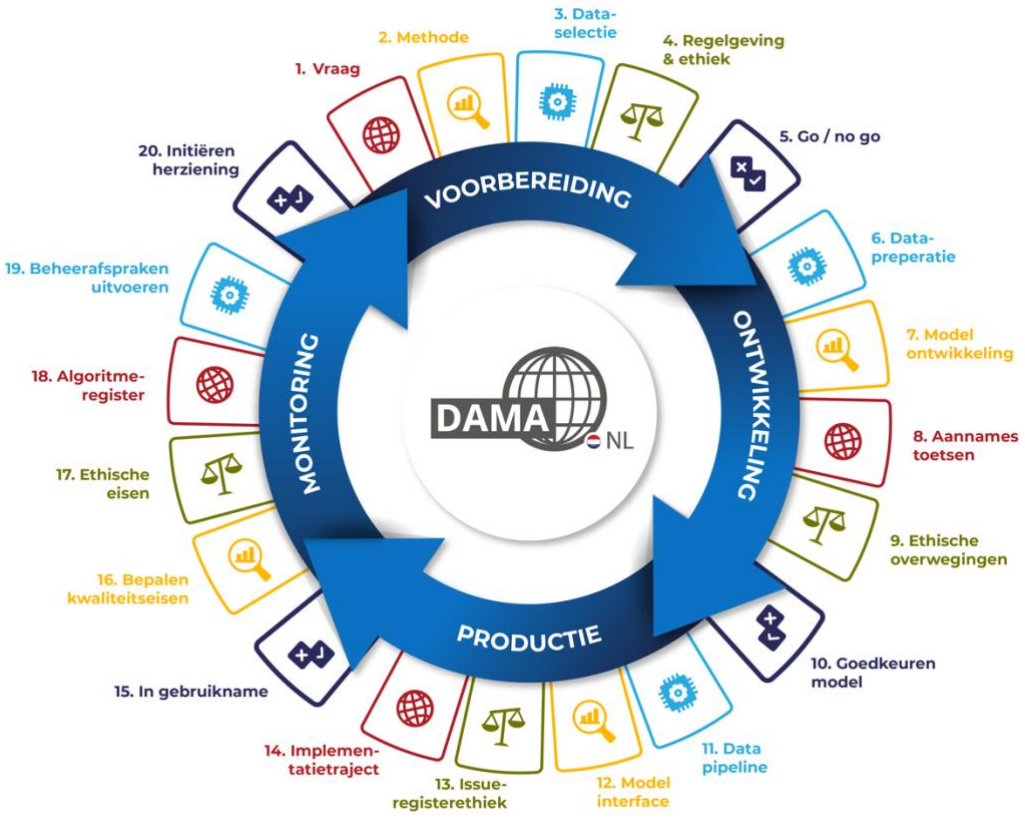
Data Science Governance Framework

EEN PRAKTISCH RAAMWERK VOOR HET BESTUREN VAN DATA SCIENCE BINNEN ORGANISATIES

DAMA NL WERKGROEP DATA SCIENCE GOVERNANCE

DATUM: 1-9-2023

VERSIE 2023.2



Toelichting veelgebruikte termen

Data Science is het gebruik van datamining, statistische analyse en machine learning samen met data-integratie en datamodellering technieken om voorspellende en voorschrijvende modellen te bouwen (DAMA DMBOK, 2017).

Data Science Governance wordt gezien als het proces om ervoor te zorgen dat Data Science wordt toegepast volgens het interne beleid, externe beleidskader, wet en overheidsrichtlijnen, en best practices. Data Science Governance gaat om het uitoefenen van controle over het toepassen van Data Science binnen de organisaties.

Voorspellende en voorschrijvende modellen worden in de praktijk vaak ”algoritmes“ genoemd. In dit document worden de termen ”algoritme” en ”model” uitwisselbaar gebruikt om het model dat in het kader van het toepassen van Data Science wordt ontwikkeld (al dan niet met behulp van algoritmen) te aanduiden.

1. Het DAMA Data Science Governance Framework

Het praktische DAMA Data Science Governance Framework biedt een handvat voor organisaties om verschillende beleidskaders en best practices voor het toepassen van Data Science om te zetten naar een werkbaar proces en dit aangesloten te houden op veranderende organisatiebehoeften. Het beschrijft in stappen hoe grip te houden op de ontwikkeling, ingebruikname, evaluatie en beheer van algoritmen, waarbij ethische aspecten per geval worden afgewogen met behoud van overzicht op complexe situaties.

In dit document beschrijven we achtereenvolgens het Data Science Governance Framework, activiteiten die worden verricht, betrokken rollen, de vertaling naar een aantal overheidsrichtlijnen (verantwoording), en tips voor gebruik.

1.1 Uitdagingen

De aandacht voor de impact van algoritmen en artificiële intelligentie (AI) op de samenleving is de laatste jaren sterk toegenomen. Met name de risico's die deze toepassingen kunnen meebrengen staan terecht hoog op de politieke en maatschappelijke agenda[1].

Inmiddels zijn er meerdere beleidskaders en instrumenten beschikbaar die organisaties kunnen helpen bij het omgaan met algoritmen. Maar hoe wordt dit in de praktijk geïmplementeerd? En hoe zorgen organisaties ervoor dat modellen niet alleen worden ontwikkeld, maar ook goed gebruikt kunnen worden en worden geëvalueerd?

Organisaties willen data-innovaties en algoritmen omarmen om betere diensten en producten aan te bieden, maar lijken te stagneren in de uitvoering door o.a. gebrek aan publiek vertrouwen, risico-aversie en complexiteit bij het toepassen van data science. Verder moeten ze ook rekening houden met ontwikkelingen in de samenleving en toenemende regelgeving om data- en algoritme op een ethische manier te gebruiken.

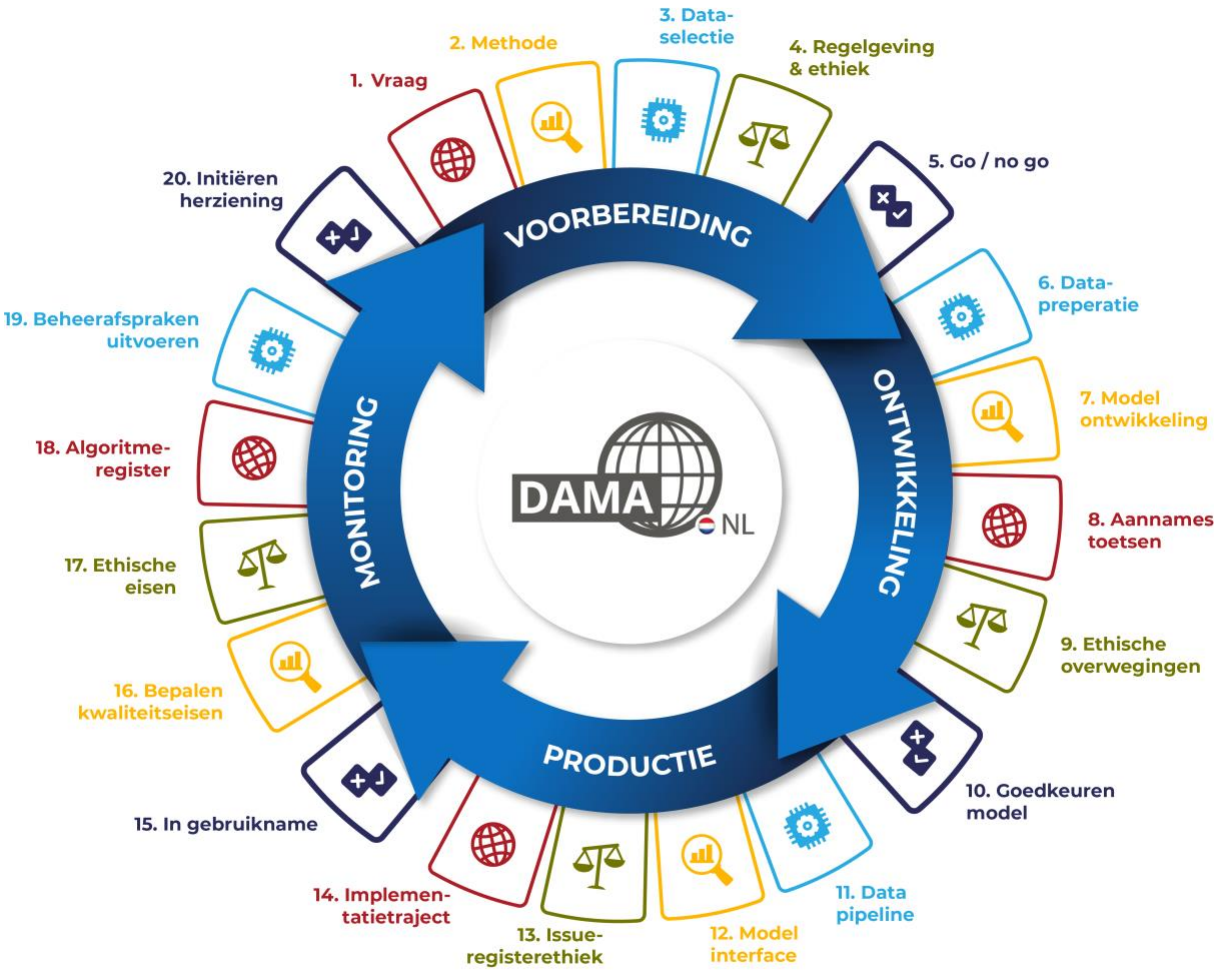
Om het vertrouwen in het toepassen van data science te vergroten, is het noodzakelijk het gebruik van gegevens transparant te maken en ethische besluitvorming te verankeren in de gehele data-waardeketen. Zonder een multidisciplinaire aanpak is het echter moeilijk om risico's tijdig te identificeren en end-to-end transparantie te garanderen.

1.2 Uitgangspunten

Omdat de meeste uitdagingen zich in de praktijk richten tot voorspellende en voorschrijvende modellen, is dit ook het primaire aandachtsgebied voor het Data Science Governance Framework. Het toepassen van data science vereist op hoofdlijnen 4 fases: voorbereiding, ontwikkeling, productie, monitoring. Elke fase benaderen we vanuit 5 disciplines: domein, data, analyse, ethiek[2] en besturing. Elke fase eindigt met een besluitvormingsstap, waardoor er kan worden begonnen aan de volgende fase.

[1] Kamerbrief, Tweede Kamer, vergaderjaar 2020-2021, 2643, nr. 765, p.1

[2] Onder ethiek wordt ook privacy en beveiliging ondergebracht. Waar 'randvoorwaarden' misschien een betere term is, kiest de werkgroep er nu voor om het invalshoek ethiek te noemen. Dat heeft te maken met de 'vergrootglas' op ethische vraagstukken.



Figuur 1.-1: Data Science Governance Framework

Het Figuur 1.1 duidt aan dat het Data Science Governance Framework volgt geen lineair proces met een vast eindpunt. Data Science Governance is een voortdurend cyclisch proces waarbij de vier fases en de bijbehorende stappen van het raamwerk gevolgd worden voor zowel de modellen die de organisatie al in gebruik heeft als de modellen die nog ontwikkeld worden.

Voor het leesgemak hebben we voor elke discipline een kleur en pictogram ontwikkeld, die terug te vinden zijn in de rest van de tekst.


Domein


Analyse


Data


Ethiek


Besturing

2. Activiteiten binnen het raamwerk

2.1 Voorbereiding

In een ideale situatie heeft een organisatie al verkennende data-analyses uitgevoerd en is er sprake van besluitvorming op basis van datagedreven inzichten. Op het moment dat er sprake is van data-analyses die verder gaan dan het beschrijven en visualiseren van de huidige situatie en/of het verleden, begint het Data Science Governance Framework.

2.1.1 Vraag (stap 1)

Het stellen van een goede vraag is essentieel voor het slagen van een project. In deze stap wordt een nulsituatie geschetst waardoor bekend is welk niveau het model minimaal moet halen om beter te werken dan de huidige situatie. Beschrijf hier wat de 'businessvraag' en de context is. Bij de context gaat het ook over een beschrijving van de huidige manier van werken, de gewenste verbetering, en de mogelijke alternatieven (incl. het aanpassen van een proces/procedure zonder het ontwikkelen van een voorspellend of voorschrijvend model) om deze verbetering te behalen. Beschrijf ook het doel van het analyseproduct.

2.1.2 Methode (stap 2)

Als er eenmaal een goede vraag is, kan een data scientist herleiden om wat voor soort analyse het gaat. Is het bijvoorbeeld een regressievraag, classificatievraag of een vorm van clustering? Er zijn meerdere analysemogelijkheden. Mogelijke beperkingen en voor- en nadelen van de verschillende analysemethoden worden beschreven en vergeleken. Waar mogelijk heeft het gebruik maken van simpele algoritmen de voorkeur boven complexe algoritmen. Tijdens de ontwikkeling fase wordt dit vervolgens uitgevoerd en blijkt welk algoritme het beste gebruikt kan worden.

2.1.3 Dataselectie (stap 3)

Voordat data wordt ontsloten naar een analyseomgeving, moet er bepaald worden welke data nodig is voor de analyse, en uit welke bronnen dit komt. Goede metadata is hiervoor essentieel. Het is belangrijk om als organisatie beschreven te hebben welke data er is, waar die data beschikbaar is, wat het betekent en wat de kwaliteit van de data is. Streven is een minimale set van data te selecteren. Minimalisatie vindt echter ook plaats tijdens de ontwikkeling van een model, door hetzelfde model te hertrainen met minder features.

2.1.4 Regelgeving en ethiek (stap 4)

Voor de AVG, art 5 en 6, is het van belang om te weten welke vraag er is, met welk doel de analyse plaatsvindt, en of datagebruik proportioneel is. Er zijn ook vereisten rondom juistheid en maatregelen (technisch en organisatorisch). Daarom vindt er een privacy toets (DPIA) en advies plaats. Ook de wenselijkheid van de analyse komt ter sprake, door een data ethicus die voor het eerst toetst op mogelijke discriminatie of stigmatisering. Ook wordt onderzocht of betrokkenen de mogelijkheid hebben om niet

onderworpen te worden aan geautomatiseerde besluitvorming. De data ethicus kan vrij van een data-analyse team opereren.

Er zijn verschillende methoden om stil te staan bij ethiek. De Nederlandse AI Coalitie heeft er 7 beschreven in theorie en praktijk[3]:

- Aanpak begeleidingsethiek (ABE)
- De ethische data assistent (DEDA)
- Impact Assessment mensenrechten en algoritmes (IAMA)
- Technology impact cycle tool (TICT)
- Data Governance Clinics (DGC)
- Assessment list for trustworthy AI (ALTAI)
- Data protection impact assessment (DPIA)

2.1.5 Go/ No Go (stap 5)

De voorgaande stappen worden vastgesteld in een voorstel. Hierdoor is het voor het management mogelijk om een akkoord te geven voor de start van het project. Indien nodig wordt op basis van een projectportfolio prioriteiten vastgesteld voor de uitvoering van verschillende projecten. In deze stap wordt het eigenaarschap van het model bepaald. Op het moment dat er meerdere partijen samenwerken, wordt bepaald welke partij de regie neemt en verantwoordelijk is voor het doorlopen van een goed proces en het inbedden van de juiste waarborgen.

2.2 Ontwikkeling

Tijdens de ontwikkeling fase wordt de stap gezet van denken over een model naar het ontwikkelen van een model.

2.2.1 Datapreparatie (stap 6)

De goedgekeurde data wordt verzameld en geplaatst in een analyseomgeving. De toegang tot de omgeving wordt ingericht. Hier wordt een dataset gemaakt dat gebruikt kan worden om een model te trainen. Deze set wordt onderverdeeld in aparte trainings-, test en validatiesets. Er vindt ook een datakwaliteitsanalyse plaats en de data wordt nogmaals onderzocht op bias. Er worden keuzes gemaakt met betrekking tot fouten in de data, outliers en missende waarden. Ook worden nieuwe variabelen gecreëerd. De gemaakte keuzes worden onderbouwd en beschreven.

2.2.2 Modelontwikkeling (stap 7)

In deze fase vindt het werkelijke trainen van (alternatieve) modellen plaats. Hier ligt de nadruk op de algemene kwaliteit van een model en een selectie van algoritmen. Tijdens de modelontwikkeling vindt nogmaals dataminimalisatie plaats als het mogelijk is om een goed model te ontwikkelen met minder variabelen. Met gebruik van de validatie dataset wordt er gekeken naar de mate waarin de getrainde modellen juiste en

[3] Ethiek en AI – Zeven methoden in theorie en praktijk; https://nlaic.com/wp-content/uploads/2022/11/Publicatie_NL_AIC_PACE_Zeven_methoden_nov_2022.pdf

onjuiste voorspellingen doen. Uiteindelijk wordt er definitief gekozen voor een model, waarbij simpel voor complex gaat. De keuze voor een bepaald algoritme wordt onderbouwd, onder andere door de 'metrics' te benoemen waarlangs wordt geëvalueerd.

2.2.3 Aannames toetsen (stap 8)

Aannames toetsen op businesswerkelijkheid gebeurt idealiter samen met de voorgaande twee stappen. Omgaan met datakwaliteitsissues, features die nuttig zijn bedenken en aannames over welke type fouten een model mag maken, zijn typisch vraagstukken die vanuit het domein worden beantwoord. De eindgebruikers en zo nodig andere belanghebbenden worden betrokken, al dan niet via een vertegenwoordiger, business analist of ethische adviseur. Tijdens deze stap wordt stilgestaan bij de relatie tot de organisatiedoelstellingen en de eindgebruikers/stakeholders.

2.2.4 Ethische overwegingen (stap 9)

Een model komt nooit helemaal overeen met de werkelijkheid. Er zitten altijd afwijkingen in. Daarom is het belangrijk om stil te staan bij bepaalde eerlijkheidsfactoren[4] als het gaat om algoritmegebruik. In deze stap wordt, met een ethisch adviseur, actief gekeken of groepen beschermd zijn tegen niet-terechte bias en profilering. Er wordt stilgestaan bij fairness en bias in zowel model als data. En er wordt een tweede keer getoetst op discriminatie of stigmatisering. Er worden ethische waarborgen ingesteld bij het ontwikkelen van het model.

2.2.5 Goedkeuren model (stap 10)

De verantwoordelijke uit het domein bepaalt uiteindelijk of en welke van de getrainde modellen goed genoeg is om gebruikt te worden. De keuze wordt gemaakt op basis van modeldocumentatie, waarin op een duidelijke manier is uitgelegd welke keuzes er zijn gemaakt bij het ontwikkelen van elk model. Op basis van de documentatie kunnen de modellen worden gecontroleerd op werking en aannames. Bij modellen met een grote impact kan een extra, onafhankelijke partij worden toegevoegd ter beoordeling, voordat het model wordt goedgekeurd.

2.3 Productie

Een succesvol model ontwikkelen is een mooie prestatie. Maar het is niet genoeg om een model ook daadwerkelijk te gebruiken. Ook voor veel grote organisaties, is het in productie brengen van modellen een grote uitdaging.[5]

[4] Eerlijkheid is in deze context geen eenduidig begrip. Wat iemand verstaat onder eerlijk hangt af van levensvisie. Zo kan je het bijvoorbeeld eerlijk vinden om gelijkheid te hebben in fouten van een model. Je kan het ook eerlijk vinden om meer fouten te hebben in een model, maar wel gelijkere uitkomsten. De werkgroep geeft geen oordeel over wat eerlijk is.
[5] https://www.forbes.com/sites/gilpress/2020/01/13/ai-stats-news-only-146-of-firms-have-deployed-ai-capabilities-in-production/?sh=697e612c2650

2.3.1 Data pipeline (stap 11)

Het ontwikkelen van een model gebeurt op een statische dataset. Om het model te gebruiken is het nodig om (structureel) data te voeden aan het model, de voorspelling uit te voeren en de werkelijke uitkomst weer terug te koppelen. Daarom worden er datastromen ontwikkeld. De organisatie heeft controle over de datastromen.

2.3.2 Model interface (stap 12)

Een voorspelling op zich doet nog niet veel voor de besluitvorming. Degene die een beslissing gaat nemen op basis van de voorspelling heeft informatie nodig over de voorspelling: op basis van welke factoren komt het tot stand en hoe sterk is de voorspelling? Een goede gebruikersinterface helpt om hier zicht op te krijgen. De interface dient als 'schil' om het model. De gebruiker heeft alleen toegang tot de interface, niet tot het model zelf. Het is bij de interface belangrijk dat de visualisatie(s) een correcte weergave is (zijn) van de output van het model. In de interface is er ook ruimte om de menselijke tussenkomst bij het model te registreren.

2.3.3 Issueregister ethiek (stap 13)

Ethische gesprekken en overwegingen voor het gebruik zijn nuttig. Organisaties kunnen gebruikers en belanghebbenden een issueregister bieden met daarin een overzicht van ethische aandachtspunten, en de wijze waarop deze zijn verwerkt. De aandachtspunten kunnen komen vanuit medewerkers en eindklanten. De data ethicus draagt zorg voor het register.

2.3.4 Implementatietraject (stap 14)

Een toekomstige gebruiker, of meerdere gebruikers, willen het model gaan gebruiken. Zij hebben wel training nodig over het doel van het model, de werking daarvan, de manier waarop het gebruikt wordt in het werkproces. Daarom vindt het implementatietraject plaats. Dit is het traject dat wordt doorlopen om gebruikers te trainen en te begeleiden bij het gebruiken en uitleggen van het model.

2.3.5 In gebruikname (stap 15)

De vraag is nu niet of het model goed is, maar of alles is georganiseerd voor het succesvol *gebruiken* van het model. Tijdens deze stap wordt de invulling van de menselijke tussenkomst gedocumenteerd en goedgekeurd voordat het product in gebruik kan worden genomen. Als het gebruik goed is georganiseerd, geeft de opdrachtgever goedkeuring voor het toepassen van het model in de praktijk.

2.4 Monitoring

Een van de kenmerken van modellen is dat de kwaliteit van een model verandert naarmate de tijd vordert. Als bijvoorbeeld de nieuwe data, waar voorspellingen op worden gedaan, veel verschilt van de trainingsdata, kan dat terug in de kwaliteit van de voorspellingen gezien

worden. Het is daarom van groot belang om de werking van de in gebruik genomen modellen goed te monitoren.

2.4.1 Bepalen kwaliteitseisen (stap 16)

Omdat de werking van modellen verandert, zal je afspraken moeten maken over de mate van veranderen dat wordt geaccepteerd. Er wordt rekening gehouden met drift bij het bepalen van kwaliteitseisen voor beheer. Bij welke verandering wordt niets gedaan en bij welke verandering wordt het model opnieuw getraind?

2.4.2 Ethische eisen (stap 17)

In deze stap worden er afspraken gemaakt over de toegestane veranderingen van de werking van het model, en wordt er getoetst op wijzigingen bij beschermde variabelen. Er worden kwaliteitseisen bepaald ten aanzien van beschermde variabelen, deze kunnen fungeren als evaluatienormen. Er wordt rekening gehouden met drift bij het bepalen van kwaliteitseisen voor beheer, waar het gaat om beschermde variabelen. Welke mate wordt geaccepteerd en wanneer vindt er hertraining plaats? Daarnaast kunnen er issues worden aangemaakt via het register. Hoe gaat de organisatie daarmee om?

2.4.3 Algoritmeregister (stap 18)

Publieke organisaties moeten openlijk communiceren naar burgers over het gebruik van algoritmes. Daarvoor is het openbare[6] algoritmeregister beschikbaar gemaakt. In het register wordt een overzicht van de belangrijkste datasets gegeven. Ook een beschrijving van de wijze waarop de data worden verwerkt en het algoritme ongelijke behandeling tegengaat. Bovendien wordt er beschreven in hoeverre mensen toezicht houden op de werking van het algoritmen, welke risico's er geïdentificeerd zijn en welke mitigerende maatregelen er genomen zijn. Het is aan te bevelen dat ook niet publieke organisaties een vergelijkbaa overzicht voor hun klanten/ belanghebbenden beschikbaar maken, bijvoorbeeld via eigen website.

2.4.4 Beheerafspraken uitvoeren (stap 19)

De beheerorganisatie neemt het model in beheer conform de afspraken die zijn gemaakt. Er wordt structureel gemonitord en op de afgesproken momenten wordt contact gezocht met bijvoorbeeld de data ethicus, analist of gebruiker (bijvoorbeeld voor periodieke evaluaties of bij veranderingen in de kwaliteit van voorspellingen). In deze fase wordt functiescheiding toegepast bij de toegang tot het algoritme en de toegangsrechten op passende wijze ingeregeld en up-to-date gehouden. Het aantal en de inhoud van de beheeraccounts wordt onderhouden, en er worden naamconventies gebruikt zodat gebruikers en beheerders kunnen worden geïdentificeerd. De functionaris gegevensbescherming heeft met een account toegang tot 'MLOPS'.

2.4.5 Initiëren herziening (stap 20)

[6] Het Algoritmeregister van de Nederlandse overheid (https://algoritmes.overheid.nl/)

Uiteindelijk zal blijken dat er aanpassingen nodig zijn aan het model. Als ze klein zijn vallen ze misschien onder het reguliere beheer. Voor grotere veranderingen is het domein weer aan zet, om die te initiëren. Hiermee ga je terug naar de voorbereiding fase, waarbij het doorlopen van de stappen waarschijnlijk veel sneller zal gaan dan bij het voor het eerst ontwikkelen van een model.

3. Rollen binnen het raamwerk

Bij de beschrijving van de fasen zijn verschillende rollen genoemd. Deze rollen worden hier beschreven.

3.1 Business rollen

3.1.1 Business owner

Manager/medewerker die opdracht geeft tot realisatie van een data-/informatieproduct. De opdrachtgever specificeert op welke wijze het data-/informatieproduct waarde levert en wordt hierbij ondersteund door de business analist en data steward. Tevens is hij/zij betrokken bij de implementatie van het data-/informatieproduct in de processen en het realiseren van de baten na oplevering van het product.

3.1.2 Business analist

De business analist is verantwoordelijk voor het beschikbaar zijn van heldere en concrete datawensen/-vragen. Hiertoe stemt de business analist af met managers/medewerkers uit de business om hun vragen verder te verhelderen en articuleren. Hij/zij werkt proactief aan het creëren van een werkvoorraad voor het data-/analyseteam door bij (potentiële) opdrachtgevers de meerwaarde van data en analyse te etaleren. De business analist ondersteunt tevens de opdrachtgever bij de implementatie van het data/-analyseproduct na oplevering. De business analist beschikt voor het uitvoeren van de rol over voldoende business kennis.

3.1.3 Data-eigenaar

De data-eigenaar is verantwoordelijk voor de definitie van data en de kwaliteit ervan binnen een bepaald organisatieonderdeel, vaak gerelateerd aan een bepaald proces waar hij/zij verantwoordelijk voor is. De data-eigenaar stuurt met dit proces als uitgangspunt op kwaliteit en beschikbaarheid van data (hier geldt: ‘proceskwaliteit is datakwaliteit’). Deze verantwoordelijkheid wordt wel eens gedelegeerd van een manager naar een uitvoerende medewerker.

3.1.4 Product owner/projectleider

De product owner/projectleider bepaalt in overleg met de opdrachtgever en andere stakeholders wat het multidisciplinaire team moet realiseren en in welke volgorde (prioriteit). Bij de rol hoort ook het specificeren van het op te leveren product en het bevorderen van en toezicht houden op de voortgang van de realisatie.

3.2 Analytics rollen

3.2.1 Data scientist

De data scientist voert analyses uit. Hij/zij zorgt ervoor dat de realisatie van de producten voldoet aan de afgesproken kwaliteitseisen. Tevens is hij/zij verantwoordelijk voor het beheer van data science producten. De data scientist heeft kennis van methoden en technieken voor het uitvoeren van analyses.

3.2.2 User experience ontwikkelaar

Deze draagt zorg voor een juiste en makkelijke ervaring bij het benutten van het eindproduct. Omdat het over dataproducten gaat, is datavisualisatie hier een onderdeel van. Het eindproduct kan bijvoorbeeld landen in een dashboard, app of website.

3.3 Datarollen

3.3.1 Data steward

De data steward is de primaire kennishouder over data binnen een bepaald inhoudsgebied. Hij/zij zorgt voor de juiste duiding, betekenis en context (documentatie, vastleggen metadata) van de data. Tevens draagt hij/zij bij aan de (her)bruikbaarheid en de kwaliteit van de data voor eindgebruikers. Hij/zij zorgt er op operationeel niveau voor dat men zich aan de data governance kaders en afspraken houdt. De data steward zorgt door heldere communicatie en coaching ervoor dat anderen beter in staat zijn om de waarde te vinden in het gebruik van data.

3.3.2 Data engineer

De data engineer werkt aan één of meerdere projecten waarbij hij/zij de ontsluiting van brondata in de analyseomgeving verzorgt, in overeenstemming met de afgesproken kwaliteitseisen. Hij/zij is verantwoordelijk voor het beheer van de analyseomgeving en voor het ontwikkelen en beheren van data pipelines. Hieronder vallen ook werkzaamheden als het opschonen, combineren en uniformeren van data.

3.4 Ethiek, privacy en security rollen

3.4.1 Data-ethicus

De data-ethicus is verantwoordelijk voor het uitvoeren van een ethische toets en het zorg dragen voor naleving van verantwoord datagebruik in alle fasen van het proces. De data-ethicus is de beheerder van het issueregister ethiek.

3.4.2 Privacy adviseur

De privacy adviseur is verantwoordelijk voor het uitvoeren van een privacy toets (DPIA) voor een nieuw data-/analyseproduct. Hij/zij ziet tevens toe op correcte toepassing van de privacy eisen bij de oplevering van een product en bij het aanpassen daarvan.

3.4.3 Security officer

De security officer is verantwoordelijk voor het uitvoeren van een security toets en het bepalen van het informatiebeveiligingsniveau voor een nieuw data-/analyseproduct. Hij/zij ziet tevens toe op correcte toepassing van de informatiebeveiligingseisen na oplevering van een product en bij het aanpassen ervan.

3.5 De inbedding van de rollen

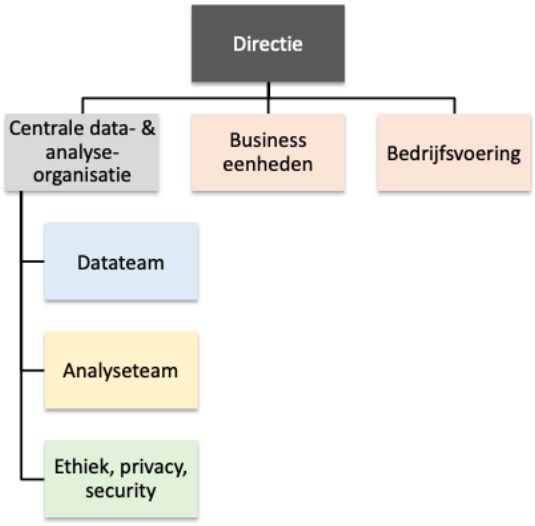
In DAMA DMBOK hoofdstuk 16 staan verschillende organisatievormen voor datamanagement toegelicht. Voor data science gelden in de basis dezelfde organisatievormen, waarbij extra complexiteit in de vorm van een analyse-organisatie wordt

toegevoegd. In deze paragraaf zijn de verschillende organisatievormen toegepast op het thema data science governance. Belangrijk uitgangspunt hierbij is dat er een “scheiding der machten” aanwezig is in de governance. Dat wil zeggen dat rollen die toezien op de juiste verwerking en analyse van data (de data-ethicus, privacy-adviseur en security officer) niet in dezelfde afdeling werken als diegenen die de data verwerken en algoritmes ontwikkelen (data steward, data engineer, data scientist).

3.5.1 Gecentraliseerd model

Binnen het gecentraliseerde model werken de analytics, data en ethiek, privacy en security rollen vanuit een centrale data- & analyseorganisatie, in verschillende afdelingen (zie figuur 3-1). Deze centrale organisatie ondersteunt de business eenheden en bedrijfsvoering met het realiseren en beheren van data (science) oplossingen.

Een belangrijk voordeel van het gecentraliseerde model is, is dat specialisten bij elkaar in één team werken. Hierdoor is het relatief makkelijk om de benodigde schaal te organiseren. Daarnaast is het eenvoudig om uniforme kwaliteitsstandaarden te implementeren en te werken aan de persoonlijke ontwikkeling van de specialisten.

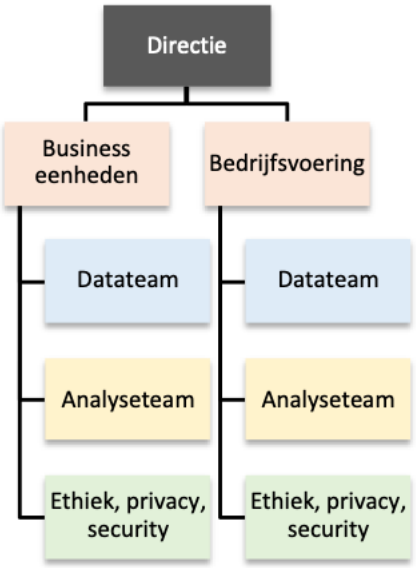


Figuur 3-1: Gecentraliseerd model

Nadeel van dit model is dat de afstand tussen de centrale data- en analysespecialisten en de business groot kan zijn. Hierdoor kan cruciale business kennis ontbreken bij de specialisten waardoor er een mismatch ontstaat tussen de problematiek in de business en de gerealiseerde data science oplossing. De grote afstand kan er tevens voor zorgen dat potentiële opdrachtgevers de centrale teams niet weten te vinden. Ook kan het lastig zijn om vanuit een centraal team (de juiste) keuzes te maken aangaande programmering en prioritering.

3.5.2 Gedecentraliseerd model

Het gedecentraliseerde model kenmerkt zich door business eenheden die hun eigen data- en analyseteams hebben (zie figuur 3-2). Er vindt geen centrale sturing plaats, waardoor iedere business eenheid de vrijheid heeft om de ondersteuning m.b.t. het realiseren van data science oplossingen zelf te organiseren. Daarnaast is de sturing op prioriteiten relatief eenvoudig en transparant.



Figuur 3-2: Gedecentraliseerd model

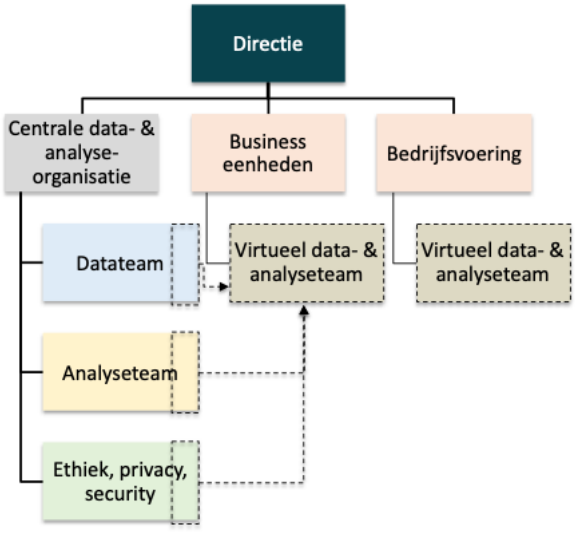
Het gedecentraliseerde model biedt de business eenheden de vrijheid de ondersteuning te organiseren die zij nodig hebben. Hierdoor kunnen zij snel capaciteit op- en afschalen. Daarnaast is de nabijheid van data- en analysespecialisten gegarandeerd en is het voor specialisten makkelijk(er) om relevante business kennis op te bouwen.

Een nadeel van het gedecentraliseerde model is een gebrek aan uniformiteit in de realisatie van data science oplossingen, wat kan leiden tot ongewenste kwaliteitsverschillen tussen business eenheden. Dit kan vervolgens het draagvlak onder de acceptatie van data science oplossingen ondermijnen. Daarnaast is het lastig om een organisatiebrede strategie te hanteren met dit model en te sturen op organisatiebrede prioriteiten.

Aanvullend nadeel van het gedecentraliseerde model is dat het voor sommige (kleinere) business eenheden lastig is om voldoende schaal te organiseren.

3.5.3 Hybride model

Zoals de naam al doet vermoeden, combineert het hybride model kenmerken van het gecentraliseerde en het gedecentraliseerde model. Een veelvoorkomende manier van werken is dat specialisten vanuit een centrale data- & analyseorganisatie zich voor langere tijd verbinden aan een specifieke business eenheid in virtuele data- & analyseteams (zie figuur 3.3). Hierdoor staan zij zowel in verbinding met elkaar als specialisten, maar werken zij ook dicht op de business.



Figuur 3-2: Hybride model

Het hybride model combineert de voordelen van het gecentraliseerde en gedecentraliseerde model. Zo kunnen (kwaliteits)standaarden makkelijk geïmplementeerd worden en vormt schaal in principe geen uitdaging. Daarnaast hebben business eenheden flexibiliteit in hoeveel dataspecialisten zij in willen zetten en is de nabijheid van de specialisten tot de business gegarandeerd. Doordat er wel één centrale data- & analyseorganisatie is, is het mogelijk om te sturen op organisatiebrede prioriteiten en strategie.

Een nadeel van het hybride model is de complexiteit van het concept. In veel organisaties is men niet gewend om te werken met meerdere aansturingslijnen. Bovendien kan het model moeilijk uitlegbaar zijn, waardoor men het lastig vindt het te hanteren.

4. Verantwoording raamwerk t.a.v. principes toepassing Data Science

Er zijn verschillende (overheids)kaders en richtlijnen die als hulpmiddel dienen voor organisaties.

De werkgroep heeft een vertaaltabel ontwikkeld waar van een aantal beleidsrichtlijnen de vertaling is gemaakt naar stappen in het raamwerk. In deze sectie wordt kort beschreven hoe het raamwerk rekening houdt met drie principes voor het toepassen van Data Science. De vertaaltabel is te vinden op: https://public.tableau.com/app/profile/arun.rampersad/viz/TableauworkbookDAMA_DSG/Dashboard1

4.1 Begrijpelijkheid

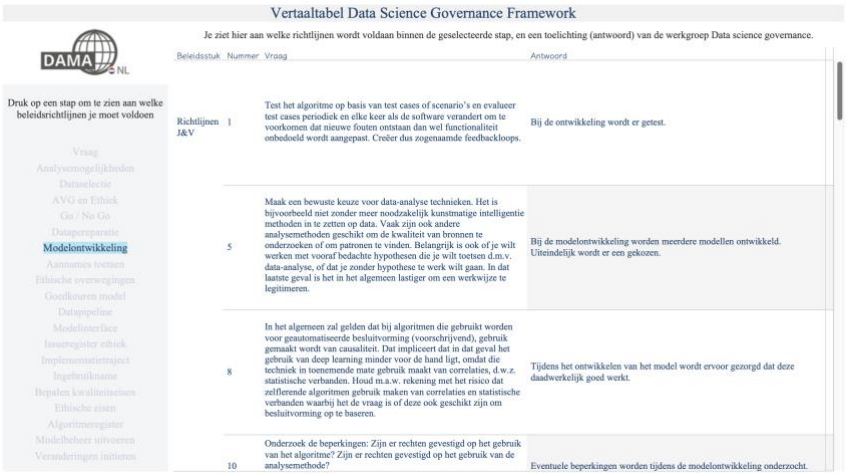
Door modeldocumentatie op te stellen en dat goed te laten keuren, is het noodzakelijk dat een model begrijpelijk is. Daarnaast is er sprake van een implementatietraject. Hier wordt elke gebruiker van het model getraind bij het gebruik. Het is daardoor niet alleen voor de ontwikkelaars maar ook voor de gebruikers begrijpelijk. Tot slot is er een algoritmeregister, waarin de werking van het model kan worden gecommuniceerd naar de gebruikers en belanghebbenden.

4.2 Eerlijkheid

Door multidisciplinair te werken kan ervoor gezorgd worden dat het niet mogelijk is dat het perspectief van 1 partij doorslaggevend is. De mening van een data scientist is niet bepalend bij het maken van een eerlijk model. Door in elke fase stil te staan bij ethiek, is het werken aan eerlijkheid een continu proces. Tot slot is er de mogelijkheid voor elke stakeholder om ethische zorgen te uiten, welke transparant kunnen worden verwerkt in het issueregister.

4.3 Kwaliteit

Tijdens de modelontwikkeling wordt de kwaliteit geborgd door domeinkennis, ethische kennis en analysekennis te bundelen. Bij modellen is het zo dat de kwaliteit daarvan verandert in de loop van tijd. Door bij de monitoring vereisten op te stellen voor veranderingen in het model en regulier onderhoud in te plannen bij het beheer, wordt de kwaliteit gewaarborgd.


Vertaaltabel Data Science Governance Framework

Je ziet hier aan welke richtlijnen wordt voldaan binnen de geselecteerde stap, en een toelichting (antwoord) van de werkgroep Data science governance.

Druk op een stap om te zien aan welke beleidsrichtlijnen je moet voldoen

Vraag
Analysemogelijkheden
Dataselectie
AVG en Ethiek
Go / No Go
Dataprocessstap
Modelontwikkeling
Aannames toetsen
Ethische overwegingen
Goedkeuren model
Datapipeline
Modelinterface
Issueregister ethiek
Implementatietraject
Ingebruikname
Bepalen kwaliteitseisen
Ethische risen
Algoritmeregister
Modelbeheer uitvoeren
Veranderingen initieren

Beleidsstuk	Nummer	Vraag	Antwoord
Richtlijnen J&V	1	Test het algoritme op basis van test cases of scenario's en evalueer test cases periodiek en elke keer als de software verandert om te voorkomen dat nieuwe fouten ontstaan dan wel functionaliteit onbedoeld wordt aangepast. Creëer dus zogenaamde feedbackloops.	Bij de ontwikkeling wordt er getest.
	5	Maak een bewuste keuze voor data-analyse technieken. Het is bijvoorbeeld niet zonder meer noodzakelijk kunstmatige intelligentie methoden in te zetten op data. Vaak zijn ook andere analysemethoden geschikt om de kwaliteit van bronnen te onderzoeken of om patronen te vinden. Belangrijk is ook of je wilt werken met vooraf bedachte hypothesen die je wilt toetsen d.m.v. data-analyse, of dat je zonder hypothese te werk wilt gaan. In dat laatste geval is het in het algemeen lastiger om een werkwijze te legitimeren.	Bij de modelontwikkeling worden meerdere modellen ontwikkeld. Uiteindelijk wordt er een gekozen.
	8	In het algemeen zal gelden dat bij algoritmen die gebruikt worden voor geautomatiseerde besluitvorming (voorschrijvend), gebruik gemaakt wordt van causaliteit. Dat impliceert dat in dat geval het gebruik van deep learning minder voor de hand ligt, omdat die techniek in toenemende mate gebruik maakt van correlaties, d.w.z. statistische verbanden. Houd m.a.w. rekening met het risico dat zelflerende algoritmen gebruik maken van correlaties en statistische verbanden waarbij het de vraag is of deze ook geschikt zijn om besluitvorming op te baseren.	Tijdens het ontwikkelen van het model wordt ervoor gezorgd dat deze daadwerkelijk goed werkt.
	10	Onderzoek de beperkingen: Zijn er rechten gevestigd op het gebruik van het algoritme? Zijn er rechten gevestigd op het gebruik van de analysemethode?	Eventuele beperkingen worden tijdens de modelontwikkeling onderzocht.

Figuur 4-1: Vertaaltabel Data Science Governance Framework

5. Aandachtspunten voor het gebruik van het raamwerk

5.1 Inbedden in het proces

Het raamwerk met de 20 stappen kan in het data science proces van de organisatie ingebed worden. Het is mogelijk om de 20 stappen op te nemen in een projectmanagementmodule, zodat een organisatie weet dat alle stappen met de daarbij verwachte output wordt geleverd. Het brengt overzicht en inzicht op het vakgebied.

5.2 Een model is niet de werkelijkheid

Een voorspelmodel benadert de werkelijkheid, maar het is niet de werkelijkheid. Zo is het ook met dit raamwerk. Gebruik het op een manier waarop het waarde toevoegt voor de eigen organisatie. Vragen en suggesties voor verbetering zijn overigens welkom en kunnen naar: arun@data-nl.org gemaild worden. Het raamwerk is een ”work-in-progress" en zal op basis van ervaringen van de leden van de werkgroep en de feedback en suggesties van de gebruikers regelmatig worden verbeterd.

5.3 Verspreiden van het raamwerk

[7] https://creativecommons.org/licenses/by/4.0/deed.nl

6. Bronnen en aanbevolen literatuur

Algemene Rekenkamer. *Aandacht voor algoritmes*. 2021
https://www.rekenkamer.nl/publicaties/rapporten/2021/01/26/aandacht-voor-algoritmes

Algemene Rekenkamer. *Algoritmes getoetst*. 2022
https://www.rekenkamer.nl/publicaties/rapporten/2022/05/18/algoritmes-getoetst

DAMA International. *DAMA-DMBOK*. 2017

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. *Impact Assessment Mensenrechten en Algoritmes*. 2021
https://open.overheid.nl/documenten/ronl-c3d7fe94-9c62-493f-b858-f56b5e246a94/pdf

Ministerie van Justitie en Veiligheid. *Richtlijnen voor het toepassen van algoritmen door overheden en publieksvoorlichting over data-analyses*. 2021
https://open.overheid.nl/documenten/ronl-1411e45f-b822-49fa-9895-2d76e663787b/pdf

Nederlandse AI Coalitie. *Ethiek en AI – Zeven methoden in theorie en praktijk*. 2022
https://nlaic.com/wp-content/uploads/2022/11/Publicatie_NL_AIC_PACE_Zeven_methoden_nov_2022.pdf

Provost, F & Fawcett, T. *Data Science for Business*. 2013

Rampersad, A. *Zicht op Algoritmen*. 2021.
https://files.cdn.thinkific.com/file_uploads/465461/attachments/5b2/d43/26b/Zicht_op_Algoritmen.pdf

Tweede Kamer. *kamerbrief*, vergaderjaar 2020-2021, 2643, nr. 765, p.1