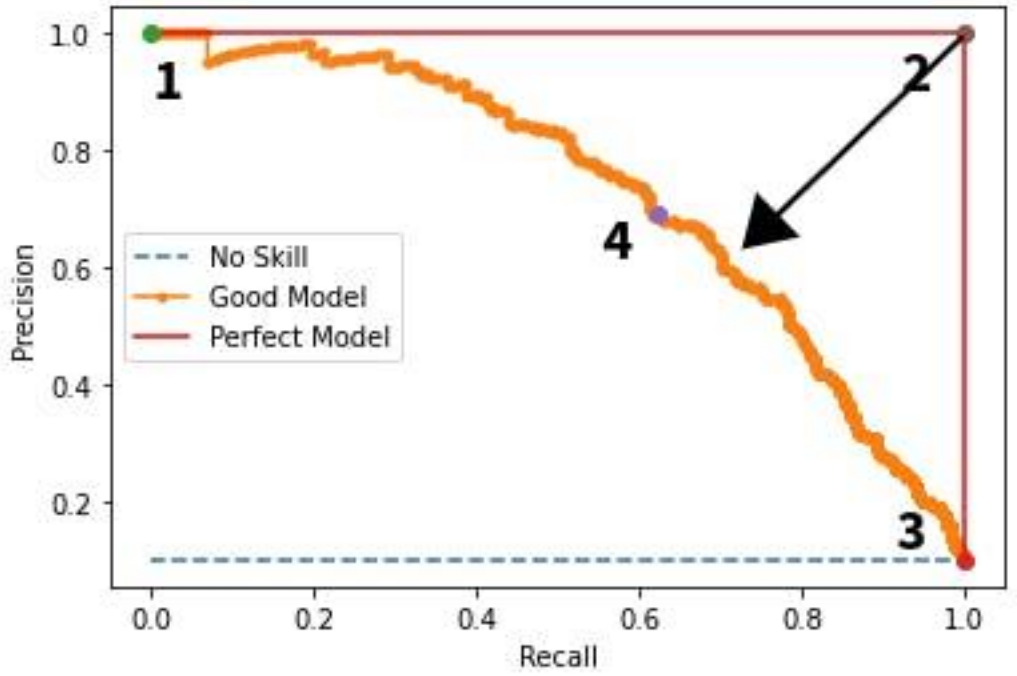
Recall (Volledigheid)

Recall meet hoe goed je systeem is in het vinden van alle echte uitvallers. Het is het percentage correct geïdentificeerde uitvallers van alle werkelijke uitvallers. Hoge recall betekent dat het model de meeste echte uitvallers vindt, ook al labelt het misschien per ongeluk wat andere studenten onterecht als uitvaller.

Formule: Recall = (Correct geïdentificeerde uitvallers) / (Alle echte uitvallers in de dataset)

Fase 2: Kwaliteit van het model

We hanteren een combinatie van precision en recall middels een curve.



Precision Recall Curve

We gebruiken als coalitie een combinatie van precision en recall om de kwaliteit van het model te bepalen.

Voordelen:	Nadelen:	Instellingsvervolg
• Uniform • Visueel en uitlegbaar • Combinatie van belangrijkste measures • Werkt bij ongebalanceerde sets	• Soms complex te interpreteren • Niet elk doel vraagt om dezelfde curve	• Beoogde (haalbare) curve opstellen