# Datagedreven sturing in de praktijk

*Grote ambities, klein beginnen*

## Voorwoord

Voor u ligt geen onderzoek naar hoe Nederlandse gemeenten datagedreven sturing ontwikkelen. Dit is namelijk in 2018 al eens op een nauwgezette manier beschreven in het rapport “Datagedreven sturing bij gemeenten, van data tot (gedeelde) informatie voor beter (samen) sturen”, wat tot stand is gekomen in samenwerking met Berenschot.

Het document dat voor u ligt probeert wel een inkijkje te geven in de huidige praktijk rondom het onderwerp datagedreven sturen binnen de gemeente. Om meer specifieker te zijn, spitst dit stuk zich toe op het opzetten van een dataproject binnen een gemeentelijke context. Hoe ontstaat bijvoorbeeld een project waar je met AI aan de slag gaat? Waar moet je op letten tijdens het opzetten van een pilot? Waarom zou je überhaupt met data aan de slag gaan?

Aan de hand van een praktijkvoorbeeld uit de gemeente Arnhem zullen de verschillende facetten die samenhangen met het uitvoeren van een dataproject de revue passeren. Dit met als doel om de leefwerelden van de beleidsmatige (en uitvoerende) kant van een gemeente te verbinden aan de wereld van de data science. Het is dan ook geen toeval dat we hiervoor iemand die een behoorlijke tijd geacteerd heeft op dit snijvlak binnen gemeenteland hebben gevraagd om zijn ervaringen te delen, namelijk Bas Bloemberg. Bas was de afgelopen jaren bestuursadviseur op het onderwerp datagedreven innovatie binnen de gemeente Arnhem, hij is data analist en werkt sinds 1 september als consultant Data & AI bij Verdonck, Klooster en Associates (VKA).

## Introductie Bas Bloemberg

In een wereld met een toenemende hoeveelheid aan data, nieuwe analysetechnieken, groeiende proceskracht, digitalisering, goedkopere dataopslag en grote maatschappelijke uitdagingen (u kent het rijtje wel), is het niet vreemd dat de aandacht voor *big data* vaak grote belangstelling geniet.[[1]](#footnote-1) Algoritmen worden geacht in staat te zijn om patronen te herkennen in grote hoeveelheden en complexe data. Anderhalf jaar geleden heb ik mij laten omscholen tot data analist, om naast mijn adviserende taken binnen de gemeente Arnhem, ook een veel beter beeld te krijgen van *hoe* deze technieken nou precies werken.

Binnen de gemeente Arnhem heb ik een lange periode met veel plezier op het snijvlak van business en IT gewerkt. Als je beide werelden steeds beter leert begrijpen, zitten daar enorme voordelen aan. Als stad hebben we onze uitdagingen (of opgaven) enerzijds en technologische mogelijkheden anderzijds. Door die bij elkaar te brengen ontstaan er soms mooie dingen. Je komt er dan ook achter dat *data* één van de ingrediënten kán zijn.

Data en algoritmen kunnen absoluut een waardevolle bijdrage leveren, als ze maar op de juiste manier ingezet - en geïnterpreteerd - worden. Misschien is het in dit licht ook niet onverstandig om eerst te kijken naar datgene wat beantwoord moet worden en de ‘job-to-be-done’, en niet meteen naar de beschikbare data. Een goede data analist zal op basis van de gevonden patronen en trends waarschijnlijk ook eerder nieuwe vragen opwerpen in plaats van beantwoorden.

In deze bijdrage voor de VNG zal ik proberen om mijn ervaringen op het gebied van datagedreven sturing zowel vanuit de business als de techniek te beschrijven en uiteen te zetten. Ik zal in dit stuk eerst toelichten wat we in mijn optiek wel en niet kunnen met AI, om vervolgens in te gaan op een praktijkvoorbeeld uit Arnhem.

# Waar hebben we machine learning (AI) voor nodig?!



Laten we beginnen met een voorbeeld waar machine learning en AI totaal overbodig zouden zijn. Stel je even voor dat je een hardloopschema hebt gevonden op het internet, met daarin een schema voor 30 dagen. Je downloadt het schema en ziet vervolgens de volgende data:

***Dag 1 4 minuten***

***Dag 2 3 minuten***

***Dag 3 11 minuten***

***Dag 4 5 minuten***

***Dag 5 3 minuten***

***……***

We veronderstellen een relatie (aanname) tussen de dag waarop je gaat hardlopen en het aantal minuten - of zoals een data scientist zou zeggen “input features” en “output prediction”. Het lijkt onwaarschijnlijk dat het aantal minuten namelijk op een random manier in het schema geplaatst zijn, dus je vermoedt dat er iets van een relatie bestaat tussen deze twee. Stel dat je nieuwsgierig bent naar het aantal minuten dat je moet lopen op dag 2, wat doe je dan?

Ik hoop dat je gewoon het schema erbij pakt en het getal opzoekt. Eenvoudiger kan niet.

## Machine learning: het ontdekken van zinvolle patronen

Nu je over een hardloopschema beschikt voor 30 dagen is er geen noodzaak om naar een patroon te zoeken binnen dit tijdsbestek. Simpelweg omdat je alle data voor deze dagen tot je beschikking hebt.

***Dag 31 ? minuten***

Maar dan het volgende. Na 30 dagen ben je nog steeds gemotiveerd om te blijven lopen, alleen het schema eindigt op de 30e dag. Is er dan ook mogelijkheid om te voorspellen hoeveel minuten je op dag 31 zou moeten lopen?

Dat hangt ervan af.

Als er geen patroon te vinden is in jouw hardloopschema, gaat het onmogelijk worden om de input te koppelen aan de output prediction; dus je voorspelling. Dit wordt duidelijk als je het hebt over het gooien met een dobbelsteen. Je kunt 30 keer gooien, maar op basis van die worpen kun je onmogelijk iets zinnigs zeggen over worp 31. Anders wordt het als er wel degelijk een patroon te vinden is in de data: kunnen we de gemiddelde temperatuur schatten voor de zomermaanden op basis van de afgelopen zomers?

Wellicht.[[2]](#footnote-2)

## Machine learning: het ontdekken van zinvolle patronen

Als we in staat zijn om zinvolle patronen te vinden in de data, dan lukt het ons misschien(!) om een antwoord te vinden voor dag 31. Machine learning kan ons helpen om die patronen te ontdekken. Maar is dat dan het hele verhaal?

Helaas niet.

Om iets zinnigs over dag 31 te zeggen moeten we een belangrijke aanname doen. Namelijk dat de condities van de eerste 30 dagen niet zullen veranderen. Onze economie zou in 2020 met x % groeien, maar toch hebben we in 2020 te maken van een wereldwijde pandemie, waardoor er geen sprake meer is van economische groei. Voorspellingen maken op basis van gevonden patronen is één ding, maar of zij te generaliseren zijn is een tweede.

In ons voorbeeld kan het goed zijn dat dit schema overgaat in een heel ander type schema. Misschien met minder afwisseling, maar juist veel langere afstanden, met een aantal rustdagen, et cetera. Hoe goed de dataset misschien ook is, je hebt altijd te maken met grote onzekerheden.

*De hamvraag: hadden we dag 31 kunnen voorspellen?*

In dit geval wel. Onze data was niet random en we konden daarom wel degelijk een patroon vinden waarmee we het aantal minuten voor dag 31 hadden kunnen bepalen. Als we ons hardloopschema zouden visualiseren in R, dan krijgen we het volgende patroon te zien:



Figuur 1. Plot van onze data. Zie jij het patroon?

Benieuwd naar het aantal minuten voor dag 31?

**Dat is 9.**

En hoe weet ik dat zo zeker?

Hierbij de code die ik in R gebruikt heb om deze dataset te creëren:



Figuur 2. Functie in R om onze fictieve data te creëren.

In dit gedeelte heb ik op basis van een simpel voorbeeld laten zien, hoe we vanuit een patroon wellicht iets zinnigs kunnen zeggen over een toekomstige gebeurtenis. Methoden als machine learning omvatten verschillende algoritmen om deze patronen te vinden, om op basis daarvan een recept te maken dat in staat is om in een nieuwe situatie - met nieuwe data - een voorspelling te doen. Als er geen patroon te vinden is, dan kunnen we ook weinig zinnigs zeggen over de toekomst.

En als de wereld verandert? Dan wordt het nóg wat lastiger.[[3]](#footnote-3)

# Aan de slag met machine learning?!

En dan. Je zou graag willen weten hoe onze economie er morgen voor staat, hoeveel uitkeringen er aangevraagd gaan worden of welke bedrijven zullen vertrekken uit de regio. Maar je hebt alleen informatie van de afgelopen jaren tot je beschikking. Je hebt dus te maken met een grote onzekerheid, want kunnen we op basis van het verleden wel iets zinnigs zeggen over de toekomst? Dit brengt me bij een uitspraak van Cassie Kozyrkov (Head of Decision Intelligence at Google, 2018):

“It’s so annoying when we can’t remember the future, right?”

Hieruit destilleren we de volgende vergelijking: data + assumpties = voorspelling. De toekomst is niet met zekerheid te voorspellen. We staan bijna altijd zwak als we beweringen over de toekomst doen.[[4]](#footnote-4) Simpelweg omdat we in heel veel situaties niet alle feiten hebben, maar ook omdat we niet zeker zijn dat alle informatie die we hebben ook juist is of omdat bepaalde patronen die zich voorgedaan hebben de afgelopen jaren ineens aan het veranderen zijn. Dat we met een coronacrisis te maken hebben terwijl onze economie toch zou groeien met 3 procent? We moeten dus verschillende aannames doen, om het gat tussen wat we weten en wat we zouden willen weten, te dichten.

## Zonder assumpties geen voorspelling

Maar een wereld zonder aannames bestaat niet. Achter elke beslissing zitten aannames over hoe zaken in elkaar steken. Als je een krant leest ga je ervan uit dat veel feiten wel gecontroleerd zullen zijn. De vakantie die je wellicht hebt moeten annuleren voor de zomer van 2020, is waarschijnlijk geboekt met de gedachten dat vliegen nog steeds tot de mogelijkheden zou behoren en we niet met een wereldwijde pandemie te maken zouden hebben. Aannames zijn een wezenlijk onderdeel van het maken van een beslissing en dus onvermijdelijk.

Bewustzijn van de aannames die je maakt, is dus cruciaal in het doen van een data-analyse. Ontelbaar veel methoden kunnen met relatief weinig inspanningen ingezet worden om met jouw eigen data (en aannames) tot een resultaat te komen.[[5]](#footnote-5) Dit is precies het punt waar de “magic happens”. Het is ook een punt waar je met gezond wantrouwen naar moet kijken, omdat elke keuze in het proces kan leiden tot een ander resultaat. De tijden waarin we nu leven laten volgens mij ook heel goed zien dat het onverwachte en onvoorstelbare niet langer uitgesloten kan worden.

# Introductie van de casus: “Van bedrijven weten”

De gemeente Arnhem wil de dienstverlening aan ondernemers verbeteren. Met ruim 15.000 ondernemers is Arnhem een ondernemende stad. De dienstverlening voor de Arnhemse ondernemers gaat met dit programma verder geprofessionaliseerd worden. Om dit te bereiken heeft de gemeente Arnhem een team van accountmanagers, die de spil zijn in deze versterkte vorm van de gemeentelijke ondernemersdienstverlening. Het team gaat in kaart brengen waar de behoeften van de bedrijven liggen en of er binnen de gemeente zaken anders georganiseerd of geregeld moeten worden om ondernemers nog beter te ontzorgen en de werkgelegenheid in de regio te stimuleren.

## Let’s begin!

Om iets meer gevoel te krijgen bij de verschillende fasen in een AI-project, heb ik de volgende visualisatie toegevoegd. Dit zijn doorgaans de stappen die je zet in een AI-project:



Bron: Verdonck, Klooster en Associates (2020)

## Define your objectives

Laten we beginnen met stap één: het definiëren van je vraag. In Arnhem zijn we met twee acquisiteurs begonnen om te onderzoeken wat datagedreven sturing zou kunnen betekenen voor het acquisitie- en accountmanagement. Voor gemeenten is de lokale werkgelegenheid belangrijk en daarmee dus ook de aanwezige bedrijvigheid in de regio. De vraag die hieruit voortvloeide was dan ook of we kunnen ‘voorspellen’ of een bedrijf de regio dreigt te verlaten? Dit bleek uiteindelijk de kernvraag te zijn van onze acquisiteurs, om daarmee proactief te kunnen handelen in plaats van reactief. Warme acquisitie zou daarmee een extra dimensie kunnen krijgen, als we op voorhand kunnen bepalen of een bedrijf een hoog risico heeft om de regio te verlaten. Bovendien is het bezoeken van alle 15.000 bedrijven niet bijzonder efficiënt. Om hier meer richting aan te geven, kan het helpen om jezelf te laten inspireren door de data.

Voordat we meteen de data induiken kan het raadzaam zijn om eerst eens te onderzoeken of er (wetenschappelijke) onderzoeken liggen op dit vlak. En dat bleek het geval te zijn. Via een docent van de Radboud Universiteit kwam ik bij een artikel terecht dat de belangrijkste variabelen beschrijft op het gebied van verhuisbewegingen. Hier kwamen kenmerken als bedrijfsgrootte, sector, bereikbaarheid en bedrijfsgroei als belangrijke ‘voorspellers’ naar voren en boden meteen een denkrichting voor het maken van een eerste model en het verzamelen van onze data.

Dan de belangrijkste stap in het definiëren van je doelstelling: wat wordt onze y-variabel?[[6]](#footnote-6) Of in meer begrijpelijkere taal, wát willen we precies voorspellen? Kijken we naar alle bedrijven die jaarlijks vertrekken? Vinden we zzp’ers eigenlijk niet zo interessant? En vinden we alle sectoren even belangrijk? Het is erg belangrijk om vooraf goed helder te hebben waar je precies op wilt sturen. Want zoals al eerder gezegd, elke keuze in dit proces kan leiden tot een totaal andere uitkomst.

## Get access to data

Omdat we weinig praktijkvoorbeelden vonden bij andere gemeenten op dit onderwerp, besloten we om eerst eens te inventariseren hoe het datalandschap op economisch terrein er eigenlijk uitzag.[[7]](#footnote-7) De provincie Gelderland houdt in haar provinciale werkgelegenheid-enquête o.a. bij welke bedrijven zich vestigen in de regio, hoeveel uitzendkrachten een bedrijf heeft, wat de man/vrouw verhouding in een bedrijf is, hoe groot een bedrijf is en wat de perceeloppervlakte is. Dit heeft zij de afgelopen jaren minutieus bijgehouden, waardoor de mutaties van de afgelopen jaren mooi inzichtelijk zijn en de data-kwaliteit behoorlijk goed is.

Verder publiceert het CBS elk jaar een grote hoeveelheid aan data die iets zeggen over kenmerken van buurten, wijken, bewoners et cetera. Hiermee valt te onderzoeken of bijvoorbeeld het vertrek van een bedrijf wellicht samenhangt met de bereikbaarheid van het gebied. Als laatste vonden we op economisch vlak ook bronnen als Jobsfeed, DUO en makelaarsdata die iets zeggen over de vastgoedprijzen in de regio. Deze laatste bron heb ik handmatig met *webscraping* binnengehaald als experiment, om deze data later toe te voegen aan het model.[[8]](#footnote-8)

Belangrijk om hier te vermelden is dat niet alle data op hetzelfde niveau ontsloten is. Data van het CBS is verkrijgbaar op bijvoorbeeld buurtniveau, de enquête vanuit de provincie Gelderland op bedrijfsniveau en de vastgoeddata op straatniveau. Bronnen als Jobsfeed leveren in eerste instantie soms interessante inzichten op, maar de koppeling met de eerder genoemde bronnen lijkt schier onmogelijk. Verschillende data-bewerkingen zijn vervolgens nodig om alles te koppelen, en van alle stappen in zo’n project, is dit waarschijnlijk één van de meest arbeidsintensieve klussen.

## Split your data

Een onmisbare stap in het toepassen van machine learning is het splitsen van de data in een training en test gedeelte - normaal gesproken komt daar ook een validatie set bij, maar om het enigszins begrijpelijk te houden laten we het in dit voorbeeld even bij deze twee datasets. We splitsen de data vanwege de volgende reden. Je maakt een model omdat je hoopt dat dit model *out of sample,* of makkelijker gezegd, buiten de trainingsomgeving, ook goed zal presteren. Een analogie om de functie hiervan wat beter uit te leggen is wellicht de volgende: studenten die leren voor een examen moeten eerst een berg informatie doorspitten om de hoofdlijnen te destilleren. De kennis die zij daarmee vergaren kan vervolgens worden toegepast in een examen, waar vragen gesteld worden die niet exact hetzelfde zijn als waarvoor ze geleerd hebben. Wel is het zo dat deze vragen over het algemeen de lading dekken en representatief zijn voor datgene wat ze geleerd hebben. Bij machine learning is het niet veel anders. We splitten de data in verschillende delen, trainen ons model op een gedeelte hiervan en testen vervolgens of het model op een ander gedeelte nog goed presteert. De testfase binnen machine learning is daarmee dus eigenlijk een eindexamen, waarin wordt bepaald op basis van een aantal criteria hoe goed het model presteert. In het voorbeeld van wegtrekkende bedrijven moet het model dus de vetrokken bedrijven op een juiste manier identificeren in de testomgeving.

## Train & test some models

Nadat alle data netjes is gekoppeld, opgeschoond, opgedeeld en gestructureerd zijn, kunnen we aan de leukste fase van een project beginnen: het bouwen van een model! Er zijn tegenwoordig ontelbaar veel machine learning algoritmen, waarbij er relatief weinig code gebruikt hoeft te worden om die toe te passen. In het vorige gedeelte heb ik al laten zien dat het vaak gaat om het vinden van een patroon in de data die we kunnen extrapoleren naar de toekomst. Om te weten of die gevonden patronen werken in een ‘echte’ omgeving, hebben we dus wat data apart gehouden voor de testfase.

Op voorhand is het niet te zeggen welk algoritme het beste gaat presteren, of zoals mijn docenten op de Erasmus Universiteit zeiden: “There ain't no such thing as a free lunch.” In de praktijk komt het er op neer dat het maken van een model veel proberen, tunen en testen is. Veel algoritmen kennen weer individuele knoppen c.q. parameters waar je aan kan draaien, om daarmee een poging te doen om de prestaties te verbeteren.[[9]](#footnote-9) Die prestaties c.q. voorspellingen van het model worden dan uiteindelijk afgezet tegen de werkelijke waarden van een testset, waarmee je uiteindelijk de volgende tabel (ook wel confusion matrix genoemd) krijgt:



Figuur 3. Confusion matrix voor het beoordelen van een ML-model.

Als data-analist beoordeel je machine-learning-modellen vaak met bovenstaande confusion matrix. Het is een geaccepteerde manier om te kijken hoe goed een model eigenlijk presteert. Zoals je in dit voorbeeld kan zien worden de voorspelde waarden altijd afgezet tegenover de werkelijke waarden. Verder zijn er veel manieren om de nauwkeurigheid van een model te bepalen op basis van deze matrix, maar het past helaas niet in de scope van dit artikel om daar veel dieper op in te gaan.

Als je genoeg vertrouwen hebt in een model dat je hebt kunnen trainen op historische data, kan het een mooi experiment zijn om dit model toe te passen op het huidige bedrijvenbestand in de regio. Op deze manier komt er een shortlist van bedrijven naar voren waarvan het model vermoedt dat deze bedrijven zullen vertrekken. Hiermee ontstaat een nieuwe vorm van acquisitie- en accountmanagement, namelijk een datagedreven vorm waarbij je je laat inspireren door de data. Als de wereld niet verandert, dan kunnen patronen uit het verleden wellicht iets zinnigs zeggen over de toekomst.

## Afsluitend

In het Arnhemse voorbeeld kwam er uiteindelijk een model naar voren dat redelijk goed in staat bleek om de vertrokken bedrijven in onze test-omgeving ook als zodanig te identificeren. Dan rijst natuurlijk de vraag: wat ga je met dit model doen?

Mijn reflectie op dit experiment is dat het absoluut een bijdrage kan leveren in termen van efficiëntie. Als accountmanager bezoek je niet alle bedrijven – wat trouwens ook een schier onmogelijke taak zou zijn - , maar richt je je op de bedrijven die volgens het model de grootste kans hebben om te vertrekken. Ook biedt het wellicht handvatten om te kijken naar de factoren waar je als overheid wellicht invloed op hebt, zoals bereikbaarheid van een locatie, lopende huurcontracten, subsidiemogelijkheden, et cetera.

Grote nadelen die ik zag rondom dit project lagen met name op het gebied van het beheer. De databronnen van het CBS en de provincie Gelderland zijn jaarlijks beschikbaar en kennen een hoge kwaliteit, maar de aanvullende bronnen (zoals prijsontwikkelingen) waren erg lastig om binnen te halen en te koppelen. Dat maakt het erg ingewikkeld en tijdsintensief om dit model echt naar productie te tillen. Toch sluit ik niet uit dat dit een mogelijkheid voor de toekomst kan zijn. Als er afspraken gemaakt worden met partners uit de regio, er data onderling gedeeld kan worden en er overlap zit in de verschillende belangen, dan is het absoluut levensvatbaar.

## Lessons learned

1. Houdt het klein.

De kloof tussen wat mensen denken dat technisch mogelijk is en wat er daadwerkelijk kan is soms best groot: “Facebook en Google weten toch ook alles?”, heb ik weleens in een vergadering gehoord. Goed verwachtingsmanagement en het nemen van kleine stappen zijn cruciaal om draagvlak te behouden voor ontwikkelingen als datagedreven sturing. Die *druk op de knop* is vaak niet realistisch. Het is trial-and-error. Alleen al het visualiseren en beschrijven van beschikbare data geven soms al mooie inzichten. Daar heb je geen machine learning voor nodig. Het gaat niet om de verschillende algoritmen of het aantal ‘voorspelmodellen’, maar veeleer dat je begrijpt dat data een meerwaarde kan bieden voor de business in het nemen van besluiten.

1. Analytics ≠ statistics

Data analytics en statistiek kunnen in mijn ogen worden beschouwd als twee takken van dezelfde boom: data science. Ze delen voor een groot gedeelte dezelfde helden en grondbeginselen. Eenvoudig een grens trekken om beide werelden te scheiden lijkt hier wat lastig. Maar waar in het algemeen de focus binnen analytics meer lijkt te liggen op het *exploreren* van de data, richt de statistiek zich meer op de conclusies die wel of niet getrokken kunnen worden op basis van de data. Het lijkt erop dat beide rollen elkaar niet altijd begrijpen in de praktijk.[[10]](#footnote-10) Mijn ervaringen binnen de overheid zijn dat er veel (goede) statistici rondlopen, maar nog weinig data-analisten. Iets wat een ontwikkeling als datagedreven sturing binnen de overheid wellicht nog wat in de weg kan zitten als het gaat over het potentieel dat benut kan worden. Mijn advies op dit vlak zou dan ook zijn om actief de opleidingsmogelijkheden te bekijken van het eigen personeel, alsook om openstaande vacatures goed te laten aansluiten bij de ontwikkelingen in de markt: heeft een onderzoeker bijvoorbeeld ook ervaring met programmeertalen zoals R of Python?

1. Laat je inspireren door data!

Datagedreven werken lijkt niet zelden geïntroduceerd te worden als een manier van werken die vrij is van subjectieve waarden. Het zou *decision makers* in staat stellen om beslissingen te nemen op een neutrale en objectieve manier. Maar in veel gevallen is zo’n beslissing al genomen en zoeken we bevestiging voor datgene wat we toch al vermoeden. Daarbij zijn data vaak allesbehalve neutraal en objectief; zaken worden bewust wel of niet geregistreerd. In de psychologie wordt dit ook wel de *confirmation bias* genoemd. Wat vrij vertaald neerkomt op de neiging om informatie te zoeken waarmee heersende overtuigingen bevestigd worden. Ik denk dat dit voor het onderwerp datagedreven werken (binnen de publieke sector) ook zeker van toepassing is. We laten ons vaak inspireren door data, maar het nemen van een beslissing hangt van zoveel andere factoren af. Iets waar je als organisatie goed over na moet denken, want voor je het weet eindig je met een uiterst dure hobby.[[11]](#footnote-11)

1. Doe het zelf!

Zorg ervoor dat je als organisatie zelf de kennis in huis hebt om te experimenteren op het gebied van data. De enorme hoeveelheid aan keuzemogelijkheden in een project, de daaruit voortvloeiende aanpak, het grote aantal analysetechnieken, de verschillende ingrediënten die nodig zijn en - niet op de laatste plaats - de onzekere uitkomsten, zorgen ervoor dat je echt zelf de expertise in huis moet hebben om aan de slag te gaan. Het kan anders een erg kostbare exercitie worden als je als gemeente afhankelijk wordt van externe partijen op dit vlak. Maar nog belangrijker, ook de opgedane kennis in een project verdwijnt dan vaak als een project is afgerond. Zonde!

1. Goed om te vermelden dat veel ‘nieuwe’ analysetechnieken, waaronder populaire machine learning algoritmen, al decennia geleden zijn bedacht. Zo dateert het bekende Random Forest algoritme, bedacht door Tin Kam Ho, uit 1995. [↑](#footnote-ref-1)
2. In data science termen wordt hier gesproken over (non-)stationary data. Met stationary wordt bedoeld dat de eigenschappen van de data onafhankelijk zijn van de tijd. [↑](#footnote-ref-2)
3. Een niet onbelangrijk aspect om in het kader van ‘voorspellingen’ te noemen is het volgende. We kunnen namelijk met toenemende hoeveelheden aan data steeds vaker correlerende variabelen vinden, waardoor we denken een goede voorspeling te kunnen doen. Het gevaar is dat je niet weet of er ook een causale relatie aanwezig is en je voorspellingen sec gebaseerd zijn op correlaties. [↑](#footnote-ref-3)
4. Geciteerd uit “Bureaucratie is een inktvis” van René ten Bos (2015) [↑](#footnote-ref-4)
5. Niet voor niets worden een aantal bekende algoritmen binnen de machine learning (afgekort ML) “decision trees” genoemd. Dit gaat deels over de manier waarop een algoritme werkt, maar het geeft ook de indruk dat je een beslissing kan nemen op basis van de output. [↑](#footnote-ref-5)
6. Veel vragen binnen een data science project moeten vaak omgezet en platgeslagen worden naar een 0 of een 1. [↑](#footnote-ref-6)
7. Deze verkenning hebben we o.a. met Berenschot Intellerts gedaan. [↑](#footnote-ref-7)
8. Webscraping is een computertechniek waarbij software wordt gebruikt om informatie van webpagina's te extraheren en al dan niet te analyseren. [↑](#footnote-ref-8)
9. Er zijn veel verschillende manieren op de prestaties van een model te verbeteren. Naast de keuzen voor een goed algoritme kun je bijvoorbeeld ook denken aan het creëren van nieuwe variabelen, dit wordt ook wel feature engineering genoemd. Een mooi voorbeeld is BMI, wat een combinatie is van lengte en gewicht. [↑](#footnote-ref-9)
10. Voorbeeld uit de praktijk: waarom hoeft er binnen machine learning niet altijd getoetst te worden op multicollineariteit? [↑](#footnote-ref-10)
11. Bron:ww. blog.upscope.io [↑](#footnote-ref-11)