Hoe werkt data analyse: een kijkje in de black box

Welke stappen moet je doorlopen om te komen tot een goed model voor data analyse?

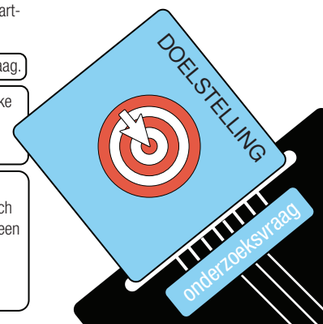
1 Doelstelling definiëren

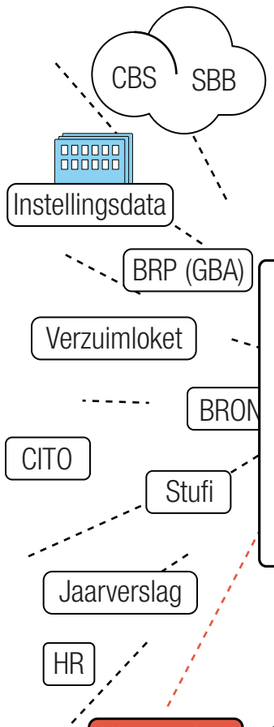
Het definiëren van een helder, haalbaar doel is het startpunt om de juiste data te verzamelen en analyseren.

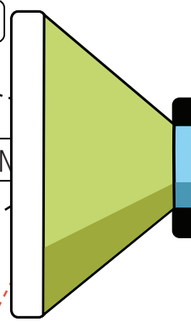
Formuleer een bijpassende, concrete onderzoeksvraag.

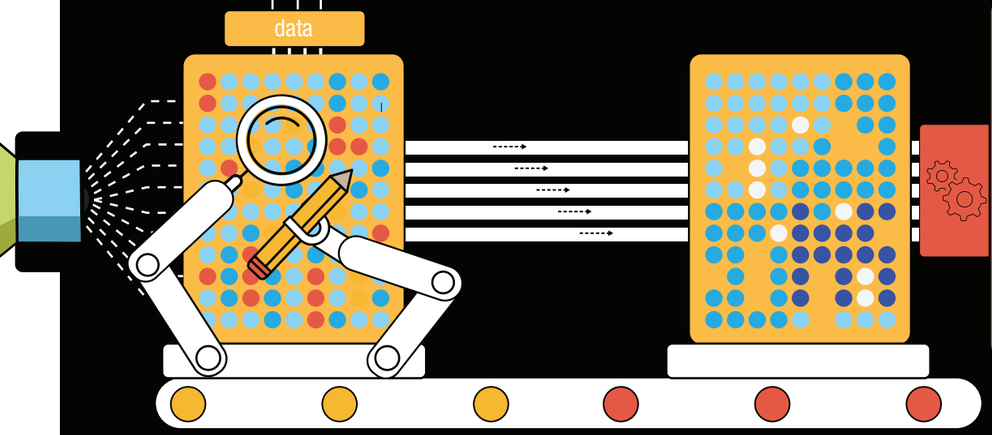
Voorbeeld: "Hoe ziet de instroompopulatie eruit, welke groepen herkennen we hierin en wat is de kans dat zij switchen in het eerste jaar?"

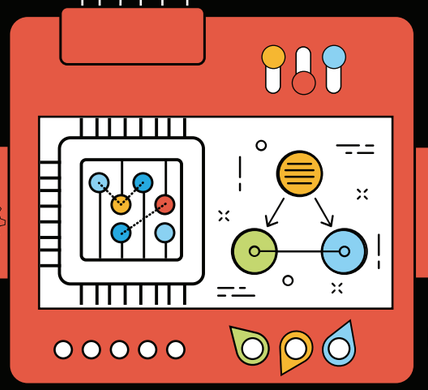
Belangrijk: stel met de betrokkenen definities vast voor wat je wilt onderzoeken. Bijvoorbeeld: met switch bedoelen wij het overstappen naar een opleiding in een ander opleidingsdomein binnen dezelfde instelling. Maak daarbij gebruik van bestaande definities waar mogelijk en onderbouw de keuzes die je maakt.

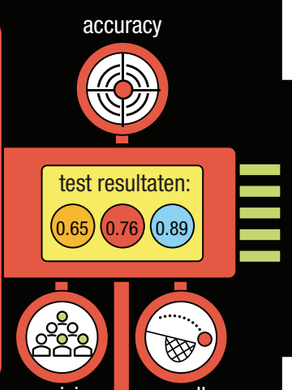


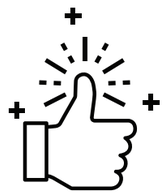












4 Selectie methode (=algoritme)

Er zijn verschillende methoden beschikbaar en welke je kiest wordt bepaald door wat je wilt onderzoeken. De geschiktheid van een algoritme is afhankelijk van type data en type vraagstelling. In de praktijk betekent het dat je middels experimenten zult ontdekken en testen wat de beste voorspellingen oplevert.

Welk soort algoritme past?
Afhankelijk van de vraag kies je voor een supervised of unsupervised algoritme. Dit bepaalt mede welke methode je kunt en wilt gebruiken.

Wat is het verschil?
Supervised: Het model leert om een bepaalde uitkomst zo goed mogelijk te voorspellen.
Unsupervised: Je werkt niet met een hypothese, maar het model gaat op zoek naar patronen in de data.

Model werkt nog niet goed genoeg!
Na het beoordelen van de kwaliteit van het model ga je op zoek welke verbeteringen mogelijk zijn. Misschien geef je andere of extra data mee aan het model, maak je andere clusters, prepareer je data anders of gebruik je een andere soort methode. Een lerend model verandert met elke iteratie.

2 Data verzamelen

Welke data heb je nodig om de vraag uit de doelstelling te kunnen beantwoorden?

Binnen DUO zijn er diverse bronnen die samen leiden tot het 1-cijferbestand. Dit kan dienen als basis voor de analyse, maar kan uiteraard aangevuld worden met extra data, bijvoorbeeld van de eigen instelling.

TIP: Beoordeel vooraf de kwaliteit en bruikbaarheid van de bronnen.

3 Data prepareren

Om data te kunnen gebruiken binnen een algoritme moet deze geschikt zijn of geschikt gemaakt worden.

Gebruikelijke procedures om data bruikbaar te maken zijn koppelen, verwijderen of vervangen van missende waarden, indelen in categorieën, controle voor outliers et cetera.

Sommige modellen werken beter als data op een specifieke manier is ingedeeld of voorbereid. Kijk dus goed welke voorwaarden methodes stellen aan de data-preparatie.

5 Een “lerend” model

Een model moet worden getraind en getest. Afhankelijk van de parameters die je meegeeft functioneert een model beter of slechter. Daarbij zoek je altijd naar een model dat niet alleen op de huidige dataset goede resultaten behaalt maar ook op nieuwe data.

Nieuwe data + testen verandert het model en is een iteratief (zichzelf herhalend) proces wat de werking van het model verbetert.

Effectiviteit

Hoe meet je hoe goed een model werkt?

De kwaliteit van het model kun je op verschillende manieren bepalen:
- *Accuracy* hoe vaak is het goed voorspeld?
- *Precision* hoe vaak komt de voorspelling overeen met de werkelijkheid?
- *Recall* hoeveel keer wordt een student gemist?

Uiteindelijk is het een afweging tussen inspanning (=kosten) en precisie. Gebruikers van het model moeten bepalen hoe het model ingezet kan worden om een zo groot mogelijk voordeel te behalen.

Resultaat

Afhankelijk van het soort onderzoek kunnen de resultaten ook uiteenlopen.

Bijvoorbeeld:
- Een indeling van studenten in onderscheidbare groepen.
- Een overzicht met kenmerken die een verband hebben met uitval in het eerste jaar.
- Een overzicht van alle historische patronen op basis van reeds afgestudeerde studenten.