

Machine Learning Pipeline Automation

Amar Kisoensingh

25 juni 2021

Voorwoord

Voor u ligt de scriptie "Machine Learning Pipeline Automation". Deze scriptie is geschreven in het kader van mijn afstuderen aan de opleiding Informatica aan de Hogeschool Rotterdam en in opdracht van het stagebedrijf NGTI. De afstudeerstage liep van februari 2021 tot juni 2021. Deze scriptie is gericht tot mensen binnen het Informatica domein geïnteresseerd in machine learning pipelines en het automatiseren en versimpelen.

De onderzoeksvraag is bedacht door mijn bedrijfsbegeleider Okke van 't Verlaat. Halverwege de afstudeerstage heeft Kolja van der Vaart de bedrijfsbegeleiding overgenomen. Terugkijkend was de vraag complex en ambitieus. Ik heb me beziggehouden met machine learning, automatisering en cloud platformen.

Bij deze wil ik mijn afstudeerbegeleiders Okke van 't Verlaat, Kolja van der Vaart en vanuit school Stelian Paraschiv en Marian Slingerland bedanken voor de zorgvuldige begeleiding. Ik heb met Okke en Kolja effectief kunnen sparren over het onderzoek. Mijn begeleiders hebben mij ook geholpen om de scriptie in de juiste richting te sturen. Tot slot wil ik mijn familie en vrienden bedanken voor de kritische en motiverende feedback.

Ik wens u veel leespleszier toe.

Amar Kisoensingh

Zoetermeer, 25 juni 2021

Samenvatting

Een probleem met machine learning (ML) is dat het gebruik ervan tijd en kennis vereist. Ook is er kans op vendor lock-in als ML wordt gebruikt op een cloud platform zoals Google Cloud of Microsoft Azure. Het is dus vrijwel lastig om te verhuizen naar een andere cloud platform.

Het doel van deze scriptie is om niet alleen te onderzoeken in hoeverre het mogelijk is om ML te automatiseren, maar ook of dit kan ongeacht het cloud platform. De onderzoeksvraag kan als volgt opgesteld worden: kan ML geautomatiseerd worden op een cloud platform-agnostische wijze?

Om deze vraag te beantwoorden is er onderzoek gedaan naar ML pipelines. Hiervoor is literatuuronderzoek en experimenteren van pas gekomen. Met ML pipelines kan op een gestructureerd manier gewerkt worden met ML. Dit verlaagt de leercurve en maakt ML toegankelijker voor nieuwe developers. Naast het onderzoek is ook een proof of concept (PoC) gebouwd om de platform-agnostische kant te valideren. Met het onderzoek en experimentatie dat is gedaan blijkt dat het mogelijk is om ML te automatiseren op een cloud platform-agnostische manier.

Het onderzoek kan verder uitgebreid worden door te kijken naar hoe ML toegankelijker gemaakt kan worden voor developers en hoe het systeem robuuster nog opgezet kan worden.

Summary

One problem with machine learning (ML) is that using it requires time and knowledge. There is also a chance of vendor lock-in if ML is being used on a cloud platform such as Google Cloud or Microsoft Azure. It is very hard to move to another cloud platform.

The aim of this thesis is not only to investigate to what extent ML can be automated, but also whether this is possible regardless of the cloud platform. The research question can be formulated as follows: can ML be automated in a cloud platform agnostic way.

To answer this question, research has been done on ML pipelines. Literature research and experimentation were utilized. With ML pipelines you can work with ML in a structured way. This not only makes ML more accessible but lowers the learning curve for new developers. In addition to the research, a proof of concept (PoC) has also been built to validate the platform-agnostic aspect. The research and experimentation that has been done shows that it is possible to automate ML in a cloud platform agnostic way.

The research can be further expanded by looking at how ML can be made more accessible for developers and how the system can be set up more robustly.

Afkortingen

API application programming interface 5, 27, 28, 35, 41, 44

CLI command line interface 41

CRUD create, read, update en delete 59

DRY don't repeat yourself 6, 59, 60

ERD entity relation diagram 5, 42

IaC infrastructure as code 5, 33, 34

MLPA machine learning pipeline automation 41

MVP minimal viable product 49–51

PaaS Platform as a Service 15, 16

REST representational state transfer 35

SPA single page application 5, 41, 44

SVC support vector classification 5, 26

UI user interface 51, 53

VM virtual machine 6, 46, 57, 58, 61, 90–92

Begrippenlijst

artifact Een artifact is de output van een pipeline proces **artifact-definition**. 6, 51, 58, 59, 61

scope creep Wanneer de omvang van een project ongecontroleerd toeneemt **scope-creep-definition**. 50

Lijst van figuren

1.1	Organogram van NGTI op 29-03-2021 [3]	11
1.2	Screenshot van de Swiss Climate Challenge app [5]	12
1.3	Screenshot van de My Swisscom App [7]	12
4.1	Machine learning in de context van andere domeinen	22
4.2	Lifecycle van een model volgens Hapke en Nelson [17, p. 4]	23
4.3	Voorbeeld van integer encoding	25
4.4	Voorbeeld van hot encoding	25
4.5	Gamma hyperparameter van een support vector classification (SVC) algoritme met waarde 0.1	26
4.6	Gamma hyperparameter van een support vector classification (SVC) algoritme met waarde 100	26
4.7	Model uitgerold met behulp van een API	28
4.8	Vorm van expliciete feedback gebruikt door YouTube	29
4.9	Machine learning pipeline met een algoritme exploratie tussenstap	30
4.10	Machine learning pipeline waarbij stappen geautomatiseerd kunnen worden voor het gemak van developers	31
5.1	Voorbeeld van een infrastructure as code (IaC) plan [27]	33
5.2	Proces van een infrastructure as code (IaC) om aan de gewenste situatie te voldoen	34
5.3	Sequence-diagram van het experiment met orkestratietool Pulumi	36
5.4	Node.js server voor het experiment met orkestratietool Pulumi	37
5.5	De stack voor het experiment	37
5.6	POST endpoint van de Pulumi experiment	38
6.1	Voorbeeld van een C4 model [41]	40
6.2	Context niveau diagram van het architectuurontwerp	41
6.3	Containerniveau-diagram van het architectuurontwerp	42
6.4	Entity relation diagram (ERD)	42
6.5	Single page application (SPA) componentniveau diagram van het architectuurontwerp	44
6.6	Node.js server componentniveau diagram van het architectuurontwerp	45
6.7	Sequence-diagram van het aanmaken van een pipeline	46
6.8	Sequence-diagram van het starten van een pipeline	47
7.1	Requirements opgesteld voor de proof of concept	49
7.2	Trello bord in het begin van sprint 1	50
7.3	Mock up van de pipeline details pagina	51
7.4	Gedrag van Pulumi met een incorrect plan	52
7.5	Aanmaak dialoog van een pipeline	53
7.6	Service om een pipeline aan te maken	54
7.7	Functie om een storage bucket aan te maken in Google Cloud	54
7.8	Status en upload-mogelijkheid van datasets en artefacts in de details-pagina van een pipeline	55
7.9	Service om datasets te uploaden naar een opslaglocatie binnen een cloud platform	55

7.10	Functie om datasets te uploaden naar een storage bucket in Google Cloud	55
7.11	Configuratiedialoog van een pipeline	56
7.12	Code om de configuratie op te slaan in de database	56
7.13	Status en startknop van een pipeline	57
7.14	Code om een virtual machine (VM) in Google Cloud te starten	57
7.15	Startup script van een virtual machine (VM) in Google Cloud	58
7.16	Model en overige artifacts van het trainproces geüpload in een storage bucket . .	59
7.17	Voorbeeld van de repository-service pattern	59
7.18	Voorbeeld van de don't repeat yourself (DRY) principe in de backend	60
7.19	Voorbeeld van het don't repeat yourself (DRY) principe in de frontend	60
7.20	Unitests van de <i>Button</i> component in de frontend	61
7.21	Unitests van de pipeline in de backend	62
7.22	Mock up van een verbeterde configuratiedialoog - Stap 1	64
7.23	Mock up van een verbeterde configuratiedialoog - Stap 2	65
7.24	Mock up van een verbeterde configuratiedialoog - Stap 3	65
7.25	Mock up van een verbeterde configuratiedialoog - Stap 4	66
7.26	Mock up van een verbeterde configuratiedialoog - Stap 5	67
1	Mindmap van machine learning algoritmes	81
2	Importeren van modules	82
3	Importeren van de dataset	82
4	Validatie van de dataset	83
5	Validatie van de dataset - resultaat 1	84
6	Validatie van de dataset - resultaat 2	84
7	Dataset voorbereiden	84
8	Dataset voorbereiden	85
9	Verschil van de kernel hyperparameter	86
10	Verschil van de gamma hyperparameter	86
11	Verschil van de c hyperparameter	87
12	Verschil van de degree hyperparameter	87
13	Dataset voorbereiden	88
14	Dataset voorbereiden	88
15	Stack configuratie van het experiment met Pulumi	89
16	Mock up van het pipeline overzicht	90
17	Mock up van het aanmaken van een pipeline	90
18	Mock up van het configuratie dialoog	91
19	Validation sequence-diagram	92
20	Code om de status van een VM op te halen in Google Cloud	93
21	Virtual machine (VM) output in de front-end	94
22	Code om een virtual machine (VM) te verwijderen in Google Cloud	95
23	Globale planning	96
24	Notities bij elke meeting	97
25	Pagina 1 van de getuigenverklaring bedrijfsbegeleider	98
26	Pagina 2 van de getuigenverklaring bedrijfsbegeleider	99

Lijst van tabellen

3.1	Onderzoeks methode deelvraag 1	19
3.2	Onderzoeks methode deelvraag 2	19
3.3	Onderzoeks methode deelvraag 3	19
3.4	Onderzoeks methode hoofdvraag	20
5.1	Knock-out criteria voor orkestratietools dat cloud computing platformen beheert	34
5.2	Knock-out criteria tegen orkestratietools dat cloud computing platformen beheerd	35
6.1	Criteria voor frameworks	43
6.2	Criteria voor frontend frameworks	43
6.3	Criteria voor backend frameworks	43
1	Scope deelvraag 1	78
2	Scope deelvraag 2	79
3	Scope deelvraag 3	79
4	Scope hoofdvraag	80
5	Voorbeeld van de iris dataset	82

Inhoudsopgave

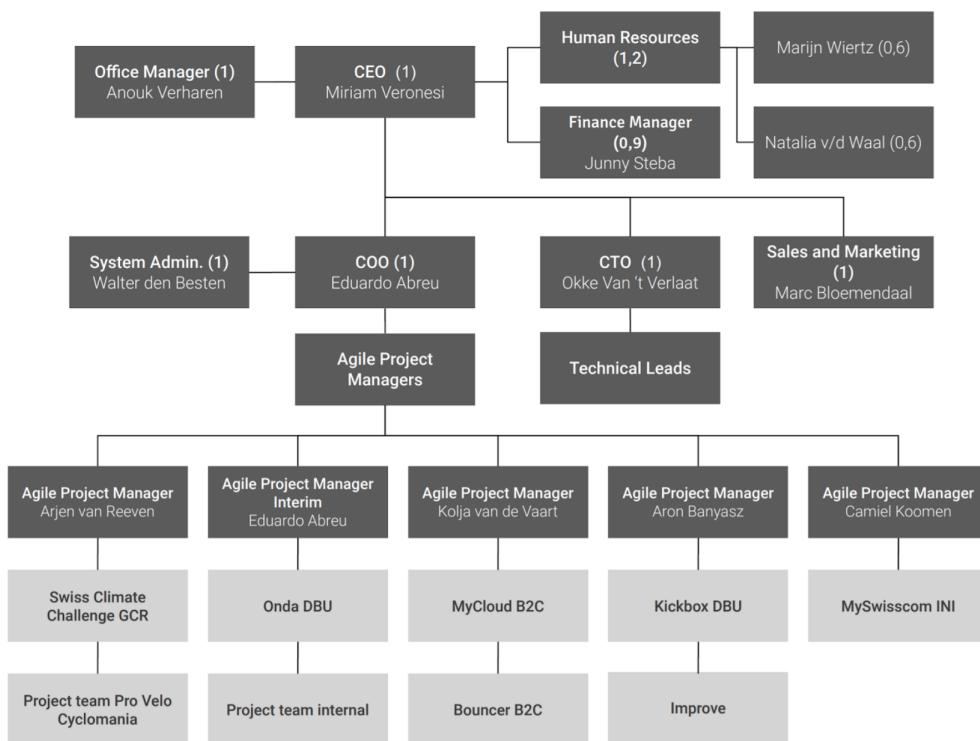
1 Inleiding	10
1.1 Projecten van NGTI	11
1.2 Tools die worden gebruikt	12
1.3 Aanleiding opdracht	13
2 Probleemanalyse	14
2.1 Obstakels voor NGTI	15
2.2 Vooronderzoek naar bestaande oplossingen	16
2.3 Doelstelling	16
2.4 Hoofd- en deelvragen	16
3 Onderzoeksmethoden en scope	18
4 Stappen in een machine learning pipeline	21
4.1 Wat is machine learning?	22
4.2 De stappen in een machine learning pipeline	23
4.3 Conclusie	29
4.4 Advies	30
5 Orkestratietools om platformen te beheren	32
5.1 Wat is infrastructure as code?	33
5.2 Orkestratietools vergeleken met elkaar	34
5.3 Experiment met de orkestratietool Pulumi	35
5.4 Conclusie	38
6 Architectuurontwerp van de oplossing	39
6.1 Het C4 model	40
6.2 Niveau 1: Context	41
6.3 Niveau 2: Container	41
6.4 Niveau 3: Component	44
6.5 Sequence-diagrammen	45
7 Proof of concept	48
7.1 User requirements verzamelen	49
7.2 Mock up van de proof of concept	51
7.3 Wijzigingen in het architectuurontwerp	51
7.4 De proof of concept	53
7.5 Kwaliteit van de code	59
7.6 Conclusie	62
7.7 Advies	63
8 Discussie	68
9 Reflectie	70

Bijlagen	77
I Scope deelvraag 1	78
II Scope deelvraag 2	79
III Scope deelvraag 3	79
IV Scope hoofdvraag	80
V Mindmap van machine learning algoritmes	81
VI Machine learning pipeline experiment	82
VII Stack configuratie van het experiment met Pulumi	89
VIII Mock up - Overzicht pipeline en pipeline aanmaken	90
IX Mock up - Configuration dialoog	91
X Validation sequence diagram	92
XI Proof of concept code - <i>gc_getRunStatus()</i>	93
XII Proof of concept code - VM output in de front-end	94
XIII Proof of concept code - <i>gc_stopRun</i>	95
XIV Globale planning	96
XV Notulen bij meetings	97
XVI Getuigenverklaring bedrijfsbegeleider	98

1 Inleiding

NGTI is een software ontwikkelbedrijf dat gevestigd is in Rotterdam. Opgericht in 2012 maakt NGTI applicaties (apps) voor mobiel en/of webgebruik. Naast het ontwikkelen werkt NGTI aan het hele traject om een app heen, namelijk de probleemstelling, mockups, wireframes en prototyping. Daarnaast levert NGTI ook support en lost bugs op nadat een app live is gegaan [1]. Ook maakt NGTI white label apps en frameworks [2].

Het bedrijf heeft, zoals de meeste bedrijven, een organogram (Figuur 1.1). In de praktijk is dit echter niet terug te vinden en wordt de structuur gezien als "plat". Collega's kunnen elkaar laagdrempelig benaderen waardoor niemand een onbekende is en verschillende disciplines makkelijk met elkaar samen kunnen werken.



Figuur 1.1: Organogram van NGTI op 29-03-2021 [3]

NGTI is een dochterbedrijf van Swisscom [4]. Sinds maart 2021 is het bekend gemaakt dat Swisscom van plan is om een afdeling, Swisscom DevOps Center, te fuseren met NGTI per 1 juli. Omdat de fusie onzekerheid met zich meebrengt voor de structuur en manier hoe NGTI werkt, zal de situatie vóór de fusie aangehouden worden gedurende het afstuderen.

1.1 Projecten van NGTI

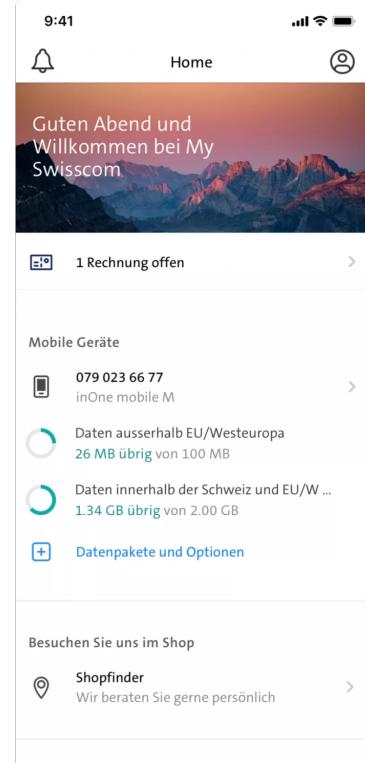
NGTI heeft een vrij breed portfolio met apps voor verschillende doeleinden. Een van deze apps is de Climate Challenge App [5]. Met deze app kunnen gebruikers hun CO₂-voetafdruk en impact in kaart brengen. Er wordt bijgehouden hoeveel kilometer de gebruiker reist en met welk vervoersmiddel. De app is onderdeel van twee bestaande nieuwsapps, Blick en Bluewin [6]. Het

doel is om de gebruiker aan te sporen om groener te reizen. Een screenshot van de app is te zien in Figuur 1.2.

Een andere oplossing is de My Swisscom App [7]. Dit is een native app voor Android en iOS waarbij Swisscom-klanten hun contract kunnen bestellen, wijzigen of beëindigen. In de app kunnen klanten ook de dataverbruik zien en instellingen voor abonnementen wijzigen. Een screenshot van de app is te zien in Figuur 1.3.



Figuur 1.2: Screenshot van de Swiss Climate Challenge app [5]



Figuur 1.3: Screenshot van de My Swisscom App [7]

1.2 Tools die worden gebruikt

Om productief te zijn, gebruikt NGTI een aantal tools en programma's om producten te maken en te communiceren met zowel collega's als klanten. De meest gebruikte en belangrijkste zijn Slack, Google Workspace, Zoom en Microsoft Teams.

1.2.1 Slack

Interne communicatie gaat via Slack. Het programma faciliteert collega's om elkaar met een lage instap te benaderen en berichten die voor het hele bedrijf relevant zijn te versturen. Ook zijn er 'channels' beschikbaar over specifieke onderwerpen, zoals: `#dev`, `#ios` en `#test-automation`.

1.2.2 Google Workspace

Met Google Workspace kunnen bestanden en documenten gemaakt, opgeslagen en gedeeld worden. Dit is mogelijk via een browser waardoor werknemers geen software hoeven te installeren. NGTI gebruikt het ook om collaboratief en parallel te werken aan hetzelfde document.

1.2.3 Microsoft Teams en Zoom

Voorheen werd Zoom alleen gebruikt om te videobellen met collega's en geïnterviewden. In de tijd van de pandemie is Zoom echter een belangrijke speler geworden om effectief samen te werken. Meetings zoals introducties van nieuwe collega's of demo's van producten worden online gehouden. Microsoft Teams wordt gebruikt om met collega's van Swisscom Zwitserland te communiceren omdat Zoom niet toegestaan is voor hun medewerkers.

1.3 Aanleiding opdracht

NGTI wil de gebruikerservaring van haar apps verbeteren en een voorsprong hebben op haar concurrenten. Dit kan NGTI op een aantal manieren doen waarvan apps "slimmer" maken er één van is. Het slimmer maken houdt voor NGTI in dat de app bijvoorbeeld beter kan anticiperen wat de gebruiker wil en nodig heeft op een gegeven moment. Dit kan onder andere door het toepassen van machine learning (ML). De opdracht kan verdeeld worden in drie onderdelen:

- Onderzoek naar het versimpelen van ML voor developers
- Onderzoek naar hoe een pipeline opgezet kan worden op een cloud computing platform door middel van een framework
- Onderzoek naar het maken van een platform-agnostische oplossing

De aanleiding zal in het volgende hoofdstuk verduidelijkt worden. Daarnaast worden de onderdelen toegelicht en wordt een beeld gegeven van het probleem.

2 Probleemanalyse

NGTI wil ML toepassen om haar apps slimmer te maken. Hier zijn een aantal redenen voor, onder andere om de gebruikerservaring te verbeteren en om een voorsprong te hebben op concurrenten. Het 'slimmer' maken van applicaties kan op verschillende manieren, maar met machine learning kan een platform gebouwd worden waarmee elke richting op gegaan kan worden. Een voorbeeld van toegepaste ML is het aanbevelingssysteem van YouTube [8], gepersonaliseerde kunst van Netflix [9] of de Smart Reply feature van Google in Gmail [10].

Om machine learning te implementeren in haar applicaties loopt NGTI tegen een aantal obstakels aan, namelijk: expertise vereist in het ML domein, benodigde tijd om een pipeline op te zetten en vendor lock-in.

2.1 Obstakels voor NGTI

NGTI loopt tegen de voorgenoemde obstakels aan omdat NGTI weinig ervaring heeft met ML. In samenwerking met een extern bedrijf worden modellen getraind die vervolgens gebruikt worden in apps van NGTI. Om zelf modellen te trainen heeft NGTI kennis over ML, ML pipelines en tijd nodig. Bovendien wilt NGTI niet bij één cloud computing platform haar infrastructuur opzetten maar gemakkelijk kunnen schakelen tussen platformen. De obstakels worden in de volgende paragrafen verduidelijkt.

2.1.1 Expertise machine learning

ML is een ingewikkeld en diepgaand onderwerp. Om een model te trainen is kennis nodig van verschillende domeinen: data mining, software engineering en statistieken. Doordat er voorkennis nodig is om een model te trainen en een pipeline goed op te zetten, is het vaak te hoogdrempelig voor developers om een start te maken met ML. De expertise is daarnaast niet in een korte tijd te vergaren.

2.1.2 Opzetten pipeline

Boven op de complexiteit van ML zelf bestaan er verschillende manieren om een model te trainen. Het opzetten van ML pipelines is daar een van. Een pipeline is een workflow dat bestaat uit een aantal stappen die doorgelopen worden om een model te trainen. In elke stap worden acties uitgevoerd, zoals het verwijderen van onbruikbare data of de prestatie van modellen vergelijken en een rapport met uitslagen genereren. Het opzetten van zo een pipeline én de actie(s) in de stappen definiëren kost tijd en vereist specifieke kennis. Daarnaast zijn de stappen en acties vaak hetzelfde voor verschillende pipelines. Het automatiseren en hergebruiken van stappen en acties tussen pipelines zou tot onder andere tijdwinst, gebruikersgemak en het verminderen van herhaling van code kunnen leiden.

2.1.3 Vendor lock-in

Er bestaan een aantal diensten, zogenoemde Platform as a Service (PaaS), waarbij je een pipeline kan opzetten en acties kan definiëren. Een van de problemen met een PaaS is vendor lock-in. Dit betekent dat, als er eenmaal een pipeline is opgezet, de overdraagbaarheid van de pipeline naar een andere PaaS lastig is. Pipelines van diensten komen niet altijd overeen met elkaar waardoor een vertaalstap genomen moet worden om de pipeline te verhuizen naar een andere dienst. Het kan voorkomen dat overstappen naar een andere dienst niet mogelijk is omdat de andere dienst

niet alle features van de originele dienst ondersteund. Ook zijn de opties en mogelijkheden om uit te breiden in de toekomst afhankelijk van wat de dienst implementeert. Om onafhankelijk te zijn is een oplossing dat platform-agnostisch is een eis.

2.2 Vooronderzoek naar bestaande oplossingen

Het gebruik van ML pipelines is geen nieuwe techniek en is mogelijk bij PaaS's en bedrijven die zich specialiseren in ML pipelines. Het gemak met zulke services is dat de gebruiker gelijk kan beginnen met het trainen van ML modellen en zich niet zorgen hoeft te maken over infrastructurele details. Echter is er sprake van vendor lock-in en kunnen services van deze bedrijven niet gebruikt worden. Vendor lock-in is een van de obstakels dat NGTI probeert te vermijden; platform-agnostisch zijn is dus een van de eisen.

Gedurende het vooronderzoek zijn "Infrastructure as Code" frameworks naar voren gekomen. Dit zijn frameworks waarmee, middels code, een infrastructure binnen een PaaS opgezet kan worden. Dit kan gebruikt worden als onderdeel van de PoC en wordt in een van de deelvragen verder onderzocht.

2.3 Doelstelling

Om haar doel te bereiken, de gebruikerservaring van apps verbeteren en een voorsprong te hebben op concurrenten, is NGTI van plan ML pipelines te gebruiken om ML toe te passen. De gewenste oplossing is een systeem waarbij developers met weinig tot geen kennis een model kunnen trainen. Het systeem moet de infrastructurele taken voor zich nemen, zoals het opzetten van een pipeline en de stappen en acties automatiseren. Daarnaast moet een ML pipeline pijnloos doorlopen kunnen worden op verschillende platformen zodat het systeem platform-agnostisch is.

2.4 Hoofd- en deelvragen

Uitgaande van de drie obstakels kan de hoofdvraag als volgt worden geformuleerd:

*In welke mate kan een machine learning pipeline worden
geautomatiseerd onafhankelijk van het onderliggende cloud computing
platform?*

De hoofdvraag kan worden onderbouwd met vier deelvragen. Om te beginnen is het verstandig om te weten welke stappen er in een machine learning pipeline zit:

Waar bestaat een machine learning pipeline uit?

Daarnaast is een framework nodig dat, door middel van code, verschillende cloud computing platformen kan beheren:

*Hoe kan een framework verschillende cloud computing platformen beheren om een
machine learning pipeline op te zetten?*

Ten slotte wordt een PoC gemaakt om te laten zien of het probleem oplosbaar is. Hiervoor is een doordachte voorbereiding onmisbaar:

Hoe ziet de architectuurblauwdruk van een applicatie, waarmee een platform-onafhankelijk machine learning pipeline opgezet kan worden, eruit?

3 Onderzoeksmethoden en scope

Om elke hoofd- en deelvraag te beantwoorden, wordt er bij elke vraag gebruik gemaakt van een onderzoeks methode. Volgens Scribbr [11] zijn er twee onderzoeks methoden: kwantitatief en kwalitatief. Bij een kwantitatief onderzoeks methode wordt data verzameld waarmee bijvoorbeeld grafieken of tabellen gemaakt kunnen worden. De focus bij een kwalitatief onderzoeks methode ligt bij het verzamelen van verschillende interpretaties en opvattingen. Hierop kan optioneel een eigen interpretaties op gemaakt worden. [12].

Onder kwantitatief en kwalitatief vallen verschillende dataverzamelings methoden. Deze beschrijft simpelweg de manier hoe data wordt verzameld. Dit kan bijvoorbeeld met een enquête, literatuuronderzoek op websites en in boeken of een onderzoek over een lange periode [12].

Elke hoofd- en deelvraag is gekoppeld aan een onderzoeks methoden. Vervolgens is beschreven welk(e) dataverzamelings methode(n) wordt gebruikt met een korte toelichting. Daarnaast wordt op een hoog niveau de scope bepaald.

Tabel 3.1: Onderzoeks methode deelvraag 1

D1: Waar bestaat een machine learning pipeline uit?	
Methode(s)	Kwalitatief
Dataverzamelings methode(n)	Literatuuronderzoek, fundamenteel onderzoek, toegepast onderzoek
Scope	Bijlage I

Tabel 3.2: Onderzoeks methode deelvraag 2

D2: Hoe kan een orkestratietool verschillende cloud computing platformen beheren om een machine learning pipeline op te zetten?	
Methode	Kwalitatief
Dataverzamelings methode(n)	Literatuuronderzoek, vergelijkend onderzoek
Scope	Bijlage II

Tabel 3.3: Onderzoeks methode deelvraag 3

D3: Hoe ziet de architectuurblauwdruk van een applicatie, waarmee een platform-onafhankelijk machine learning pipeline opgezet kan worden, eruit?	
Methode	Kwalitatief
Dataverzamelings methode(n)	Literatuuronderzoek
Scope	Bijlage III

Tabel 3.4: Onderzoeks methode hoofdvraag

H: In welke mate kan een machine learning pipeline worden geautomatiseerd onafhankelijk van de onderliggende cloud computing platform?	
Methode	Kwalitatief
Dataverzamelingsmethode(n)	Literatuuronderzoek
Scope	Bijlage IV

4 Stappen in een machine learning pipeline

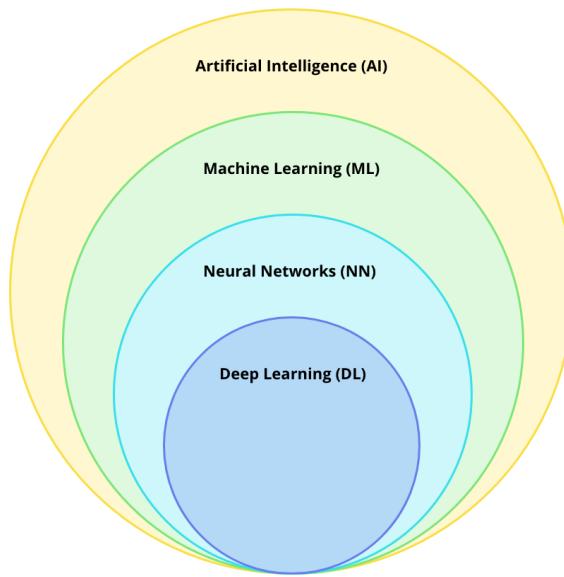
Een ML pipeline is zoals beknopt beschreven in deelparagraaf 2.1.2 een collectie van stappen dat wordt doorlopen om een model te trainen. Elke stap bevat een aantal acties dat wordt uitgevoerd, zoals onbruikbare data weghalen of de prestatie analyseren. De stappen en acties worden in paragraaf 4.2 uitgelegd met behulp van theorie en een experiment dat is uitgevoerd. Tijdens het experiment wordt een model getraind om de familie van een bloem te herkennen aan de hand van afmetingen van de kelk- en bloemblad. Het volledige experiment is te vinden in bijlage VI. Een van de stappen in een pipeline is het trainen van het model. Om een idee te krijgen van ML en het trainen van modellen zal dit kort uitgelegd worden.

4.1 Wat is machine learning?

ML houdt in dat een computer een taak kan uitvoeren zonder daarvoor expliciet geprogrammeerd te zijn. Dit wordt gedaan door een ML model te laten leren van een gegeven dataset. Vervolgens kan er een voorspelling worden gemaakt [13, p. 1-3].

De domeinen deep learning (DL), neural networks (NN) en artificial intelligence (AI) komen vaak voor als het over ML gaat. Zoals weergegeven in Figuur 4.1 is te zien dat ML een subset is van AI, NN een subset van ML en als laatste DL dat een subset is van NN [14].

Bij AI wordt niet alleen ML toegepast, maar ook concepten zoals beredeneren, plannen, vooruitdenken, onthouden en terug refereren. Een voorbeeld hiervan is dat een ML model kan voorspellen wat het volgende woord in een zin kan zijn, maar een AI kan beredeneren waarom de zin gebouwd is zoals het is en hoe het binnen de context van de alinea past [15].



Figuur 4.1: Machine learning in de context van andere domeinen

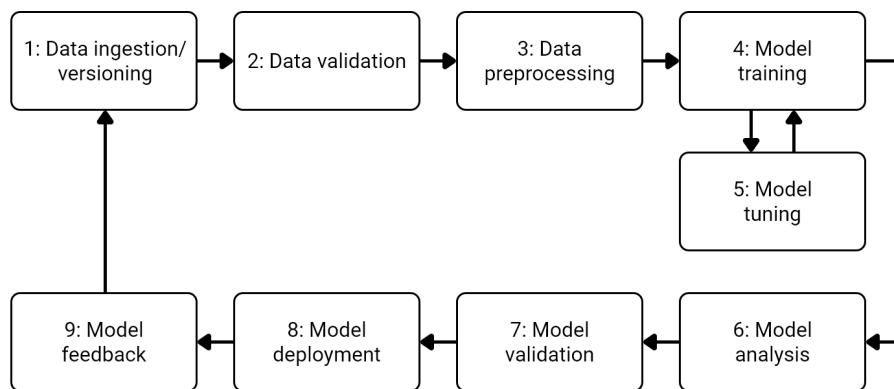
Een NN bestaat uit een collectie van nodes dat gemodelleerd is naar de hersenen. Een NN heeft minimaal 3 lagen: een inputlaag, een verborgen laag en een outputlaag. Elke laag bevat neuronen dat data als input kan krijgen en data als output aan de volgende laag meegeeft. DL is een NN dat meerdere verborgen lagen bevat [16].

4.1.1 Een manier om een machine learning model te trainen

Het trainen van een ML model is een proces waarbij vooraf data wordt opgeschoond en achteraf de prestatie van het model wordt gevalideerd. Normaliter wordt dit gedaan door specifieke code te schrijven voor deze taken. Een valkuil is dat de code niet bij elke developer werkt en schaalt niet in alle gevallen naar een productieomgeving. Een manier om dit wel te behalen is om te werken met ML pipelines. Zoals kort uitgelegd in deelparagraaf 2.1.2 bestaat een ML pipeline uit stappen en acties. Er is echter geen consensus binnen het ML domein over wat de juiste stappen en acties in een pipeline zijn. Voor de scriptie is gekozen voor een pipeline volgens Hapke en Nelson uit het boek "Building Machine Learning Pipelines". Het boek is gemaakt en gepubliceerd door O'Reilly; een bekend en gecrediteerd bedrijf dat boeken maakt binnen het software engineering domein.

4.2 De stappen in een machine learning pipeline

Een ML pipeline begint met het opnemen van data en eindigt met het ontvangen van feedback om de prestatie van het model te verbeteren. De pipeline bevat een aantal stappen zoals data voorbereiden, het model trainen en het uitrollen van het model (Figuur 4.2).



Figuur 4.2: Lifecycle van een model volgens Hapke en Nelson [17, p. 4]

In totaal zijn er negen stappen die elke keer doorlopen kunnen worden om een model te trainen. In de volgende deelpragrafen zullen de stappen worden doorlopen met een korte uitleg over wat er gebeurt in een stap.

4.2.1 Stap 1: Dataopname en versiebeheer (Data ingestion/versioning)

De eerste stap in de pipeline is het opnemen van data. Met deze data zal het model getraind, gevalideerd en getest worden. De dataset kan van een of meerdere bronnen komen, zoals lokaal, een online opslag locatie of van een database. De dataset kan gekopieerd worden en op een plek worden opgeslagen waar alle datasets te vinden zijn. Zodra de data op een bereikbare plek is opgeslagen en is ingeladen in de code, moet het verdeeld worden tussen een train-, validatie- en testdataset. Normaal gebeurt dit met een splitratio van 6:2:2. De train dataset is 60% en de validatie en testdatasets zijn allebei 20% van de originele dataset [17, p. 27-37].

Een use case van een pipeline is dat een nieuw model getraind kan worden door een nieuw dataset te gebruiken. Dit wordt gedaan door de voorgaande dataset te gebruiken, waarbij nieuwe data is toegevoegd. Door het gebruik van verschillende datasets is het verstandig om versiebeheer toe te passen. Zo is goed te zien welk dataset welke model produceert. Een versie geven aan een dataset gebeurt voordat de dataset wordt ingeladen [17, p. 39-40]. Versiebeheer voor datasets kan bijvoorbeeld met DVC [18] of Pachyderm [19]. Beide zijn frameworks om bij te houden waar datasets zijn opgeslagen, welke versie het is en welk model het heeft geproduceerd.

4.2.2 Stap 2: Datavalidatie (Data validation)

Nu de dataset verdeeld is, een versie heeft en op een bereikbare plek is, kan de data gevalideerd worden. Deze stap is vooral belangrijk om te voorkomen dat een model wordt getraind dat niet nuttig is aangezien het trainen veel tijd in beslag kan nemen. Een bekende uitdrukking is "garbage in = garbage out". Dit betekent dat als de dataset niet goed is, het model ook niet goed zal presteren [17, p. 43]. Tijdens de validatiestap wordt gecontroleerd op het volgende:

- Afwijkingen in de dataset
- Wijzigingen in de structuur
- Algemene statistieken in vergelijkingen met voorgaande datasets [17, p. 44]

Bij het controleren van afwijkingen in de dataset wordt gekeken naar waardes die opmerkelijk zijn. Afwijkende waardes liggen te ver van het gemiddelde en kunnen een verkeerd beeld schetsen bij het trainen van het model. Deze uitschieters kunnen simpelweg uit de dataset gefilterd worden.

Het kan voorkomen dat bij een nieuwe dataset het type van waardes zijn gewijzigd. Een *int* kan bijvoorbeeld veranderd zijn in een *string* of *boolean*. Er is dan sprake van een wijziging in de structuur van de dataset. Dit is problematisch omdat er een vertaalstap gemaakt moet worden naar iets bruikbaars. Als dit niet mogelijk is moeten de waardes uitgefilterd worden wat de prestatie van het model negatief kan beïnvloeden.

De algemene statistieken is een hulpmiddel om te controleren op afwijkingen en wijzigingen. Vaak kan de controle in een oogopslag gedaan worden.

4.2.3 Stap 3: Data voorbereiden (Data preprocessing)

Het voorbereiden van de dataset is een stap dat de prestatie van het model verbetert en het proces van het trainen versneld. Deze stap kan verdeeld worden in twee substappen: het opschonen en het optimaliseren van de dataset.

Bij het opschonen worden bijvoorbeeld duplicaten of waardes die onbruikbaar zijn uit de dataset weggehaald. Onbruikbare waardes zijn waardes die simpelweg niet kloppen of verkeerd zijn ingevoerd. Hierbij kan bijvoorbeeld gedacht worden aan een medewerker die de interactietijd met een klant moet bijhouden, maar is vergeten om de eindtijd te noteren. Voor het algoritme zal het dan lijken alsof de medewerker een klant tot sluitingstijd heeft geholpen.

Datasets kunnen geoptimaliseerd worden om twee redenen: algoritmes werken sneller met waardes die dichter bij 0 liggen en algoritmes kunnen niet met elke waarde in de dataset omgaan. Om de efficiëntie van het algoritme te verbeteren kunnen een aantal technieken gebruikt worden:

- Schalen

Bij het schalen van data worden de waarden getransformeerd die liggen tussen een schaal, zoals tussen 0 en 100 of 0 en 1. Bij het schalen wordt het **bereik** getransformeerd [20].

- Normaliseren

Als een dataset wordt gestandaardiseerd, worden de waarden getransformeerd in een standaard normale verdeling waarbij het gemiddelde 0 en de afwijking 1 is. Hierbij wordt de **vorm** van de dataset getransformeerd [20]. Het normaliseren is belangrijk als het algoritme waarden vergelijkt met verschillende eenheden [21]. In sommige gevallen is normalisering een vereiste bij een aantal algoritmes [22].

- Categorische codering

Een groot aantal algoritmes kan alleen omgaan met numerieke waarden. Het kan voorkomen dat datasets categorische waarden bevatten. Dit zijn waarden die een label voorstellen, zoals *first*, *second* of *third*. Om deze waarden te gebruiken moeten ze worden gecodeerd als een numerieke waarde. De label kan bijvoorbeeld als de waarde 1, 2 of 3 gecodeerd worden. Deze techniek heet label encoding of integer encoding en kan gebruikt worden als de volgorde van de codering klopt (deelparagraaf 4.2.3).

Categorical label	Encoded
<i>first</i>	1
<i>second</i>	2
<i>third</i>	3

Figuur 4.3: Voorbeeld van integer encoding

- Hot encoding

Een andere techniek om waarden te coderen is hot encoding. Deze techniek wordt gebruikt waarbij de labels geen relatie met elkaar hebben en maakt gebruik van meerdere waarden om een label te beschrijven (Figuur 4.2.3).

Animal	C1	C2	C3
<i>cat</i>	1	0	0
<i>dog</i>	0	1	0
<i>mouse</i>	0	0	1

Figuur 4.4: Voorbeeld van hot encoding

Dit is een goed moment om de getransformeerde dataset op te slaan als "tussen-dataset" zodat deze stap niet herhaald hoeft te worden. Het voorbereiden van grote datasets kan een grote

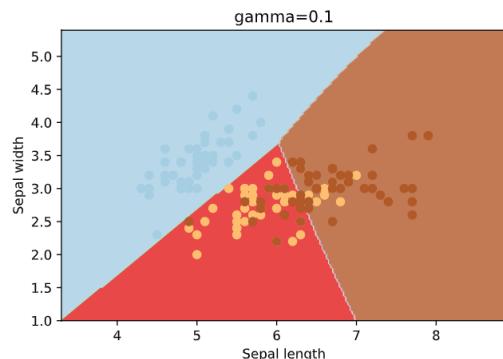
hoeveelheid tijd in beslag nemen. Nu de dataset is gevalideerd en voorbereid, kan het model getraind en getuned worden.

4.2.4 Stap 4 en 5: Model trainen en tunen (Model training and Model tuning)

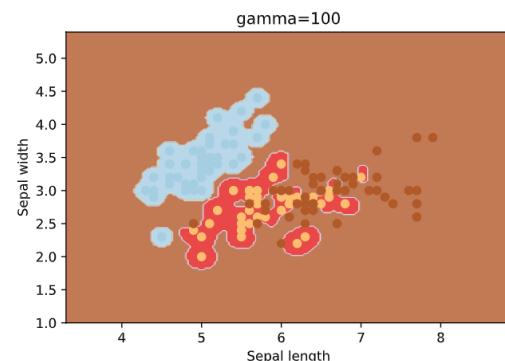
Het trainen van een model gaat door middel van een ML algoritme. Een ML algoritme is vergelijkbaar met een conventioneel algoritme in software engineering zoals bijvoorbeeld binary search, merge sort en depth first search. Het algoritme slaat na het trainen met de dataset regels, nummers en algoritme specifieke datastructuren op in de vorm van een model. Het algoritme kan worden gezien als een specifiek programma waarmee, gegeven een input, voorspellingen mee gedaan kunnen worden [23].

Elk algoritme heeft variabelen om er voor te zorgen dat het algoritme beter aansluit op de dataset. Deze variabelen heten *hyperparameters*. Het algoritme heeft standaard hyperparameters die niet altijd de optimale prestatie behalen. Op een heuristische wijze kunnen de optimale hyperparameters gevonden worden [24]. Dit proces heet *tunen*.

Tijdens het experiment zijn diagrammen gegenereerd met verschillende hyperparameters. In Figuur 4.5 en Figuur 4.6 is te zien hoe de *gamma* hyperparameter van een support vector classification (SVC) algoritme invloed heeft op het model. Dit algoritme classificeert op basis van gegeven waarden. De waarden worden geplot en zal binnen een van de drie kleuren, licht blauw, bruin of rood, vallen. Het model met als hyperparameter waarde 100 zal vaker de classificatie met de bruine kleur als voorspelling geven dan als de waarde 0.1 zou zijn.



Figuur 4.5: Gamma hyperparameter van een support vector classification (SVC) algoritme met waarde 0.1



Figuur 4.6: Gamma hyperparameter van een support vector classification (SVC) algoritme met waarde 100

In het ML domein bestaan talloze algoritmes om voorspellingen te maken. In bijlage V is een mind-map te vinden van de algoritmes die gedurende de scriptie naar voren zijn gekomen.

4.2.5 Stap 6 en 7: Modelanalyse en -validatie (Model analysis and Model validation)

Na het trainen en tunen kan analyse en validatie plaatsvinden. Bij het analyseren wordt de prestatie van het model vergeleken met voorgaande modellen. Om dit te doen kunnen, net

als in deelparagraaf 4.2.2, statistieken gegenereerd worden van het model. De soort statistiek hangt af van wat voor soort algoritme is gebruikt. Een classificatie-algoritme heeft bijvoorbeeld andere statistieken dan een regressie-algoritme. Op basis van de uitkomst kunnen de dataset of hyperparameters aangepast worden om de accuraatheid te verhogen.

Met de validatie van een model moet er gekeken worden met een genuanceerd perspectief. Validatie gaat namelijk over hoe eerlijk een model is als er gekeken wordt naar een bepaalde groep. Een groep kan gedefinieerd worden als geslacht, etniciteit, locatie of leeftijd. Het kan namelijk voorkomen dat de dataset voornamelijk bestaat uit bijvoorbeeld vrouwen. Dit *kan* er voor zorgen dat het model minder accurate voorspellingen maakt voor mannen [13, p. 109-110]. Een voorbeeld waar het gebruik van ML grote negatieve gevolgen had was de toeslagenaffaire in 2020. De Nederlandse overheid maakte gebruik van een systeem dat aangaf of een burger mogelijk bijstandsfraude had gepleegd. Het systeem maakte gebruik van ML om fraude aan te geven. Jaren na het gebruik bleek dat er sprake was van etnische profilering. De dataset bestond voornamelijk uit immigranten uit Islamitische landen. Het systeem heeft hierdoor een *bias* [25].

4.2.6 Stap 8: Model uitrollen (Model deployment)

Na het analyseren en valideren kan het model uitgerold worden. Het uitrollen kan op twee manieren: inbakken in een app of serveren met een application programming interface (API).

Met het inbakken in een app wordt bedoeld dat het model meegenomen wordt als de app gebouwd wordt. Voordelen van deze methode is dat het model offline bruikbaar is en de rekenkracht van het apparaat gebruikt kan worden om te voorspellen of classificeren. Een nadeel is dat het model niet makkelijk geüpdatet kan worden.

De andere optie is om het model te serveren door middel van een API. De app die het model wil gebruiken zal een request moeten doen met input. De API geeft de input door aan het model. Het model geeft een predictie terug dat de API doorstuurt naar de applicatie. Het updaten van het model gaat gemakkelijker dan het model inbakken met een app. Een nadeel is de eis dat de app een internetverbinding moet hebben om het model te gebruiken [13, p. 130].

Bij het experiment is gekozen voor het uitrollen met een API. In Figuur 4.7 is te zien hoe een model geserveerd wordt via een API. Een endpoint genaamd "classify" is gedefinieerd waarop een *POST* request gedaan kan worden. De endpoint haalt de input uit de request en geeft het door aan het model. Het model maakt vervolgens een predictie dat uiteindelijk naar de app teruggestuurd wordt.

```
1 @app.route("/classify", methods=["POST"])
2 def classify():
3     # Get input from request
4     json = request.get_json()
5     data = json["data"]
6
7     # Get prediction from model
8     classification = model.predict([data])
9
10    # Return prediction to app
11    return jsonify({ "family": classification[0] })
```

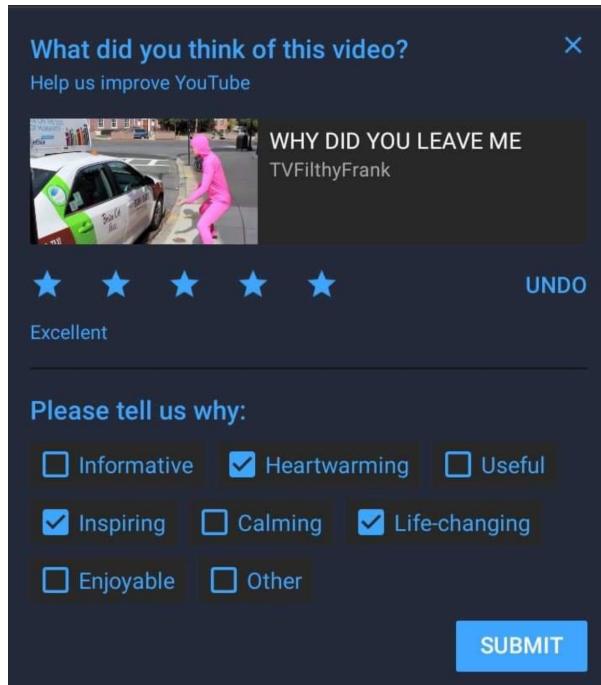
Figuur 4.7: Model uitgerold met behulp van een API

4.2.7 Stap 9: Feedbackloop (Model feedback)

Om het model continue te verbeteren kan feedback verzameld worden. De feedback geeft een beeld van de effectiviteit van het model in een productieomgeving. Feedback kan verdeeld worden in twee soorten: impliciet en expliciet [13, p. 264].

Het verzamelen van impliciete feedback wordt gedaan zonder dat de gebruiker zich daar bewust van is. Dit wordt gedaan door bij te houden of een voorspelling correct was volgens een gebruiker. Een voorbeeld van deze methode is het bijhouden of een gebruiker een video kijkt dat door het algoritme voorgesteld is. Als een gebruiker de voorgestelde video bekijkt, wordt dit gezien als een succesvolle voorspelling. Mocht de gebruiker niet een voorgestelde video bekijken, wordt dit gezien als onsuccesvol.

In tegenstelling tot impliciet wordt bij expliciete feedback gevraagd aan de gebruiker of een voorspelling gepast is. Een voorbeeld van deze vorm in de praktijk is hoe YouTube expliciete feedback vraagt (Figuur 4.8). Naast het aangeven of de voorspelling gepast was, kan de gebruiker ook aangeven waarom de video gepast was.



Figuur 4.8: Vorm van expliciete feedback gebruikt door YouTube

Het verzamelen van feedback is niet noodzakelijk maar helpt met het verbeteren van het model. Deze output van deze stap gaat samen met een nieuwe dataset naar de eerste stap in de lifecycle (Figuur 4.2). Met de nieuwe dataset en de feedback kan een nieuw model getraind worden dat beter presteert dan de voorganger.

4.3 Conclusie

In dit hoofdstuk is er onderzoek gedaan naar het antwoord op de deelvraag: **D1: Waar bestaat een machine learning pipeline uit?** Het onderzoek is uitgevoerd met behulp van zowel digitale bronnen, een boek en een experiment.

Een ML pipeline bestaat uit verschillende stappen die op hun beurt bestaan uit acties. De stappen zijn gericht op het waarborgen van de kwaliteit van de dataset en model, het voorbereiden van de dataset en het trainen van het model. Daarnaast kan het model uitgerold worden in een productieomgeving en kan er feedback verzameld worden over de prestatie van het model volgens gebruikers.

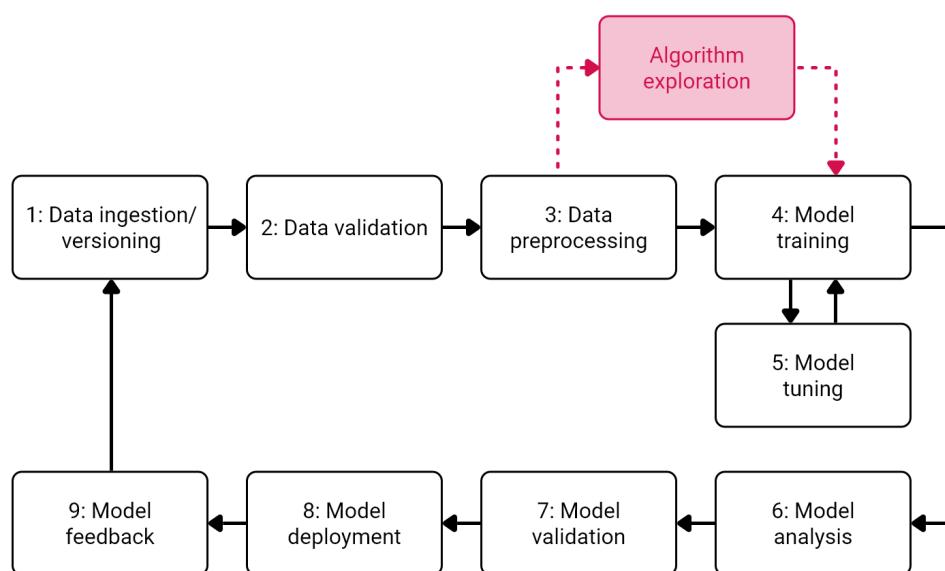
De acties in de stappen moeten bij het eerste gebruik van de pipeline gedefinieerd worden. De acties kunnen bijvoorbeeld het filteren van onbruikbare data, een model trainen of statistieken genereren zijn. Als dit vast staat, kan de pipeline hergebruikt worden om nieuwe modellen te trainen met nieuwe datasets. Op deze manier is een ML pipeline een gestructureerde en reproduceerbaar proces.

4.4 Advies

Het in gebruik nemen van een ML pipeline werkwijze is initieel ingewikkeld, maar lonend op de lange termijn. Het opzetten van de pipeline en de acties definiëren kost tijd en kennis. Echter als dit eenmaal gedaan is, kan eenvoudig een verbeterd model worden getraind door een nieuwe dataset aan te leveren. Een verbeterd model is vrijwel gegarandeerd omdat de stappen in de ML pipeline de dataset en het model analyseren en valideren.

4.4.1 Een betere keuze maken tussen machine learning algoritmes

Zoals eerder aangegeven in paragraaf 4.1 bestaat er niet één ML pipeline dat correct is. Een ML pipeline is een werkwijze waarbij de stappen anders geïnterpreteerd kunnen worden. Hierdoor is er geen regelmaat tussen verschillende pipelines. Een pipeline kan stappen of acties weglaten of extra hebben. Overeenkomsten tussen pipelines zijn er ook niet aangezien stappen gemaakt kunnen worden die specifiek zijn voor het algoritme of workflow van het bedrijf. De stappen in de pipeline van Hapke en Nelson (Figuur 4.2) is een valide basis, maar kan uitgebreid worden om de ervaring van developers te verbeteren. In stap 4 en 5 (deelparagraaf 4.2.4) wordt een model getraind waarbij het beste algoritme om het ML probleem op te lossen al bekend is. De aannname is dat de keuze of het onderzoek vóór het starten van de pipeline is gedaan. Tussen stap 3 en 4 kan echter een stap tussen zitten om te experimenteren met een combinatie van verschillende algoritmes en hyperparameters. In Figuur 4.9 is te zien hoe zo een stap zou passen in de lifecycle. De stap is met een stippe lijn aangegeven omdat de stap de eerste keer meegenomen kan worden als de pipeline wordt doorlopen. Na de eerste keer zijn de beste algoritme en hyperparameters al bekend en kan deze stap overgeslagen worden.



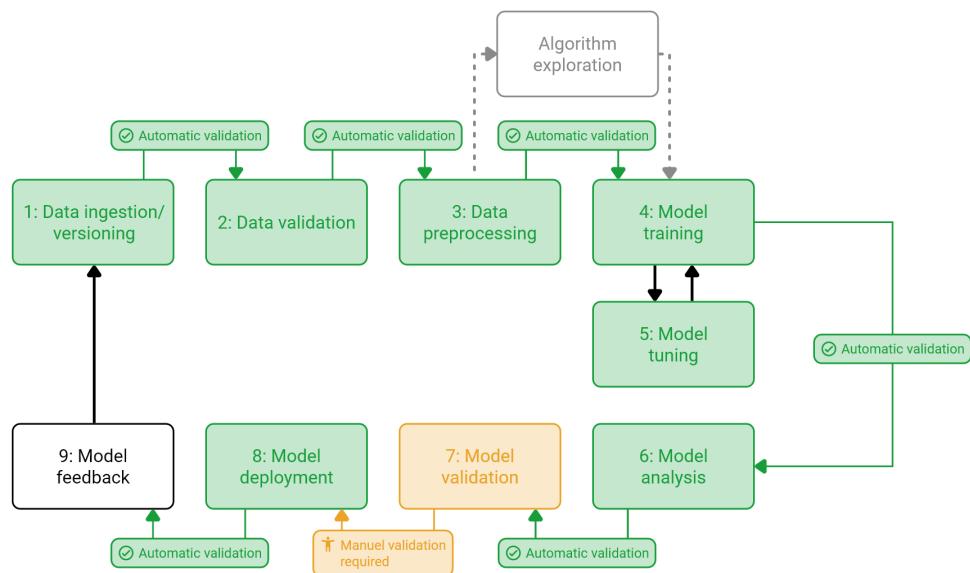
Figuur 4.9: Machine learning pipeline met een algoritme exploratie tussenstap

In de "Algorithm exploration" stap kunnen verschillende algoritmes die hetzelfde doel bereiken en een variatie van hyperparameters voor elk algoritme getest worden tegen de dataset. Uit deze tests kan een prestatiescore gegenereerd worden om vervolgens te kiezen voor het algoritme dat

het meest geschikt lijkt. Na de keuze kan de pipeline verder gaan met de volgende stappen.

4.4.2 Machine learning versimpelen

Een van de focuspunten is om te onderzoeken in hoeverre ML te vergemakkelijken is voor developers. Met de huidige pipeline is er vrij veel kennis vereist over ML om ermee te werken. Toch kan een groot deel van de stappen geautomatiseerd worden. Tussen elke stap kan een validatie plaatsvinden met een aantal randvoorwaarden waaraan voldaan moeten worden om door te gaan naar de volgende stap. In Figuur 4.10 is te zien hoe tussen elke stap een validatie plaats vindt. Om bijvoorbeeld van stap 6 (Model analysis) naar stap 7 (Model validation) te gaan, moet het model even goed of zelfs beter presteren dan de voorganger. Hoe veel beter het model moet kunnen presteren is aan een developer. Mocht het zo zijn dat het model niet voldoet, kan de pipeline stopgezet worden.



Figuur 4.10: Machine learning pipeline waarbij stappen geautomatiseerd kunnen worden voor het gemak van developers

Niet alle aspecten van de pipeline kunnen versimpeld worden. Een van de stappen die niet kan is stap 7 (Model validation) in Figuur 4.10. Deze stap vereist input van een mens aangezien de bias van het model wordt beoordeeld. Wel kunnen rapporten over de bias gegenereerd worden zodat de workflow van developers gestroomlijnd wordt.

5 Orkestratietools om platformen te beheren

Om resources in een cloud platform te beheren kan gebruik gemaakt worden van een portaal. In een portaal kunnen databases, servers en netwerken aangemaakt worden om een infrastructuur te creëren. Omdat een portaal uitgebreid en complex kan zijn, moet de developer een infrastructuur kunnen opzetten zonder het gebruik van een portaal. Dit is mogelijk met infrastructure as code (IaC) orkestratietools. IaC zal als eerst worden uitgelegd. Vervolgens worden verschillende IaC orkestratietools met elkaar vergeleken. Als laatst wordt een experiment uitgevoerd om de haalbaarheid met een orkestratietool te valideren.

5.1 Wat is infrastructure as code?

IaC is een manier om een infrastructuur op te zetten binnen een cloud platform zonder gebruik te maken van een interface zoals een portaal. De gewenste infrastructuur wordt beschreven waarna de orkestratietool de gewenste situatie een realiteit probeert te maken [26].

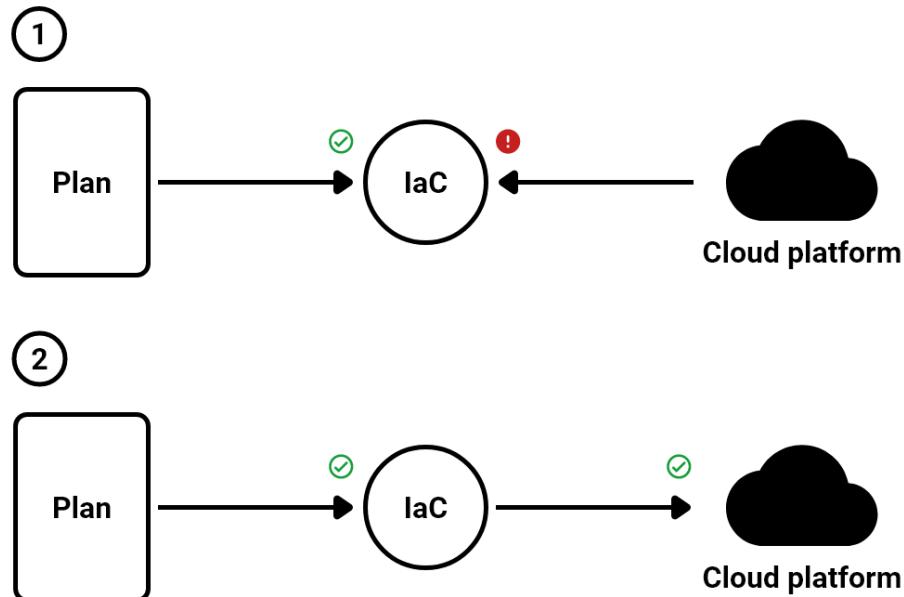
De manier waarmee een orkestratietool de infrastructuur aanmaakt, wijzigt of verwijderd is met een plan. In het plan staat welke resources moeten bestaan om aan de gewenste situatie te voldoen. Vaak gaat dit gepaard met configuratie zoals een naam, locatie of tags. In Figuur 5.1 is een voorbeeld van een plan als een YAML-bestand te zien. YAML is een bestandstype dat vaak wordt gebruikt voor configuratie. Het plan beschrijft op welke cloud platform dit uitgevoerd moet worden (regel 12) en wat er aangemaakt moet worden. In dit geval is het een resource group (regel 16) met als configuratie een naam (*myTFResourceGroup*) en locatie (*westus2*).

```
1 terraform {
2   required_providers {
3     azurerm = {
4       source  = "hashicorp/azurerm"
5       version = "> 2.26"
6     }
7   }
8
9   required_version = "> 0.14.9"
10 }
11
12 provider "azurerm" {
13   features {}
14 }
15
16 resource "azurerm_resource_group" "rg" {
17   name      = "myTFResourceGroup"
18   location  = "westus2"
19 }
```

Figuur 5.1: Voorbeeld van een infrastructure as code (IaC) plan [27]

Een IaC volgt een proces om aan de gewenste situatie te voldoen. In Figuur 5.2 is te zien hoe zo een proces verloopt. Zodra de orkestratietool start, wordt zowel het plan als de situatie in de cloud platform gecontroleerd en met elkaar vergeleken. In stap 1 is het plan valide, maar de situatie in de cloud platform niet. Dit betekent bijvoorbeeld dat er in het plan staat dat er een resource group moet bestaan met een naam en locatie zoals het beschreven staat in Figuur 5.1, maar dit niet bestaat in de cloud platform. In de volgende stap wordt de resource aangemaakt

door de IaC om aan de gewenste situatie te voldoen. De wijzigingen gebeuren programmatisch en wordt gedaan door de IaC. Er is geen tussenkomst van een persoon of interface nodig.



Figuur 5.2: Proces van een infrastructure as code (IaC) om aan de gewenste situatie te voldoen

5.2 Orkestratietools vergeleken met elkaar

Om een keuze te maken tussen IaC orkestratietools kunnen zogenoemde "knock-out" criteria opgesteld worden. Een IaC orkestratietool valt af zodra de tool niet voldoet aan één van de criteria. In Tabel 5.1 zijn de knock-out criteria te vinden. De criteria zijn samenwerking met NGTI opgesteld.

Tabel 5.1: Knock-out criteria voor orkestratietools dat cloud computing platformen beheert

Criteria	Toelichting
Programmatisch beheren	Het orkestratietool moet via code een resources kunnen aanmaken, wijzigen en verwijderen.
Ondersteuning voor cloud computing platformen	Om platform-agnostisch te zijn moet de orkestratietool minstens twee cloud computing platformen ondersteunen waarop een ML model getraind kan worden.
Uitgebreide documentatie	De documentatie moet toegankelijk en duidelijk zijn. Ook moet de documentatie moet tutorials, concepten en een reference bevatten.

Voor de keuzes uit de tools is onderzoek gedaan op het internet en een enquête geraadpleegd van Stack Overflow [28]. De enquête is ingevuld door developers en bevat vragen over de ervaring met frameworks, technologieën en de website zelf. Uit de vraag over welke frameworks, libraries en tools het meest gebruikt werkt, kwam Ansible, Terraform, Puppet en Chef naar boven [29]. Volgens een vraag over de meest geliefd en gevreesde frameworks, libraries en tools zijn Ansible en Terraform het meest geliefd en Puppet en Chef het meest gevreesd [30]. Uit algemeen onderzoek naar orkestratietools is, naast de bovengenoemde tools, Pulumi naar voren gekomen als kandidaat [31]. Ansible, Terraform en Pulumi zijn in Tabel 5.2 tegen de knock-out criteria gehouden.

Tabel 5.2: Knock-out criteria tegen orkestratietools dat cloud computing platformen beheerd

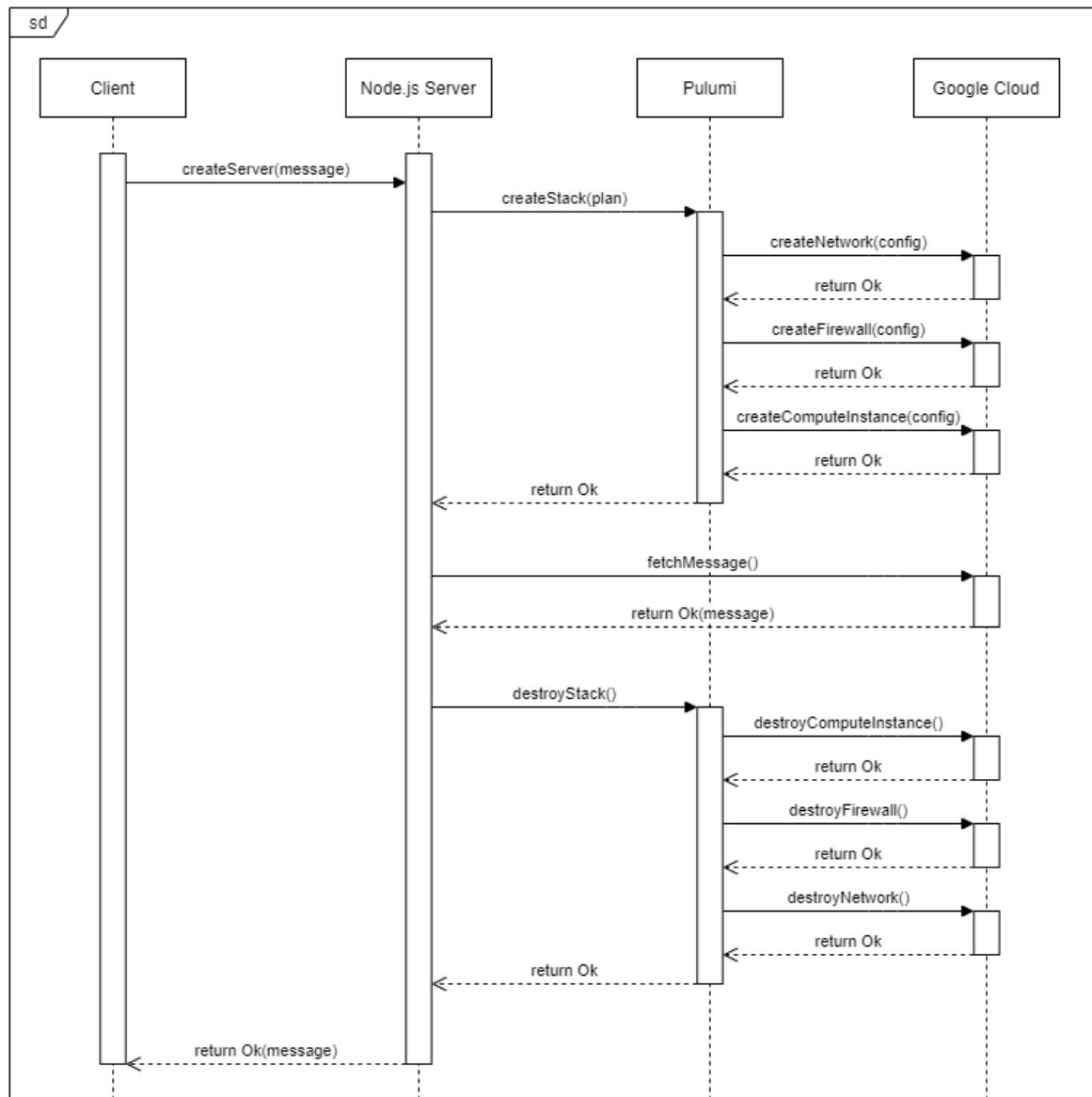
Orkestratie-tools	Programmatisch beheren	Ondersteuning voor cloud computing platformen	Uitgebreide documentatie
Ansible	Nee, orkestratie gaat via een YAML-bestand [32]	AWS, Google Cloud en Azure. 11 platformen in totaal [33]	Uitgebreide documentatie met tutorials, references en migration guide [34]
Pulumi	Ja, orkestratie via CLI of TypeScript, JavaScript, Python, .NET en Go [35]	AWS, Google Cloud en Azure. 13 platformen in totaal [36]	Uitgebreide documentatie met tutorials en references [37]
Terraform	Nee, orkestratie gaat via een YAML-bestand [38]	AWS, Google Cloud en Azure. 5 platformen in totaal [39]	Uitgebreide documentatie met tutorials, references en video's [40]

In Tabel 5.2 is te zien dat de drie orkestratietools gelijk zijn als het gaat om de ondersteuning van cloud platform en documentatie. Programmatisch beheren kan met Ansible en Terraform alleen door middel van een YAML-bestand in tegenstelling tot Pulumi dat via code een plan kan uitvoeren. Omdat het praktischer voor de PoC is om via code een plan uit te voeren, is er voor gekozen om met Pulumi verder te werken. In het volgende hoofdstuk zal een experiment met Pulumi uitgevoerd worden om te valideren of Pulumi geschikt is.

5.3 Experiment met de orkestratietool Pulumi

Het idee achter het experiment is om te achterhalen of Pulumi een geschikte orkestratietool is. Op dit moment in het afstudeerproces wordt er verwacht dat de PoC een frontend krijgt dat praat door middel van een representational state transfer (REST) API met een backend. De backend zal vervolgens gebruik maken van de orkestratietool om te praten met een cloud platform. Om de PoC na te bootsen op kleinere schaal wordt er bij het experiment gebruik gemaakt van een client waarbij een opdracht met behulp van een REST API naar een Node.js server wordt verstuurd. De Node.js server stuurt om zijn beurt opdrachten naar Pulumi die communiceert met Google Cloud. Voor het experiment is arbitrair gekozen voor het cloud platform; de PoC moet ten slotte platform-agnostisch zijn.

Het experiment is als sequence-diagram uitgewerkt in Figuur 5.3. Hierin is goed te zien welke acties gedaan worden en door welke actor. In de sequence-diagram is ook te zien wat voor taak wordt uitgevoerd aan de hand van de naam tussen de actoren.



Figuur 5.3: Sequence-diagram van het experiment met orkestratietool Pulumi

De Node.js server is vrij simpel en bevat één endpoint op regel 12 genaamd /api/run-python-code (Figuur 5.4). In deze *POST* endpoint wordt een stack aangemaakt, een *GET* request uitgevoerd naar de aangemaakte compute instance en vervolgens wordt de stack weer verwijderd. Een stack is de manier hoe Pulumi een plan beschrijft.

```
1 import express, { response } from 'express'
2 import fetch from 'node-fetch'
3 import { compute } from '@pulumi/gcp'
4 import { LocalWorkspace } from "@pulumi/pulumi/automation"
5
6 const app = express()
7 const port = 8080
8
9 app.use(express.json())
10 app.use(express.urlencoded({ extended: true }))
11
12 app.post("/api/run-python-code", async (req, res) => {
13
14   app.listen(port, () => {
15     console.log(`server started at http://localhost:${port}`)
16   })
17 })
```

Figuur 5.4: Node.js server voor het experiment met orkestratietool Pulumi

In de stack worden drie resources aangemaakt: een network, een firewall en een compute instance (Figuur 5.5). De compute instance moet een HMTL bestand serveren met een bericht dat de client heeft gespecificeerd.

```
1 const CreateResource = (message: string) => async () => {
2   // Create a network
3   const computeNetwork = new compute.Network("network", {
4     // Configuration
5   })
6
7   // Create a firewall in the network
8   const computeFirewall = new compute.Firewall("firewall", {
9     // Configuration
10  })
11
12  // Define script that runs when the compute instance runs
13  const startupScript = `#!/bin/bash
14 echo "${message}" > index.html
15 nohup python -m SimpleHTTPServer 80 &`  

16
17  // Create a compute instance that runs startupScript
18  const computeInstance = new compute.Instance("instance", {
19    // Configuration
20  }, { dependsOn: [computeFirewall] })
21
22  const ip = computeInstance.networkInterfaces.apply(ni => ni[0].accessConfigs![0].natIp);
23
24  return { name: computeInstance.name, ip: ip }
25 }
```

Figuur 5.5: De stack voor het experiment.

Elke resource heeft zijn eigen configuratie om te zorgen dat het HTML bestand beschikbaar is voor de buitenwereld. De code met de volledige configuratie is te vinden in bijlage VII.

In Figuur 5.6 is te zien wat er in de POST endpoint gebeurt. Als eerst wordt een stack in geheugen aangemaakt met behulp van de code in Figuur 5.5. Na het aanmaken wordt de stack uitgevoerd met de *stack.up()* functie op regel 12. Nadat de stack is aangemaakt in Google Cloud wordt een *GET* request uitgevoerd naar de server op regel 15. Nadat de request voltooid is wordt de stack verwijderd uit Google Cloud (regel 21) met *stack.destroy()* en wordt het bericht teruggestuurd naar de client.

```
1 app.post("/api/run-python-code", async (req, res) => {
2   const stackName = 'mlpa-py'
3   const projectName = 'disco-sky-312109'
4
5   const stack = await LocalWorkspace.createOrSelectStack({
6     stackName,
7     projectName,
8     program: CreateResource(req.body.message)
9   })
10
11 // Comparing stack against Google Cloud
12 const upRes = await stack.up({ onOutput: console.info })
13
14 // Fetch HTML file with message
15 const action_response = await fetch(`http://${upRes.outputs.ip.value}`)
16   .then(async r => await r.text())
17   .then(t => t)
18   .catch(e => console.log(e))
19
20 // Destroying stack in Google Cloud
21 await stack.destroy({ onOutput: console.info })
22
23 res.send({ returned_response: action_response })
24 })
```

Figuur 5.6: POST endpoint van de Pulumi experiment

5.4 Conclusie

In dit hoofdstuk is er onderzoek gedaan naar het antwoord op de deelvraag: **D2: Hoe kan een orkestratietool verschillende cloud computing platformen beheren om een machine learning pipeline op te zetten?** Het onderzoek is uitgevoerd met behulp van digitale bronnen en een experiment.

Een orkestratietool is een manier om resources aan te maken, aan te passen en te verwijderen. De tool doet dit zonder het gebruik van een online portaal of interventie van een developer. Het managen van resources wordt gedaan op basis van een plan en de huidige situatie binnen de cloud platform. In het plan staat welke resources moeten bestaan en met welke configuratie. Het plan kan beschreven worden in een YAML-bestand of door middel van code. De orkestratietool vergelijkt vervolgens het plan met de huidige situatie en zorgt er voor dat de situatie voldoet aan het plan.

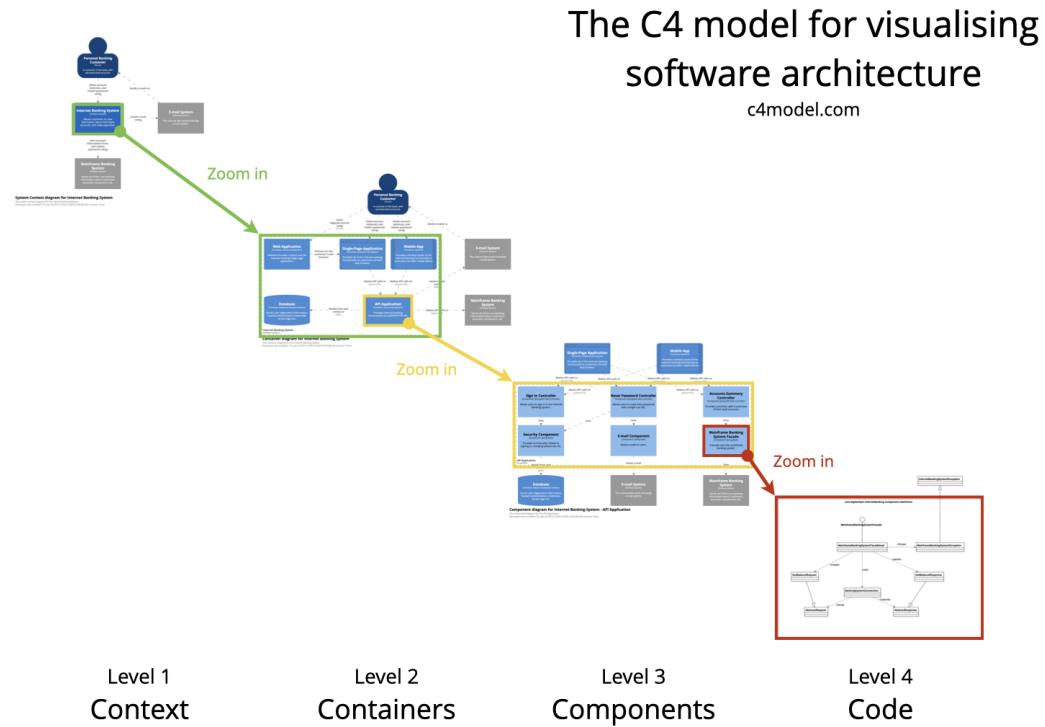
De orkestratietool Pulumi is een valide optie uitgaand van het experiment. Pulumi heeft uitstekende documentatie en laat middels output zien wat er gebeurd als Pulumi wordt uitgevoerd. Ook is de manier waarop Pulumi werkt (Figuur 5.5) simpel en makkelijk te volgen.

6 Architectuurontwerp van de oplossing

Nu het bekend is hoe ML pipelines en orkestratietools werken, kan er gekeken worden naar hoe de PoC architectonisch eruit gaat zien. De technische tekeningen laten zien hoe verschillende componenten in de PoC interacteren met elkaar en met componenten buiten de scope van de PoC. De tekeningen zijn gemaakt volgens het C4 model [41]. In dit hoofdstuk zal worden uitgelegd hoe C4 gelezen moet worden waarna de technische tekeningen wordt uitgelegd.

6.1 Het C4 model

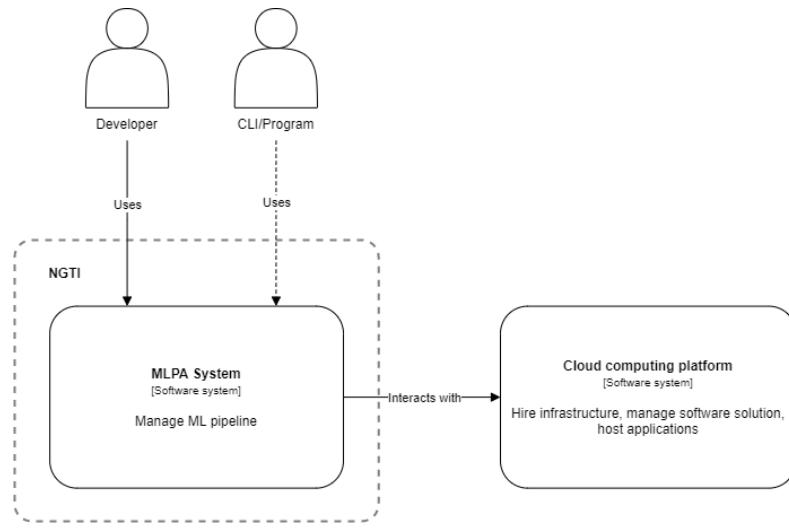
Het C4 model is een notatietechniek om de architectuur van een systeem te communiceren aan andere developers en is bedacht door Simon Brown [41]. Bij het tekenen van de modellen wordt er gedacht in vier niveaus: context, containers, components en code. In Figuur 6.1 is een voorbeeld van een C4 model te zien. De context level is het eerste niveau waarbij wordt gekeken naar het systeem, gebruikers en contexten buiten de scope. Het systeem binnen de scope kan uitgebreid worden door middel van het volgende niveau: containers. Containers laten op een hoog niveau zien waar het systeem uit bestaat. Elk container kan verder verduidelijkt worden met het componentenniveau. Hierbij worden containers verder opgebroken in components. Het laatste niveau is het code niveau en wordt doorgaans weergegeven via een klassendiagram.



Figuur 6.1: Voorbeeld van een C4 model [41]

6.2 Niveau 1: Context

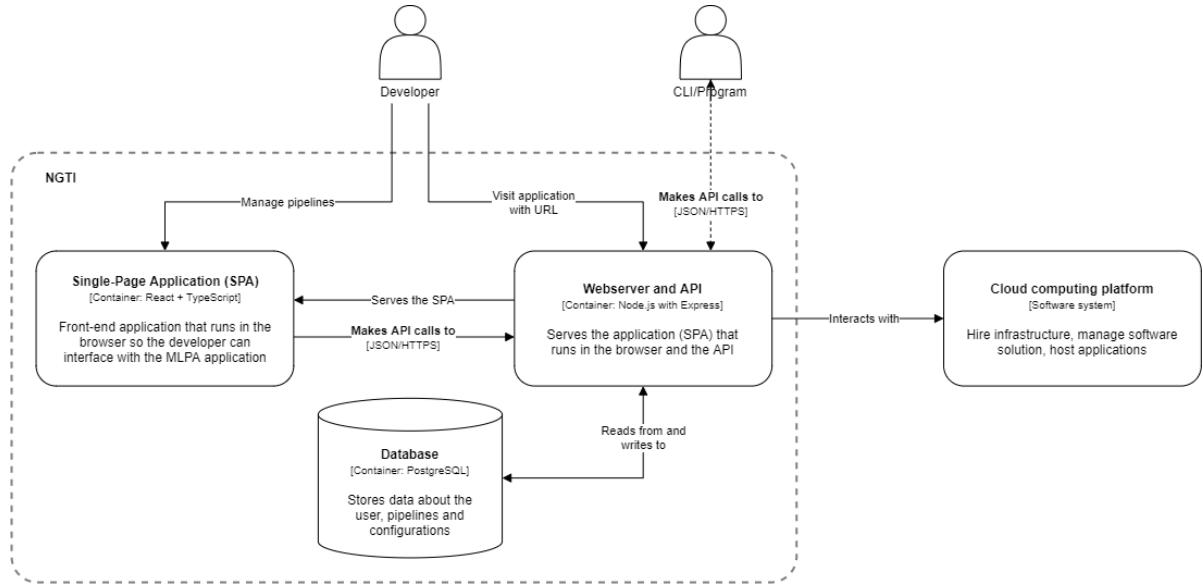
Op contextniveau is te zien dat developers gebruik maken van het machine learning pipeline automation (MLPA) systeem (Figuur 6.2). Het MLPA systeem interacteert met cloud platformen. In dit diagram is ook te zien dat developers alleen gebruik maken van het systeem en indirect van cloud platformen. In de context diagram is ook een command line interface (CLI)/program te zien dat gebruik kan maken van het MLPA systeem. Deze valt echter buiten de scope en is daarom aangegeven met een stippellijn.



Figuur 6.2: Context niveau diagram van het architectuurontwerp

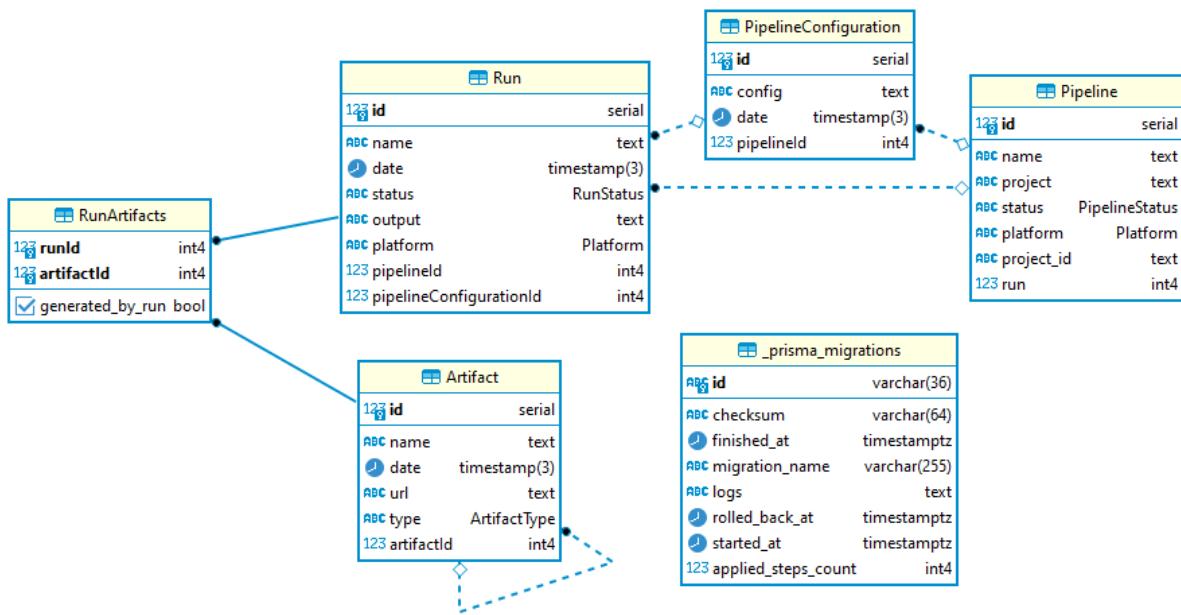
6.3 Niveau 2: Container

In Figuur 6.3 is het container niveau van het MLPA systeem te zien. Het systeem bestaat uit een single page application (SPA) frontend, een Node.js met Express backend en een PostgreSQL database. Een developer krijgt van de backend de SPA geserveerd waarmee een pipeline beheert kan worden. De frontend stuurt API requests naar de backend die vervolgens interacteert met de database en cloud platformen.



Figuur 6.3: Containerniveau-diagram van het architectuurontwerp

Voor het database is een entity relation diagram (ERD) gemaakt (Figuur 6.4). De ERD laat zien welke tabellen het database bevat en hoe de tabellen in relatie tot elkaar staan. De *Pipeline* tabel heeft een one-to-many relatie met tabellen *PipelineConfiguration* en *Run*. Artificaten worden in de *Artifact* tabel opgeslagen en heeft een many-to-many relatie met *Run*. In de *_prisma_migrations* tabel worden migraties van het database bijgehouden.



Figuur 6.4: Entity relation diagram (ERD)

Voor de keuze voor de frameworks voor elke container is gekeken naar de populariteit en de kwaliteit van documentatie. In Tabel 6.1 zijn de criteria verder toegelicht.

Tabel 6.1: Criteria voor frameworks

Criteria	Toelichting
Populariteit	De populariteit zegt veel over hoe robuust een framework is. Hierbij wordt gekeken naar hoe vaak het framework wordt gebruikt en wie er achter het framework zit.
Documentatie	De documentatie moet toegankelijk en duidelijk zijn. Ook moet de documentatie moet tutorials, concepten en een reference bevatten.

Voor de opties is de enquête van Stack Overflow geraadpleegd [28]. Voor de frontend zijn React, Angular en Vue de meest gebruikte frameworks [42]. Angular is het minst populair uit de drie opties [43] en is daardoor niet meegenomen in de keuze. Voor de backend zijn er twee opties: Node.js en .NET Core [29]. In Tabel 6.2 zijn de frontend frameworks tegen de criteria gehouden en in Tabel 6.3 de backend frameworks.

Tabel 6.2: Criteria voor frontend frameworks

Framework	Populariteit	Documentatie
React	Gebruikt door 35.9% van developers volgens enquête [42]. Framework is gemaakt door Facebook.	Zeer uitgebreide documentatie met tutorials, references en migration guides [44]
Vue	Gebruikt door 17.3% van developers volgens enquête [42]. Framework is gemaakt door Evan You.	Zeer uitgebreide documentatie met tutorials, references en migration guides [45]

Tabel 6.3: Criteria voor backend frameworks

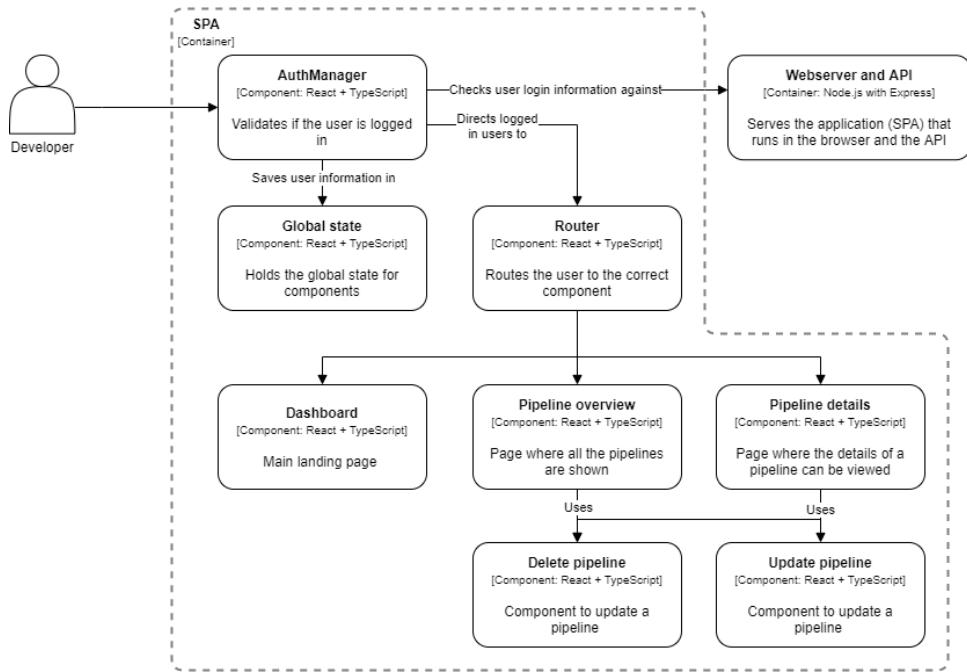
Framework	Populariteit	Documentatie
Node.js	Gebruikt door 51.4% van developers volgens enquête [29]. Framework is gemaakt door OpenJS Foundation.	Beknopte documentatie met reference en tutorials [46].
.NET Core	Gebruikt door 26.7% van developers volgens enquête [29]. Framework is gemaakt door Microsoft.	Uitgebreide maar overweldigende documentatie [47].

React en Vue hebben beide sterke documentatie. Echter wordt React vaker gebruikt dan Vue; voor de frontend is daarom gekozen voor React. De keuze voor het framework van de backend is makkelijker te maken. De documentatie van .NET Core is uitgebreid maar kan snel overweldigend omdat er zo veel is. Voor de backend is daarom gekozen voor Node.js.

Voor de database zijn er geen specifieke requirements vanuit NGTI. Volgens Stack Overflow is MySQL de meeste gebruikte database met 55.6% gevolgd door PostgreSQL met 36.1% [48]. Na verder onderzoek naar beide databases blijkt dat het verschil nihil is. De voorkeur gaat daarom naar PostgreSQL omdat met deze database eerder is gewerkt.

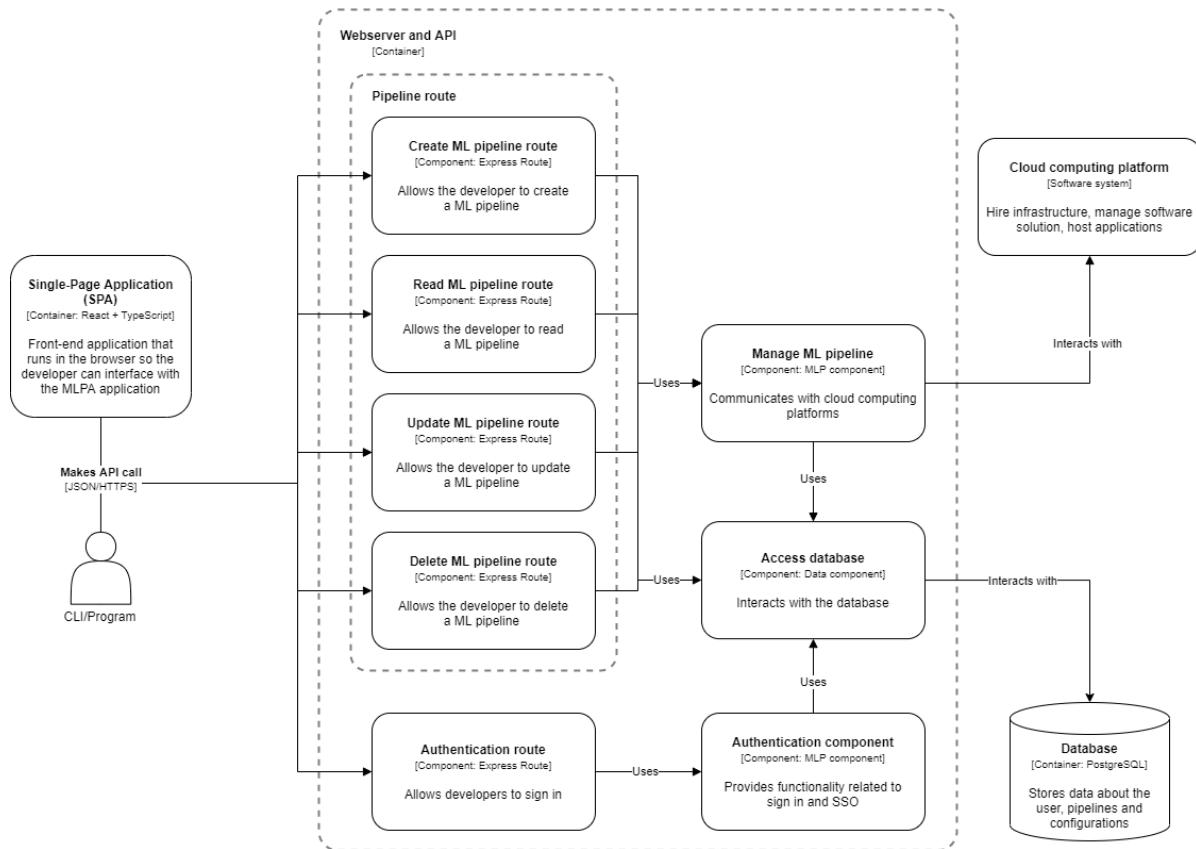
6.4 Niveau 3: Component

De component overview van de SPA is te zien in Figuur 6.5 en van de webserver met de API in Figuur 6.6. De SPA bevat een *AuthManager* component dat controleert of de developer is ingelogd. Als dat het geval is, stuurt de *Router* component de developer door naar de *Dashboard*, *Pipeline overview* of *Pipeline details* pagina.



Figuur 6.5: Single page application (SPA) componentniveau diagram van het architectuurontwerp

Het component diagram van de Node.js server bevat het complexe gedeelte (Figuur 6.6). Hier komt een API request binnen in een *ExpressRoute*. De route maakt vervolgens gebruik van Pulumi en de database om taken uit te voeren zoals het starten van een pipeline, het uploaden van een dataset of de status van een run bekijken. De keuze voor Pulumi is toegelicht in paragraaf 5.2.



Figuur 6.6: Node.js server componentniveau diagram van het architectuurontwerp

De architectuur van de Node.js server is ontworpen volgens het repository-service pattern [49]. Dit is een pattern dat bepaalde zaken van elkaar scheidt zodat het project onderhoudbaar blijft. In paragraaf 7.5 zal dit uitgelegd worden met voorbeelden.

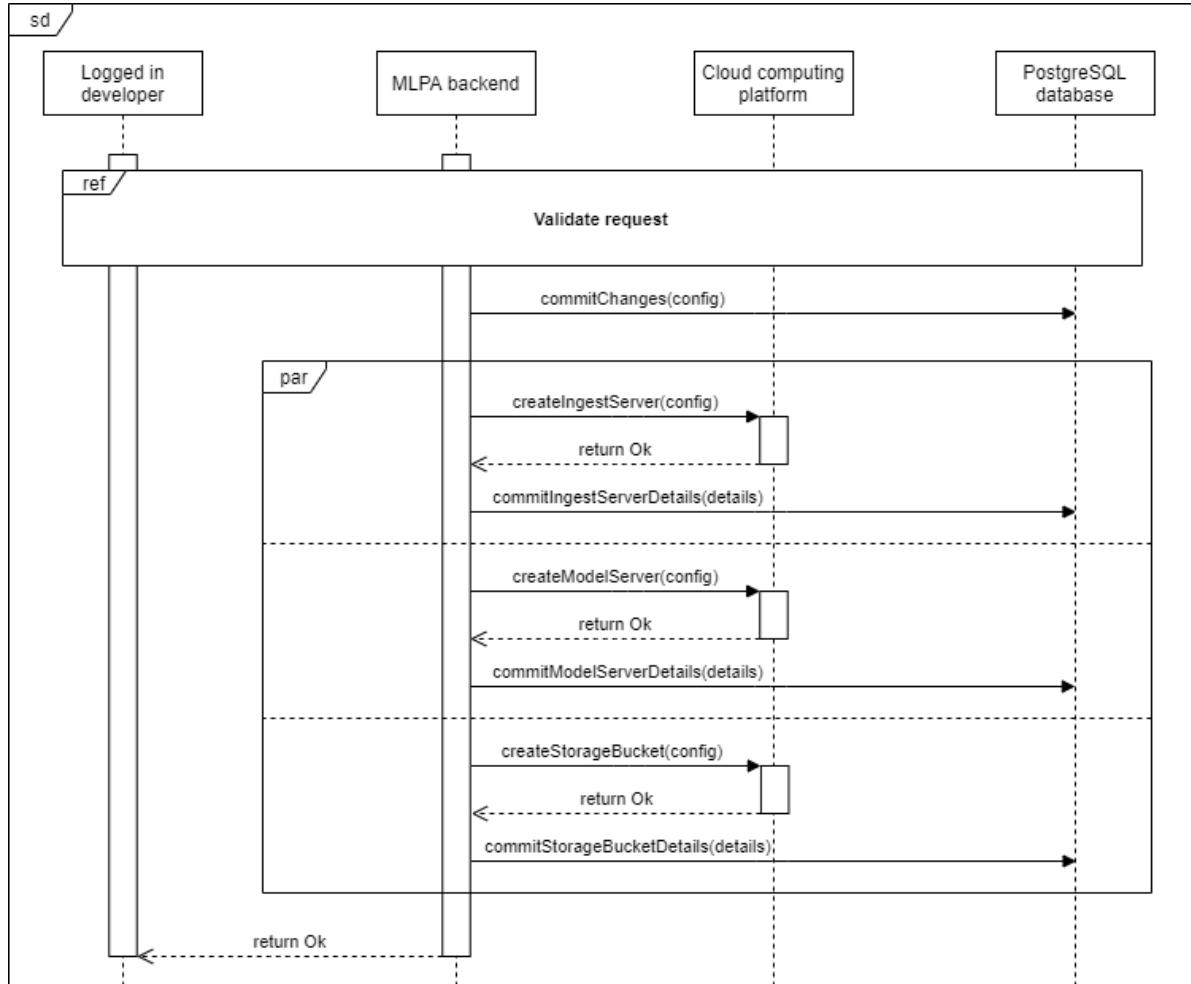
Diagrammen voor het laatste niveau, code, zijn achterwege gelaten omdat klassen niet worden gemaakt. In plaats daarvan worden functies gebruikt om een taak uit te voeren. Brown geeft ook als advies om de klassendiagrammen niet te maken of automatisch te laten genereren [50] omdat de diagrammen vaak tijdens het programmeerproces zullen veranderen. Als alternatief voor de klassendiagrammen is gekozen om sequence-diagrammen te maken om een beeld te geven wat er gebeurt op codeniveau.

6.5 Sequence-diagrammen

sequence-diagrammen zijn toegevoegd om het laatste niveau uit te leggen voor belangrijke acties zoals het aanmaken of uitvoeren van een pipeline. In deze paragraaf wordt er door een aantal diagrammen gelopen.

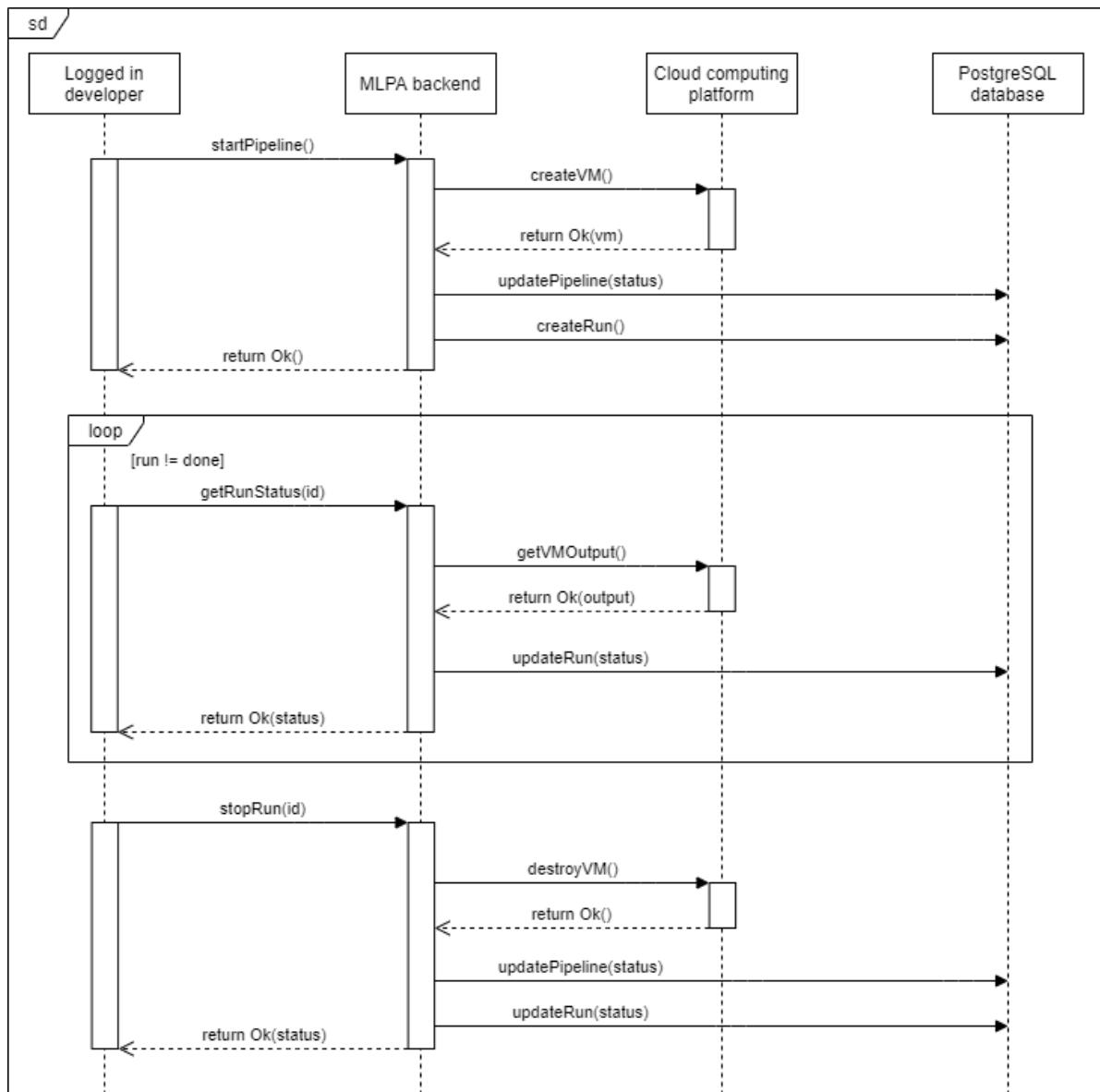
In Figuur 6.7 is te zien wat er gebeurt als een pipeline wordt aangemaakt. De developer stuurt een request naar de backend waar het wordt gevalideerd. Het validatie sequence-diagram is te vinden in bijlage X. De backend slaat vervolgens de gegevens op waarna, in parallel, een ingest server, model server en storage bucket worden aangemaakt. De ingest server zorgt voor de intake

van datasets, de model server traint het model en in de storage bucket worden alle datasets en modellen opgeslagen. Als deze drie resources zijn aangemaakt, wordt de request voltooid.



Figuur 6.7: Sequence-diagram van het aanmaken van een pipeline

Een sequence-diagram is gemaakt om te laten zien wat er gebeurt als een pipeline wordt gestart (Figuur 6.8). De developer zal een request maken om de pipeline te starten. Om dit te doen stuurt de backend een request naar het cloud platform om een virtual machine (VM) te starten. Na het opstarten wordt het trainen van het model automatisch gestart. De backend past de status van de pipeline aan en voltooit de request. De developer weet nu dat een VM is gestart en kan vragen naar de output van de VM. In de output is te zien waar de VM mee bezig is. Dit gebeurt continu totdat het model getraind is en de artefacten veilig zijn opgeslagen. Dit zijn afbeeldingen, modellen en datasets die bewaard moeten worden. Hierna zal de developer een laatste request sturen om de VM te verwijderen en de status in de database aan te passen.



Figuur 6.8: Sequence-diagram van het starten van een pipeline

7 Proof of concept

Om de PoC daadwerkelijk te maken moet er eerst gedefinieerd worden wat de scope is en hoe de PoC eruit gaat zien. De scope brengt in beeld wat er gedaan moet worden en wat het minimal viable product (MVP) is. Op het MVP kan de mock up gebaseerd worden. In dit hoofdstuk wordt de scope bepaalt, delen van de mock up doorgelopen, uitleg gegeven over de PoC en kwaliteit van de code. Het hoofdstuk wordt afgesloten met een conclusie en advies

7.1 User requirements verzamelen

Om te begrijpen wat er gebouwd moet worden kunnen de behoeftes opgeschreven worden in de vorm van user stories. Een user story is de behoefte opgeschreven in een specifiek formaat om belangrijke informatie te achterhalen [51]. Een bekend formaat is de *Connextra template* van Mike Cohn [51]: **”As a [role], I can [capability], so that [receive benefit]”**. De template zorgt er voor dat de rol (role), functie (capability) en reden (benefit) bekend is. In Figuur 7.1 zijn alle user stories opgesteld, gegroepeerd in Epics. Bij elke user story is ook aangegeven of het MVP is, welk prioriteit het heeft en de T-shirt size. De T-shirt size is een manier om op een hoog niveau aan te geven hoeveel werk een user story vereist. Een **S** vereist weinig werk terwijl **XL** veel werk vereist.

ID	Epic	As a/an...	i want to...	so that...	MVP [Y/N]	P [L/M/H]	T-Shirt [S/M/L/XL]	Notes
1	Manage pipeline	Developer	create a pipeline with a name, project and platform	I can create a pipeline configuration and train ML models	Y	H	M	
2	Manage pipeline	Developer	see the status of a pipeline	I know if a pipeline works or not	Y	L	S	
3	Manage pipeline	Developer	switch between platforms	I am not vendor-locked in	N	H	L	
4	Ingest server	Developer	see the status of the ingest server	I know if the ingest server is available and works correctly	N		S	
5	Ingest server	Developer	upload data from an application to the ingest server with specification on how it's saved	I can use the data in a dataset when a new ML model is being trained	N		M	
6	Model deployment server	Developer	see the status of the model deployment server	I know if the model deployment server is available and works correctly	N		S	
7	Model deployment server	Developer	get a prediction from the model deployment server via an endpoint	I can use a ML model without downloading it and using it locally	N		M	
8	Datasets/artifacts storage bucket	Developer	upload datasets	I can use the dataset when training ML models	Y	H	M	
9	Datasets/artifacts storage bucket	Developer	download datasets	I can experiment locally	Y	L	S	
10	Datasets/artifacts storage bucket	Developer	download dataset artifacts	I can experiment and debug locally	Y	L	S	
11	Datasets/artifacts storage bucket	Developer	download model artifacts	I can use a ML model offline	Y	M	S	
12	Pipeline configuration	Developer	be warned with detailed explanation if the configuration is invalid	I can correct the errors	N		M	
13	Pipeline runs	Developer	see the steps that have been performed in a run	I know what the MLPA did	N	H	XL	
14	Pipeline runs	Developer	see the output a step might have generated	I can see the analysis of datasets and models in a run	N	H	L	
15	Pipeline runs	Developer	start a pipeline run	I can train a model	Y	H	L	
16	Platform management	Developer	save platform keys or pats	I can access resources in platforms when using pipelines	N		S	

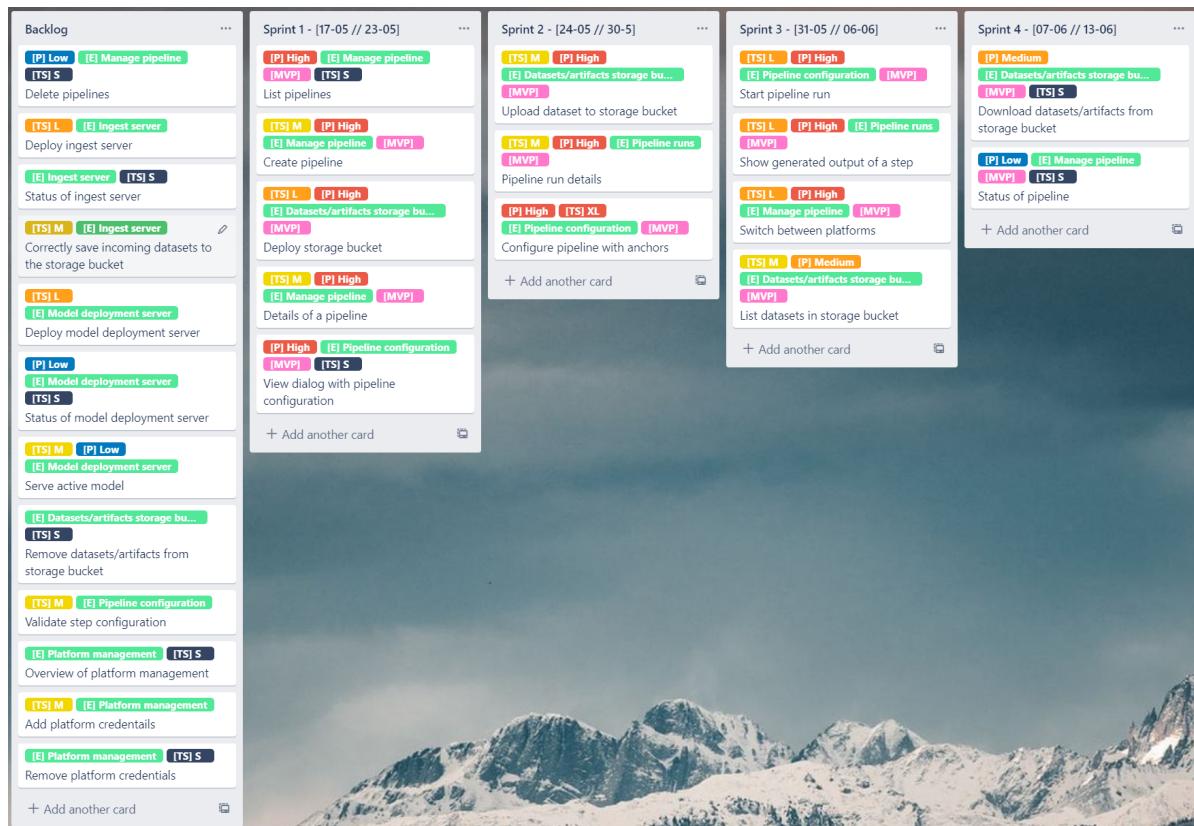
Figuur 7.1: Requirements opgesteld voor de proof of concept

De user stories zijn opgesteld in samenwerking met de begeleiders van NGTI over het hele afstudeerproces. Nu de requirements zijn opgesteld, kan een Kanban bord opgericht worden om sprints te maken.

7.1.1 Requirements vertalen naar een scrum bord

Een Kanban bord is een werkwijze om visueel aan te geven in welke stadia een taak is in het proces [52]. Taken in een Kanban bord vallen altijd onder een kolom om de status aan te geven. Een kolom kan bijvoorbeeld "todo", "doing", of "done" zijn. Een scrum bord is een vorm van een Kanban bord waarbij taken in een sprint zijn ingepland [53]. Een sprint is een vaste periode van bijvoorbeeld een of twee weken waarin gewerkt wordt aan de ingeplande taken. Dit verschilt met een Kanban bord waarbij taken op elk moment toegevoegd kunnen worden. In software development kan dit niet; sprints zorgen dat er geen scope creep plaats vindt.

De user stories uit Figuur 7.1 zijn vertaald naar tickets in het scrum bord. Een user story is opgesplitst in meerdere tickets. Dit is het geval met bijvoorbeeld user story "upload datasets" (ID 8). Om deze functionaliteit te bouwen moet een storage bucket beschikbaar zijn en in de frontend de mogelijkheid zijn om dit te doen. In Figuur 7.2 zijn de tickets aangemaakt met relevante labels om aan te geven bij welke epic een ticket hoort, of het MVP is, de prioriteit en T-shirt size.



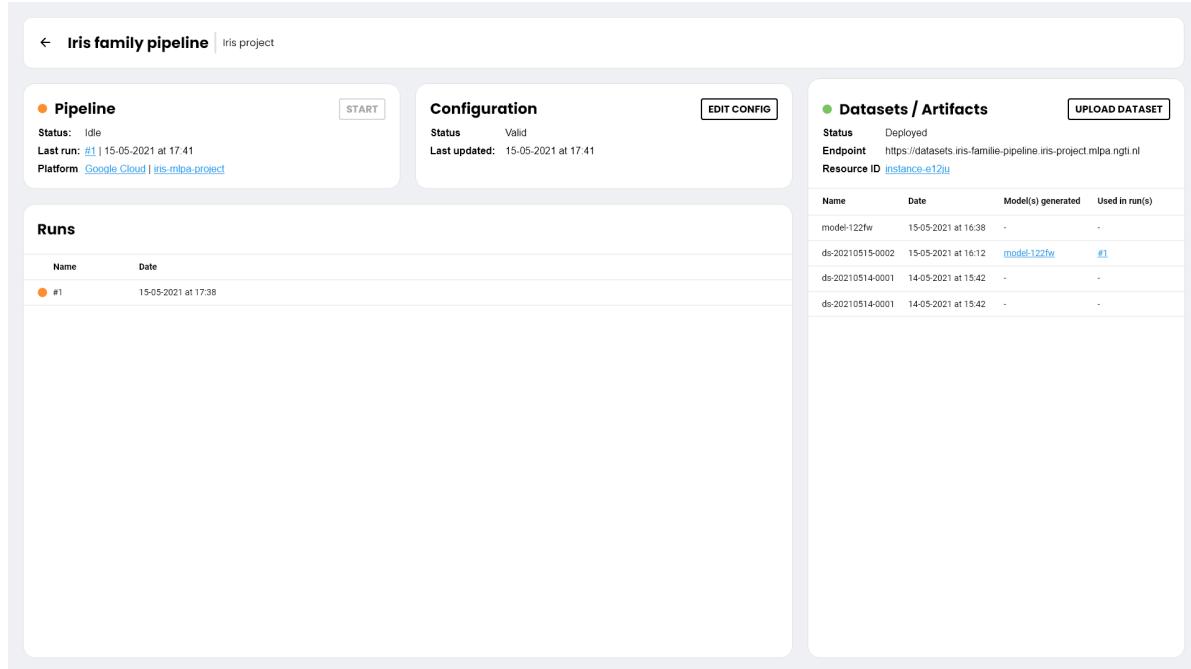
Figuur 7.2: Trello bord in het begin van sprint 1

In het ontwikkelproces is na elke sprint het scrum bord opnieuw bekijken om te reflecteren op de afgelopen sprint en de komende sprints opnieuw in te plannen als dit nodig was. Nu de tickets aangemaakt zijn en het MVP is vastgesteld, kan de mock up gemaakt worden.

7.2 Mock up van de proof of concept

De mock up is bedoeld om te laten zien hoe de frontend van de PoC eruit gaat zien. Dit geeft ook de kans om te controleren of de user interface (UI) logisch in elkaar zit. Om de mock up te ontwerpen is gebruik gemaakt van Adobe XD [54]. De mock up bevat alleen functionaliteiten die MVP zijn.

Een pipeline wordt aangemaakt door het invullen van de naam en project (bijlage VIII). Na het aanmaken van een pipeline kan erop geklikt worden. In Figuur 7.3 is de mock up van de details van een pipeline te zien. Op deze pagina is de status van de pipeline, configuration, dataset/artifacts en runs te zien.



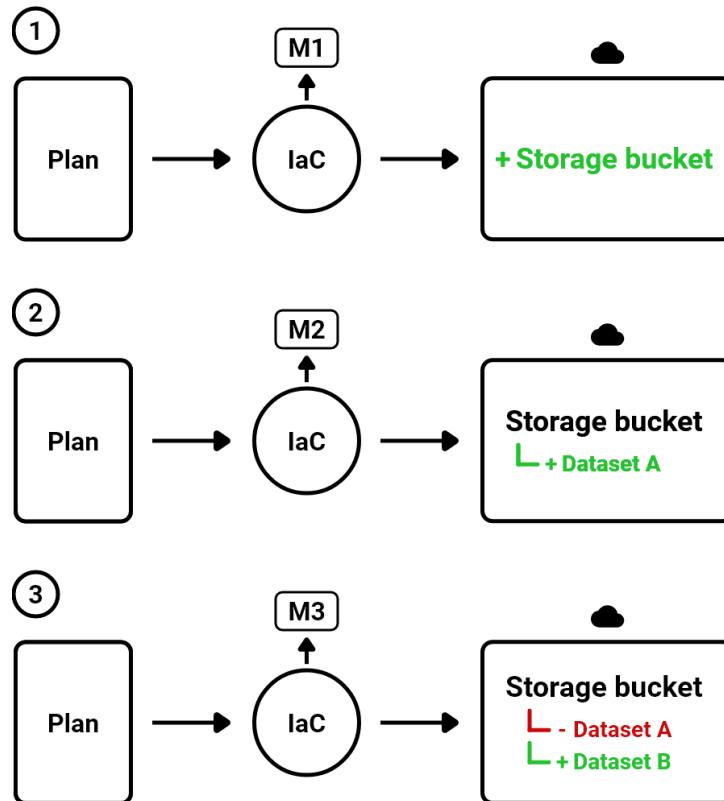
Figuur 7.3: Mock up van de pipeline details pagina

De configuration is een Python script dat door de developer is gedefinieerd. Het script wordt uitgevoerd als de pipeline wordt gestart en bevat de stappen uit Figuur 4.2. Stappen zoals "Model deployment" en "Model feedback" vallen uit de scope. In de datasets/artifacts sectie kan de developer datasets en artefacts up- en downloaden. Als laatste kan de developer de status van de pipeline en run zien. Een run is een individuele run als een pipeline wordt gestart. Mock ups van andere gedeeltes van de PoC is te vinden in bijlage IX. Nu het voorwerk is gedaan, kan code voor de PoC geschreven worden.

7.3 Wijzigingen in het architectuurontwerp

Tijdens het ontwikkelproces is naar voren gekomen dat het orkestreren van infrastructuur in een cloud platform met een tool niet de juiste werkwijze is. Een orkestratietool werkt zoals uitgelegd in paragraaf 5.1 met een plan. De tool controleert de infrastructuur in een cloud platform aan de hand van het plan en past zo nodig de infrastructuur aan.

Gedurende het experiment kwamen geen problemen naar boven; echter bij het implementeren van een feature vertoonde Pulumi gedrag dat onverwacht was. In Figuur 7.4 is een illustratie te zien ter ondersteuning van de uitleg. In de eerste stap (1) wordt middels code een plan beschreven waarbij een storage bucket aangemaakt moet worden. Pulumi valideert het plan en maakt vervolgens de storage bucket aan. Pulumi houdt intern bij met een manifest (M1) wat is aangemaakt, in dit geval een storage bucket. Bij de tweede stap (2) wordt Dataset A geüpload in de storage bucket. Pulumi houdt weer bij wat er aangemaakt is in de storage bucket met het manifest (M2). Bij de derde stap staat in het plan dat alleen Dataset B geüpload moet worden. Pulumi vergelijkt het plan met het manifest (M2) en ziet dat Dataset A niet in het plan staat. Pulumi verwijderd vervolgens Dataset A en upload Dataset B.



Figuur 7.4: Gedrag van Pulumi met een incorrect plan

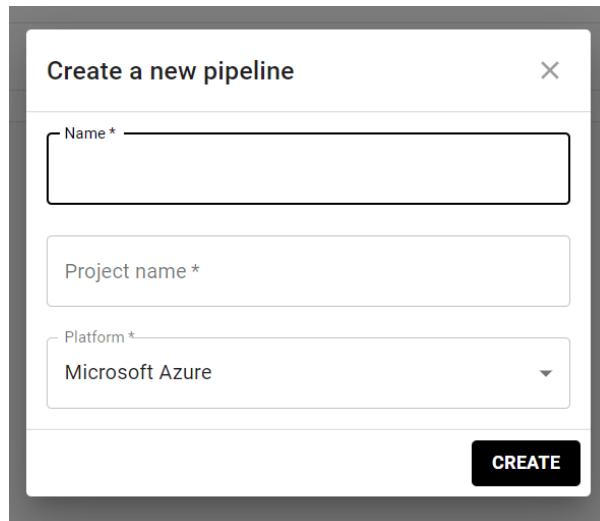
Pulumi verwacht dat de volledige infrastructuur in het plan is beschreven. Het is niet mogelijk om incrementeel de infrastructuur aan te passen. Hier was ook niet omheen te werken. Voor de PoC is daarom gekozen om verder te werken met libraries van de cloud platformen AWS, Google Cloud en Azure zelf. Na het onderzoeken of het mogelijk is om infrastructuur in te richten en ook gekeken is naar de kwaliteit van de documentatie is er voor gekozen om te werken met de libraries van Google Cloud en Azure. De documentatie van AWS is onoverzichtelijker en Amazon gebruikt unieke termen voor hun resources en services wat de leercurve hoger legt dan de andere twee libraries [55].

7.4 De proof of concept

In paragraaf 6.2 is kort uitgelegd dat de backend gebouwd wordt met Node.js en Express en de frontend met React.js en TypeScript. Voor de PoC is een GitHub repository (<https://github.com/UNRULYTHON/mlpa-poc>) aangemaakt waarin de front- en backend geïnitialiseerd is met behulp van de documentatie van de frameworks. In de volgende paragrafen wordt in een logische volgorde door de applicatie gelopen waarbij uitleg over de werking wordt gegeven met screenshots van de UI en code als ondersteuning.

7.4.1 Aanmaken van pipelines

Om te beginnen moet een pipeline aangemaakt kunnen worden. In de frontend gaat dit met een dialoog (Figuur 7.5). Hierin kan de naam, project en cloud platform aangegeven worden. De PoC zal zowel Google Cloud als Microsoft Azure ondersteunen. Voor het gemak wordt vanaf hier Google Cloud gebruikt.



Figuur 7.5: Aanmaak dialoog van een pipeline

Zodra de developer klikt op de **Create** knop, wordt een request verstuurt van de frontend naar de backend. De backend stuurt de request naar een service dat er voor zorgt dat de juiste acties uitgevoerd wordt. De service in Figuur 7.6 controleert welk platform er gespecificeerd is en spreekt de relevante functie aan. Als het platform bijvoorbeeld Google Cloud is, wordt de functie *gc_createBucket()* op regel 6 aangeroepen.

```

1 export const createPipeline = async (data: DTO_CreatePipeline) => {
2   try {
3
4     const pipeline = await createPipelineInDb(data)
5     if (data.platform === 'GOOGLE') {
6       await gc_createBucket({ name: data.name, project_id: data.project_id })
7     } else if (data.platform === 'AZURE') {
8       await azure_createResourceGroup({ name: data.name, project: data.project })
9       await azure_createBucket({ name: data.name, project: data.project })
10    }
11
12    return pipeline
13  } catch (e) {
14    throw new Error(e)
15  }
16}

```

Figuur 7.6: Service om een pipeline aan te maken

Om een pipeline in Google Cloud aan te maken hoeft alleen een storage bucket aangemaakt te worden. De functie die deze actie uitvoert is te zien in Figuur 7.7. De bucket heeft een naam nodig en de locatie. Op regel 5 van Figuur 7.7 is ”**EUROPE-WEST4**” gespecificeerd. Dit is een locatie in Nederland.

```

1 export const gc_createBucket = async (data: { name: string, project_id: string }) => {
2   const storage = new Storage()
3
4   return await storage.createBucket(createBucketName(data.name), {
5     location: 'EUROPE-WEST4'
6   })
7 }

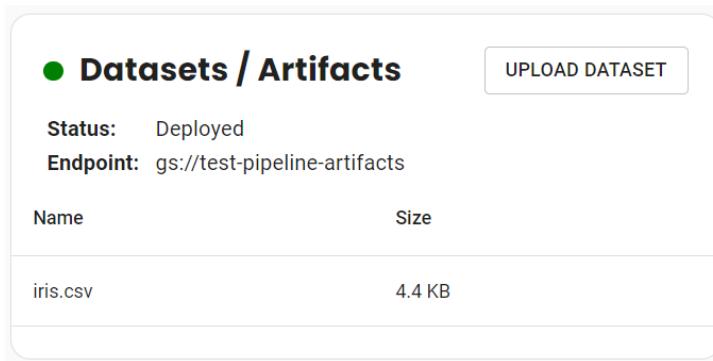
```

Figuur 7.7: Functie om een storage bucket aan te maken in Google Cloud

Naast het aanmaken van de storage bucket in Google Cloud wordt de details van de pipeline opgeslagen in de PostgreSQL database (Figuur 7.6, regel 4). Na het aanmaken van een pipeline kan een dataset geupload worden.

7.4.2 Datasets uploaden

Tijdens het aanmaken van de pipeline is een opslaglocatie in het cloud platform aangemaakt. In de details-weergave van een pipeline is een sectie weergegeven zoals in Figuur 7.8 waar de developer datasets kan uploaden. Na het uploaden krijgt de developer een confirmatie te zien en wordt de lijst met bestaande datasets en artifacts ververst.



Figuur 7.8: Status en upload-mogelijkheid van datasets en artefacts in de details-pagina van een pipeline

De code in de backend voor het uploaden is vergelijkbaar met de code in Figuur 7.6 en Figuur 7.7. In Figuur 7.9 is de service te zien dat wordt gebruikt als de developer een dataset uploadt in de frontend. De service selecteert de juiste functie om de dataset te uploaden om basis van het gekozen platform.

```

1 export const uploadData = async (id: number, files: fileType[]) => {
2   try {
3     const p = await pipeline(id)
4
5     if (p.platform === 'GOOGLE') {
6       await gc_uploadToBucket({ name: p.name, project_id: p.project_id }, files)
7     } else if (p.platform === 'AZURE') {
8       await azure_uploadToBucket({ name: p.name, project: p.project }, files)
9     }
10
11    return {}
12  } catch (e) {
13    throw new Error(e)
14  }
15}

```

Figuur 7.9: Service om datasets te uploaden naar een opslaglocatie binnen een cloud platform

```

1 export const gc_uploadToBucket = async (data: { name: string, project_id: string }, files: fileType[]) => {
2   const gc_bucket = gc_storage.bucket(createBucketName(data.name))
3
4   console.log(`Uploading to bucket: ${gc_bucket.name}`)
5
6   files.map(file => {
7     gc_bucket.upload(file.path, { destination: file.name, public: true })
8       .then(r => console.log('done uploading'))
9       .catch(e => console.log(e))
10  })
11
12  return {}
13}

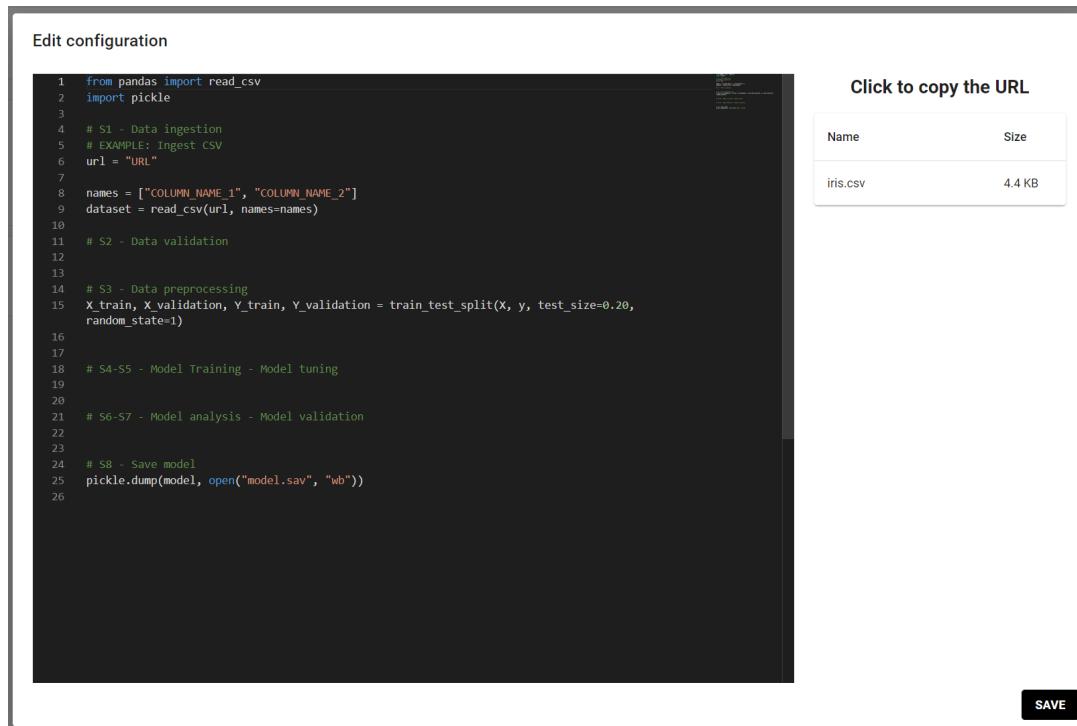
```

Figuur 7.10: Functie om datasets te uploaden naar een storage bucket in Google Cloud

De backend maakt een instantie van de bucket in Figuur 7.10 op regel 2 en voert de *upload()* functie uit op regel 7. De geüploade datasets zijn vervolgens beschikbaar om te gebruiken in de configuratie van de pipeline.

7.4.3 Configuratie definiëren

Het laatste wat de developer moet doen om de pipeline te starten is het opgeven van de configuratie. Zoals aangegeven in paragraaf 7.2 is de configuratie een Python bestand dat uitgevoerd wordt als de pipeline gestart wordt. In de PoC is de configuratie bewerkbaar door middel van een dialoog (Figuur 7.11) met rechts een lijst van de datasets. De standaard configuratie dient als startpunt voor de developer. Een link naar de dataset kan gekopieerd worden door op een dataset te klikken.



Figuur 7.11: Configuratiedialoog van een pipeline

Zodra de developer op de knop **Save** klikt, wordt een request gedaan naar de backend om de configuratie op te slaan. De code om de configuratie op te slaan is te zien in Figuur 7.12. Hier wordt een nieuwe configuratie aangemaakt en gelinkt aan de pipeline.

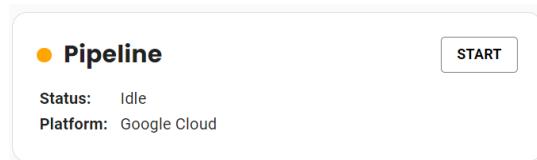
```
1 export const createPipelineConfiguration = async (pipeline_id: number, config: string) => {
2   return await prisma.pipelineConfiguration.create({
3     data: {
4       config: config,
5       pipeline: {
6         connect: {
7           id: pipeline_id
8         }
9       }
10     }
11   })
12 }
```

Figuur 7.12: Code om de configuratie op te slaan in de database

Het is mogelijk om code te schrijven in andere talen dan Python om modellen te trainen. Om de complexiteit en scope klein en realistisch te houden is er gekozen voor Python. De taal is een van de populairste talen [30]. Daarnaast heeft Python een complete set van libraries om data te transformeren, plotten van diagrammen en modellen te trainen [56]. Nu de configuratie is opgeslagen zijn alle benodigdheden beschikbaar om de pipeline te starten.

7.4.4 Pipeline starten

Zodra de opslaglocatie en configuratie beschikbaar zijn, krijgt de developer de mogelijkheid om de pipeline te starten Figuur 7.13. Zoals uitgelegd in Figuur 6.8 worden drie stappen uitgevoerd: het aanmaken van een VM, het continu ophalen van de output en vervolgens het verwijderen van de VM.



Figuur 7.13: Status en startknop van een pipeline

De request van de frontend komt in de backend binnen om een pipeline te starten en komt terecht bij de functie `gc_startPipeline()` in Figuur 7.14. Deze functie maakt een configuratie aan op regel 7 waarin het besturingssysteem, soort VM in Google Cloud en het startup script staat gedefinieerd. De configuratie wordt gebruikt als de VM wordt aangemaakt op regel 27. De status van de pipeline wordt aangepast naar **"RUNNING"** en een nieuwe Run wordt aangemaakt in de database.

```
1 export const gc_startPipeline = async (pipeline_id: number, name: string, run: number) => {
2   const compute = new Compute()
3   const zone = compute.zone('europe-west4-a')
4   const pipeline = await fetchPipeline(pipeline_id)
5   const configuration = await currentPipelineConfiguration(pipeline_id)
6
7   const config = {
8     os: 'ubuntu',
9     http: true,
10    machineType: 'e2-small',
11    metadata: {
12      items: [
13        {
14          key: 'startup-script',
15          value: `...`,
16        }
17      ]
18    }
19  }
20
21  const vm = zone.vm(`${name}-${run}`)
22
23  await updatePipeline(pipeline.id, { ...pipeline, status: 'RUNNING', run: pipeline.run + 1 })
24  const _run = await newRun(pipeline_id, pipeline.platform, `${name}-${run}`, 'PROVISIONING')
25
26  console.log(`Creating VM ${name}-${run} ...`);
27  await vm.create(config)
28
29  return { run_id: _run.id }
30 }
```

Figuur 7.14: Code om een virtual machine (VM) in Google Cloud te starten

In de configuratie is een startup script gedefinieerd (Figuur 7.14, regel 15). Het script wordt uitgevoerd zodra de VM beschikbaar is en zorgt er voor dat het model getraind kan worden en artifacts opgeslagen worden in de opslaglocatie indien dit nodig is. Op regel 18 in Figuur 7.15 wordt de configuratie die de developer heeft opgeslagen in deelparagraaf 7.4.3 in het script gezet. Na het trainen van het model worden alle artifacts in de *output* map geüpload naar de opslaglocatie. Dit gebeurt met bash commando's op regel 34 en 35.

```
1 #! /bin/bash
2 apt-get update
3 apt-get install -y python3-pip
4 echo '[MLPA] - INSTALLING PIP PACKAGES'
5
6 sudo pip3 install pandas
7 sudo pip3 install numpy
8 sudo pip3 install matplotlib
9 sudo pip3 install scikit-learn
10
11 sudo mkdir ml
12
13 cd ml
14 sudo mkdir output
15 echo '[MLPA] - CREATED OUTPUT FOLDER'
16
17 sudo cat <<'EOF' >> main.py
18 ${configuration.config}
19 EOF
20 echo '[MLPA] - CREATED PYTHON FILE'
21
22 sudo cat <<'EOF' >> key.json
23 ${JSON.stringify(key, null, 2)}
24 EOF
25 echo '[MLPA] - CREATED KEY FILE'
26
27 sudo python3 main.py
28
29 gcloud auth activate-service-account --key-file key.json
30
31 cd output
32
33 echo '[MLPA] - UPLOADING TO STORAGE BUCKET ... '
34 for FILE in *; do sudo mv $FILE "$HOSTNAME-$FILE"; done
35 for FILE in *; do gsutil cp $FILE gs://test-pipeline-artifacts; done
36
37 echo '[MLPA] - DONE'
```

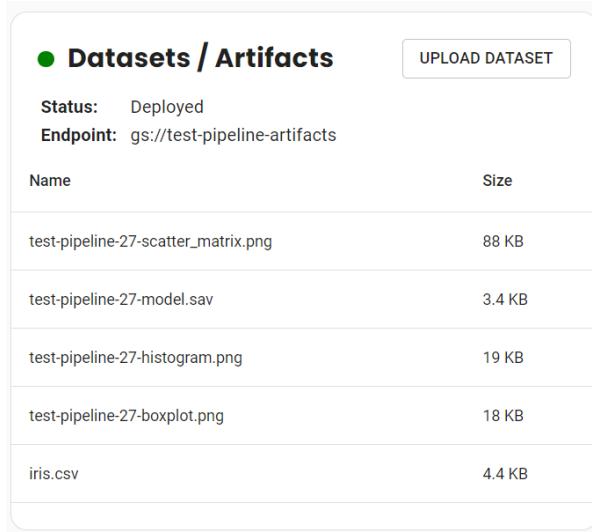
Figuur 7.15: Startup script van een virtual machine (VM) in Google Cloud

Gedurende het uitvoeren van het script in Figuur 7.14 vraagt de frontend continu naar de output van de VM. Deze request wordt afgehandeld door de *gc_getRunStatus()* functie in bijlage XI. De functie haalt de output op (regel 9) en slaat het op in de database voordat het terug gestuurd wordt naar de frontend. De frontend laat de output zien (bijlage XII) en controleert of de output '[MLPA] - DONE' bevat. Als de output deze tekst bevat, weet het systeem dat het model is getraind en alle artifacts in de output map geüpload zijn in de opslaglocatie. Als dit niet het geval is, wordt de output opnieuw opgehaald en geüpdatet in de frontend. Dezelfde controle op de output vindt vervolgens weer plaats. Als de output wel '[MLPA] - DONE' bevat, wordt er middels een request aan de backend de VM verwijderd (bijlage XIII). In het verwijderproces

wordt de status in de database geüpdatet.

7.4.5 Model downloaden

Nadat het model getraind is, worden artifacts in de *output* map geüpload naar de opslaglocatie. Het is aan de developer om in het Python script te definiëren dat artifacts opgeslagen worden in de *output* map. De lijst met datasets/artifacts wordt geüpdatet zodra het startup script klaar is. De artifacts zijn downloadbaar door erop te klikken.



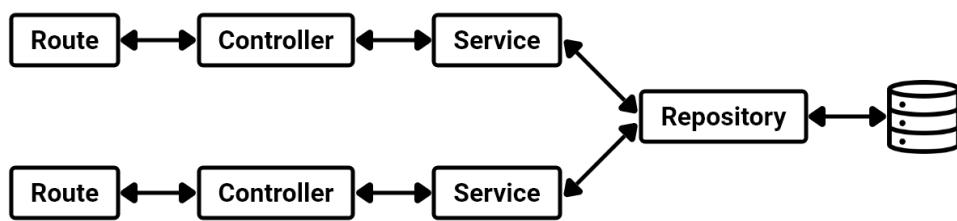
The screenshot shows a list of uploaded files in a storage bucket. At the top, there is a header with a green dot icon and the text "Datasets / Artifacts". To the right of the header is a button labeled "UPLOAD DATASET". Below the header, there are two lines of text: "Status: Deployed" and "Endpoint: gs://test-pipeline-artifacts". The main area contains a table with two columns: "Name" and "Size". The table lists six files:

Name	Size
test-pipeline-27-scatter_matrix.png	88 KB
test-pipeline-27-model.sav	3.4 KB
test-pipeline-27-histogram.png	19 KB
test-pipeline-27-boxplot.png	18 KB
iris.csv	4.4 KB

Figuur 7.16: Model en overige artifacts van het trainproces geüpload in een storage bucket

7.5 Kwaliteit van de code

Gedurende het ontwikkelproces is gebruik gemaakt van het repository-service pattern, verschillende principes en "good practices". In paragraaf 6.2 is het repository-service pattern kort uitgelegd. Het idee achter deze pattern is dat het ophalen van data uit de database en complexe logica gescheiden van elkaar blijft [49]. De repository laag in Figuur 7.17 bevat create, read, update en delete (CRUD) functies die in verschillende services hergebruikt kan worden.



Figuur 7.17: Voorbeeld van de repository-service pattern

Doordat de architectuur ontworpen is met het repository-service pattern, voldoet de PoC ook aan het principe don't repeat yourself (DRY) [57]. Volgens dit principe moet er geen code geschreven

worden dat al geschreven is. Functies zoals het ophalen van een pipeline kan hergebruikt worden met behulp van de repository laag (Figuur 7.18).

```
1 const getPipeline: RouteFunction = async (req, res, next) => {
2   const id = Number(req.params.id)
3
4   try {
5     const pipeline = await fetchPipeline(id)
6
7     if (pipeline) {
8       res.json(pipeline)
9     } else {
10       res.status(404).end()
11     }
12   } catch (e) {
13     return res.status(500).json(e)
14   }
15 }
16
17 const startPipeline: RouteFunction = async (req, res) => {
18   const id = Number(req.params.id)
19
20   try {
21     const p = await fetchPipeline(id)
22
23     const run_id = await runPipeline(p.id)
24
25     return res.status(200).json(run_id)
26   } catch (e) {
27     return res.status(500).json(e)
28   }
29 }
```

Figuur 7.18: Voorbeeld van de don't repeat yourself (DRY) principe in de backend

De toepassing van de DRY principe komt ook voor in de code van de frontend. Zoals weergegeven in Figuur 7.19 wordt bijvoorbeeld de container waar elke sectie in zit hergebruikt. Met deze werkwijze kunnen elementen in de frontend sneller en consistenter gebouwd worden.

The screenshot shows a web-based application interface for managing a pipeline. At the top, there's a header with a back arrow and the text '← test-pipeline | test-project'. Below the header, there are three main sections: 'Pipeline', 'Configuration', and 'Datasets / Artifacts'.

- Pipeline:** Shows status as 'Idle' and platform as 'Google Cloud'. There is a 'START' button.
- Configuration:** Shows status as 'Valid' and last updated as '11/06/2021, 15:55:00'. There is an 'EDIT CONFIG' button.
- Datasets / Artifacts:** Shows status as 'Deployed' and endpoint as 'gs://test-pipeline-artifacts'. There is a 'UPLOAD DATASET' button. A table lists the datasets and artifacts uploaded:

Name	Size
test-pipeline-27-scatter_matrix.png	88 KB
test-pipeline-27-model.sav	3.4 KB
test-pipeline-27-histogram.png	19 KB
test-pipeline-27-boxplot.png	18 KB
iris.csv	4.4 KB

Below these sections, there is a large empty area labeled 'Run'.

Figuur 7.19: Voorbeeld van het don't repeat yourself (DRY) principe in de frontend

Een tweede manier om de kwaliteit van de code te waarborgen is het gebruik van unittesten. Voor de front- en backend zijn aparte unittests geschreven. In Figuur 7.20 is een unittest suite te zien van de *Button* component in de frontend. Een unittest suite is een collectie van unittests. De naam van de suite is *Button* en de naam van de tests is aangegeven met de éérste parameter van de `it()` functie op regel 2 en 12. Voor de component zijn twee unittests geschreven. De éérste unittest valideert of de component door React gerendeerd kan worden met een specifiek tekst. De tweede unittest controleert of de component als *disabled* gerendeerd wordt als dit word gespecificeerd op regel 14.

```
1 describe('Button', () => {
2     it('renders with text', () => {
3         act(() => {
4             render(<Button>hello world</Button>, container)
5         })
6
7         const button = document.querySelector('button')
8
9         expect(button.textContent).toBe('hello world')
10    })
11
12    it('renders disabled', () => {
13        act(() => {
14            render(<Button disabled>hello world</Button>, container)
15        })
16
17        const button = document.querySelector('button')
18
19        expect(button.textContent).toBe('hello world')
20        expect(button).toBeDisabled()
21    })
22})
23
```

Figuur 7.20: Unittests van de *Button* component in de frontend

Een unittest suite van de backend is te zien in Figuur 7.21. Deze suite test de functies om een pipeline aan te maken, te updaten en te verwijderen in de database. Op regel 2 wordt een pipeline geïnitialiseerd met als variabelnaam *data*. Op regel 11 wordt een variabel aangemaakt om de pipeline buiten de tests te bewaren. Hierdoor kan elke test dezelfde data en aangemaakte pipeline gebruiken.

```

1 describe('DB - Pipeline', () => {
2   let data = {
3     id: 1,
4     status: PipelineStatus.IDLE,
5     name: 'test-pipeline',
6     project: 'test-project',
7     project_id: 'test-project-id',
8     platform: Platform.GOOGLE,
9     run: 0
10  }
11  let pipeline
12
13  it('should create a new pipeline', async () => {
14    prismaMock.pipeline.create.mockResolvedValue(data)
15
16    pipeline = await createPipelineInDb(data)
17    data = { ...data, id: pipeline.id }
18
19    await expect(pipeline).toEqual(data)
20  })
21
22  it('should update a pipeline', async () => {
23    prismaMock.pipeline.create.mockResolvedValue(data)
24
25    data = { ...data, name: 'updated-test-pipeline' }
26    pipeline = await updatePipeline(pipeline.id, data)
27
28    await expect(pipeline).toHaveProperty('name', 'updated-test-pipeline')
29  })
30
31  it('should remove a pipeline', async () => {
32    prismaMock.pipeline.create.mockResolvedValue(data)
33
34    pipeline = await removePipeline(pipeline.id)
35
36    await expect(pipeline).toEqual(data)
37  })
38})
39

```

Figuur 7.21: Unitests van de pipeline in de backend

7.6 Conclusie

In dit hoofdstuk is onderzoek gedaan naar het antwoord op de hoofdvraag: **H: In welke mate kan een machine learning pipeline worden geautomatiseerd onafhankelijk van de onderliggende cloud computing platform?** Het onderzoek is uitgevoerd met het verzamelen van requirements, ontwerpen van een mock up en een PoC bouwen.

Middels het implementeren is duidelijk geworden dat een orkestratietool niet de correcte werkwijze is om infrastructuur aan te passen. De tool verwacht dat alle resources, ook resources die voorheen aangemaakt zijn, in het plan zijn gedefinieerd. Resources worden verwijderd als dit het geval is. Om dit op te lossen is gekozen om te werken met libraries van de cloud platformen zelf. Documentatie van de libraries is aanwezig, echter laat documentatie van een aantal libraries te wensen over. De functionaliteit om resources aan te maken is wel aanwezig en is dus een valide vervanging.

De scope van de PoC bevat de stappen uit Figuur 4.2 om een model te trainen. Het trainen gaat met een Python bestand dat als configuratie wordt opgeslagen. Eerst moet een pipeline aangemaakt worden zodat de PoC de benodigdheden klaar kan zetten voor de developer. Dat is de opslaglocatie in de gekozen cloud platform en informatie over de pipeline in de database. Vervolgens kan de developer een dataset uploaden, de configuratie opslaan en de pipeline starten. Na het starten maakt de PoC een VM aan in het cloud platform en voert een script uit zodat het model getraind en opgeslagen kan worden samen met overige artifacts in de opslaglocatie. Tot slot wordt de VM verwijderd en is het model beschikbaar om te gebruiken in een webserver of app om voorspellingen of classificaties te maken.

Met de PoC kan een model getraind worden op een cloud platform naar keuze. Door een dataset en configuratie op te geven kan de pipeline gestart worden. De developer hoeft verder geen acties te ondernemen; de pipeline slaat alle artifacten op en verwijderd de aangemaakte resources zodra deze niet meer nodig zijn.

7.7 Advies

De PoC laat zien dat het automatiseren van een pipeline mogelijk is. Echter heeft een developer van NGTI vrij veel kennis nodig van ML om gebruik te maken van de PoC. De lat van vereiste kennis kan lager gezet worden door de manier hoe een developer configuratie opgeeft om het model te trainen te veranderen. De PoC vraagt aan de developer om Python code op te geven. Daarnaast wordt de mogelijkheid aangeboden om een link te kopiëren van een dataset om te gebruiken in een van de stappen. Het advies is om dit aanzienlijk te verbeteren door hier stappen van te maken zoals in Figuur 4.2. In Figuur 7.22 is een mock up van een verbeterd configuratiedialoog te zien. In plaats van de Python code op te geven, loopt de developer door vijf stappen. De stappen zijn links in de mock up te zien. De eerste stap vraagt naar het type ML probleem. De PoC converteert vervolgens de gekozen opties in code waarmee een model getraind kan worden.

Pipeline configuration

1 - Choose algorithm kind

2 - Dataset structure

3 - Pre-process tasks

4 - Algorithm exploration

5 - Model validation

What type of ML problem is it?

Classification
Pick one of N labels

Regression
Predict numeric values

Clustering
Group similar examples

Association rule learning
Infer likely association patterns in data

Structured output
Create complex output

Ranking
Identify position on a scale or status

SAVE CHANGES

NEXT

X

Figuur 7.22: Mock up van een verbeterde configuratiedialoog - Stap 1

Vervolgens kan de structuur van de dataset aangegeven worden in de volgende stap (Figuur 7.23). De developer ziet de eerste drie regels van het dataset en kan kolomnamen opgeven. Onder de definitie is de validatie van de dataset te zien. Bij deze stap kan de developer al controleren of de dataset geen onverwachte waardes bevat.

Pipeline configuration

1 - Choose algorithm kind	Define dataset structure				
2 - Dataset structure					
3 - Pre-process tasks					
4 - Algorithm exploration					
5 - Model validation					

iris.csv

UPLOAD TO INFER

sepal-length	sepal-width	petal-length	column_4	column_5	
5.1	3.5	1.4	0.2	Iris-setosa	
4.9	3.0	1.4	0.2	Iris-setosa	
4.7	3.2	1.3	0.2	Iris-setosa	

SAVE CHANGES

Dataset validation

```
(150, 5)
   sepal-length  sepal-width  petal-length  petal-width  class
0      5.1        3.5        1.4        0.2  Iris-setosa
1      4.9        3.0        1.4        0.2  Iris-setosa
2      4.7        3.2        1.3        0.2  Iris-setosa
3      4.6        3.1        1.5        0.2  Iris-setosa
4      5.0        3.6        1.4        0.2  Iris-setosa
   sepal-length  sepal-width  petal-length  petal-width
count  150.000000  150.000000  150.000000  150.000000
mean   5.843333  3.054000  3.758667  1.198667
std    0.828066  0.433594  1.764420  0.763161
min    4.300000  2.000000  1.000000  0.100000
25%   5.100000  2.800000  1.600000  0.300000
50%   5.800000  3.000000  4.350000  1.300000
75%   6.400000  3.300000  5.100000  1.800000
max    7.900000  4.400000  6.900000  2.500000
class
Iris-setosa      50
```

Figuur 7.23: Mock up van een verbeterde configuratiedialoog - Stap 2

Nu het systeem meer weet over de structuur van de dataset kan bij de volgende stap taken gemaakt worden om de dataset voor te bereiden (Figuur 7.24). Elke taak wordt in volgorde uitgevoerd waarbij de developer definieert wat er gedaan moet worden bij elk kolumn. In de mock up worden de getallen afgerond op twee decimalen.

Pipeline configuration

1 - Choose algorithm kind	Pre-process tasks				
2 - Dataset structure					
3 - Pre-process tasks					
4 - Algorithm exploration					
5 - Model validation					

SAVE CHANGES

Tasks

Order	Name
1	convert_strings_to_lower_case
2	round_numbers

+ Add task

Task definition

Name: round_numbers

Column	Type	Action
sepal-length	double	Round to two digits
sepal-width	double	Round to two digits
petal-length	double	Round to two digits
petal-width	double	Round to two digits
class	string	Nothing

Figuur 7.24: Mock up van een verbeterde configuratiedialoog - Stap 3

De vierde stap is een stap die alleen uitgevoerd wordt als een pipeline voor de eerste keer draait (Figuur 7.25). Bij deze stap wordt onderzocht welke algoritme het beste resultaat oplevert. De developer krijgt een lijst met algoritmes als suggestie op basis van het opgegeven probleem in stap 1. Daarnaast kiest de developer welke hyperparameters gebruikt moeten worden om te testen welke algoritme en hyperparameter combinatie het beste presteert.

The dialog box is titled "Pipeline configuration". On the left, a vertical sidebar lists steps: 1 - Choose algorithm kind, 2 - Dataset structure, 3 - Pre-process tasks, 4 - Algorithm exploration (which is selected), and 5 - Model validation. A "SAVE CHANGES" button is at the bottom of this sidebar. The main area is divided into two sections: "Algorithms to include" and "Hyperparameters".

Algorithms to include:

Algorithms
Logistic Regression Algorithm
Support Vector Machine Algorithm
+ Add algorithm

Hyperparameters:

Algorithm: Logistic Regression Algorithm	
Hyperparameter	Value
solver	['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']
penalty	['none', 'l1', 'l2', 'elasticnet']
C	[100, 10, 1.0, 0.1, 0.01]

Figuur 7.25: Mock up van een verbeterde configuratiedialoog - Stap 4

Als laatste stap wordt door de developer opgegeven hoe goed het nieuwe model moet presteren zodat de pipeline naar de volgende stap kan gaan. (Figuur 7.26). Dit kan simpelweg door een meetpunt te definiëren waarbij wordt aangegeven wat de minimale waarde moet zijn. Het meetpunt kan ook gaan over de prestatie van voorgaande modellen; het nieuwe model moet net zo goed of zelfs beter presteren dan zijn voorganger.

Pipeline configuration

1 - Choose algorithm kind
2 - Dataset structure
3 - Pre-process tasks
4 - Algorithm exploration
5 - Model validation

Performance measurements

measurement	operator	value
accuracy	=	0.975

+ Add performance check

SAVE CHANGES

Figuur 7.26: Mock up van een verbeterde configuratiedialoog - Stap 5

Met de verbeterde configuratiedialoog heeft de developer hetzelfde resultaat bereikt als een Python bestand waarin acties staan om een model te trainen. De manier met de verbeterde configuratiedialoog is gebruiksvriendelijker doordat er gevraagd wordt wat de developer wilt doen. Het systeem zorgt er voor dat code gegenereerd wordt om de gedefinieerde actie uit te voeren.

8 Discussie

Gedurende de scriptie is literatuuronderzoek gedaan naar ML pipelines en orkestratietools. Daarnaast zijn een aantal experimenten uitgevoerd om de potentie van een ML pipeline en orkestratietool te valideren.

De resultaten van het experiment met ML pipelines laten zien dat het trainen van modellen op een gestructureerde en reproduceerbare wijze gedaan kan worden. In paragraaf 4.4 is uitgelegd hoe ML versimpeld kan worden voor de developer door extra stappen toe te voegen. Dit kan niet alleen gebruikt worden door NGTI maar ook andere bedrijven die geïnteresseerd zijn in het toepassen van ML. Een stap verder zou zijn dat een standaard wordt geïntroduceerd dat wordt toegepast door industrieleiders zoals Google, Microsoft en Azure. Momenteel hebben ze hun eigen variant wat betekent dat specifieke kennis voor één cloud platform nodig is. Verhuizen naar een ander cloud platform betekent essentieel dat de werking van een ander ML pipeline geleerd moet worden. Bij het onderzoek zou ook het versimpelen van ML onderzocht worden. Dit is helaas niet aan bod gekomen door tijdsgebrek. In deelparagraaf 4.4.2 is wel advies gegeven over het versimpelen. Aan de hand van het advies zou een experiment opgesteld worden om het advies te beproeven.

Uit het onderzoek bleek dat de orkestratietool Pulumi een geschikte keuze is om verder mee te gaan. Het implementeren van de tool sprak echter het onderzoek tegen. Dit komt doordat bij het onderzoek een experiment is uitgevoerd waarbij één taak wordt uitgevoerd. Zodra er meerdere taken achter elkaar gedaan moeten worden, vertoont de tool ongewenst gedrag. Met een uitgebreider experiment waarbij meerdere taken worden uitgevoerd zou dit probleem wellicht eerder naar boven zijn gekomen.

9 Reflectie

Het afstudeerproces is ook een groot leerproces voor mij geweest. De behandelde materie was niet alleen nieuw maar ook zeer complex. Daarnaast heb ik een aantal fouten gemaakt waardoor ik in tijdsnood kwam.

In de eerste helft van de afstudeerstage lag de focus te veel op ML. Ik deed ontzettend veel onderzoek naar ML omdat ik weinig tot niks wist van het onderwerp. Gedurende het onderzoek was ik theoretisch bezig en programmeerde of experimenteerde ik niet; er was geen balans tussen theoretisch en praktisch werk. In de initiële planning was de tijd gelijkmatig over elk deel- en hoofdvraag verdeeld (bijlage XIV). Naarmate er werd gewerkt aan de scriptie nam de theorie over ML langzaam de planning over en schoof de overige deel- en hoofdvraag verder op om meer ruimte te maken. Het onderzoek was gelukkig niet voor niets. De stappen in een pipeline worden genomen omdat deze nodig zijn om een model te trainen. Door de kennis die ik heb opgedaan werd dat snel duidelijk en begreep ik wat er gebeurde in een pipeline.

Na een stap terug te hebben gedaan werd het duidelijk dat ML pipelines de focus was, en niet ML zelf. Halverwege de afstudeerstage heb ik samen met mijn bedrijfsbegeleiders gezeten om de focus en scope te bepalen. Hierdoor werd het snel duidelijk waarnaar ik onderzoek moest doen en wat ik moest bouwen. Dit heeft enorm geholpen met mijn productiviteit. Door de duidelijke grens van wat wel en niet binnen de scope viel deed ik geen onderzoek naar onderwerpen die niet relevant waren. Dit is achteraf gezien een grote maar niet duidelijke valkuil in het ML domein. Er zijn talloze algoritmes met elk hun eigen theorie en er zijn ontzettend veel manieren om dingen te doen. Terugkijkend heb ik stukken geschreven die niet geschreven hoefde te worden en ook niet in deze scriptie zijn beland.

Een andere valkuil was de manier waarop ik aan de scriptie werkte. Doordat ik vooral theoretisch bezig was had ik moeite met de structuur van hoofdstukken en de samenhang tussen paragrafen. Ook halverwege heb ik besloten met zowel de school- als bedrijfsbegeleiders dat ik praktischer bezig zou zijn en wat minder tijd aan de scriptie zou besteden. Het praktisch bezig zijn hield voor mij in dat er meer en sneller experimenten werden gedaan. In de scriptie zette ik dan steekwoorden/zinnen om de kern van een alinea te definiëren. Via deze manier kon ik de structuur van een hoofdstuk goed bepalen en kon er weer theoretisch gewerkt worden.

Iets wat ik vanaf het begin had moeten doen was het maken van een Trello bord om de status van taken bij te houden. Nadat ik een Trello bord had ingericht voor de laatste vier weken om aan de PoC te werken, wist ik dat een Trello bord beter zou werken voor mij. Bij elke meeting (45 in totaal) heb ik genotuleerd, maar uiteindelijk heb ik niet veel met de notities gedaan (bijlage XV). Met een Trello bord zou ik suggesties en andere taken die in een meeting voorbij schieten kunnen organiseren. Ook dient het als een herinnering zodat ik taken niet vergeet te doen. Tijdens de laatste vier weken ben ik ook begonnen met het gebruik van de Pomodorotechniek. Dit is een werkwijze wat helpt om de focus bij het schrijven van de scriptie te houden. Met deze techniek maak je een lijst met taken. Daarna start je een timer en na 25 minuten neem je een pauze van 5 minuten. Dit doe je vier keer waarna je een pauze kan nemen van 15 minuten. Ikzelf heb pomofocus.io gebruikt. Met deze website kan je inschatten hoeveel tijd een taak kost. Vervolgens geeft pomofocus een verwachte eindtijd.

Voor een volgende scriptie zou ik een Trello bord inrichten voor taken en gebruik maken van de Pomodorotechniek om de focus bij het werk te houden. Notuleren tijdens meetings heeft niet veel zin en suggesties/taken kunnen gelijk in het Trello bord. Daarnaast is de focus en scope belangrijk voor de scriptie; zonder dit val ik in de diepte en verspil ik tijd.

Bibliografie

-
- [1] NGTI. „Breng je business in beweging en versnel je groei. Met de mobile expertise van NGTI.” (), adres: <https://www.ngti.nl/diensten/> (bezocht op 22-03-2021).
 - [2] N. B.V. „Bouw je business, bereik je doelen. Met de mobile oplossingen van NGTI.” (), adres: <https://www.ngti.nl/oplossingen/> (bezocht op 29-03-2021).
 - [3] NGTI, „Organogram.”
 - [4] Swisscom. „Other divisions and participations.” (), adres: <https://www.swisscom.ch/en/about/beteiligungen-swisscom-uebersicht.html> (bezocht op 29-03-2021).
 - [5] NGTI. „Verplaats jezelf klimaatneutraler met de Climate Challenge App.” (), adres: <https://www.ngti.nl/cases/swiss-climate-challenge/> (bezocht op 29-03-2021).
 - [6] Swisscom. (), adres: <https://www.swissclimatechallenge.ch> (bezocht op 29-03-2021).
 - [7] NGTI. „Beheer eenvoudig al je provider data en instellingen vanuit één app.” (), adres: <https://www.ngti.nl/cases/my-swisscom-app/> (bezocht op 29-03-2021).
 - [8] P. Covington, J. Adams en E. Sargin. „Deep Neural Networks for YouTube Recommendations.” (), adres: <https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45530.pdf> (bezocht op 20-06-2021).
 - [9] A. Chandrashekhar, F. Amat, J. Basilico en T. Jebara. „Artwork Personalization at Netflix.” (7 dec 2017), adres: <https://docs.microsoft.com/en-us/azure/devops/pipelines/artifacts/pipeline-artifacts?view=azure-devops&tabs=yaml> (bezocht op 20-06-2021).
 - [10] A. e. a. Kannan. „Smart Reply: Automated Response Suggestion for Email.” (17 aug 2016), adres: <https://storage.googleapis.com/pub-tools-public-publication-data/pdf/45189.pdf> (bezocht op 20-06-2021).
 - [11] L. Binders. „Methodologie in je scriptie duidelijk beschrijven.” (7 jun 2018), adres: <https://www.scribbr.nl/scriptie-structuur/methodologie-in-je-scriptie/> (bezocht op 25-03-2021).
 - [12] Scribbr. „Wat is kwalitatief en kwantitatief onderzoek?” (10 okt 2013), adres: <https://www.scribbr.nl/onderzoeksmethoden/kwalitatief-vs-kwantitatief-onderzoek/> (bezocht op 25-03-2021).
 - [13] E. Alpaydin, *Introduction to Machine Learning*, 4de ed., ISBN: 9780262043793. (bezocht op 2020).
 - [14] Y. Gavrilova. „Artificial Intelligence vs. Machine Learning vs. Deep Learning: Essentials.” (8 apr 2020), adres: <https://serokell.io/blog/ai-ml-dl-difference> (bezocht op 01-06-2021).
 - [15] J. Brownlee. „How to Think About Machine Learning.” (2 apr 2018), adres: <https://machinelearningmastery.com/think-machine-learning/> (bezocht op 16-03-2021).
 - [16] C. Nicholson. „A Beginner’s Guide to Neural Networks and Deep Learning.” (), adres: <https://wiki.pathmind.com/neural-network> (bezocht op 16-03-2021).
 - [17] H. Hapke en C. Nelson, *Building Machine Learning Pipelines*, 1ste ed., ISBN: 9781492053194. (bezocht op 2021).
 - [18] DVG. (), adres: <https://dvc.org> (bezocht op 06-04-2021).
 - [19] Pachyderm. (), adres: <https://www.pachyderm.com> (bezocht op 06-04-2021).
-

-
- [20] R. Tatman. „Data Cleaning Challenge: Scale and Normalize Data.” (30 mrt 2018), adres: <https://www.kaggle.com/rtatman/data-cleaning-challenge-scale-and-normalize-data> (bezocht op 27-05-2021).
 - [21] R. S. „About Feature Scaling and Normalization.” (11 jul 2014), adres: https://sebastianraschka.com/Articles/2014_about_feature_scaling.html (bezocht op 27-05-2021).
 - [22] C. Lui. „Data Transformation: Standardization vs Normalization.” (2020), adres: <https://www.kdnuggets.com/2020/04/data-transformation-standardization-normalization.html> (bezocht op 27-05-2021).
 - [23] J. Brownlee. „Difference Between Algorithm and Model in Machine Learning.” (29 apr 2020), adres: <https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning> (bezocht op 28-05-2021).
 - [24] ——, „What is the Difference Between a Parameter and a Hyperparameter?” (26 jul 2017), adres: <https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/> (bezocht op 12-03-2021).
 - [25] J. Kahn. „A.I. and tackling the risk of “digital redlining”.” (11 feb 2020), adres: <https://fortune.com/2020/02/11/a-i-fairness-eye-on-a-i/> (bezocht op 30-05-2021).
 - [26] M. Wittig en A. Wittig, *Amazon Web Services in Action*. Manning Press, 1ste ed., ISBN: 9781617292880. (bezocht op 2021).
 - [27] HashiCorp. „Build Infrastructure.” (), adres: <https://learn.hashicorp.com/tutorials/terraform/azure-build?in=terraform/azure-get-started> (bezocht op 01-06-2021).
 - [28] S. Overflow. „2020 Developer Survey.” (), adres: <https://insights.stackoverflow.com/survey/2020> (bezocht op 03-06-2021).
 - [29] ——, „2020 Developer Survey.” (), adres: <https://insights.stackoverflow.com/survey/2020#technology-other-frameworks-libraries-and-tools> (bezocht op 03-06-2021).
 - [30] ——, „2020 Developer Survey.” (), adres: <https://insights.stackoverflow.com/survey/2020#technology-most-loved-dreaded-and-wanted-other-frameworks-libraries-and-tools-loved3> (bezocht op 03-06-2021).
 - [31] Pulumi. (), adres: <https://www.pulumi.com/> (bezocht op 03-06-2021).
 - [32] Ansible. „A system administrator’s guide to getting started with Ansible - FAST!” (12 mrt 2018), adres: <https://www.redhat.com/en/blog/system-administrators-guide-getting-started-ansible-fast> (bezocht op 03-06-2021).
 - [33] ——, „Public Cloud Guides.” (4 jul 2021), adres: https://docs.ansible.com/ansible/latest/scenario_guides/cloud_guides.html (bezocht op 03-06-2021).
 - [34] ——, (), adres: <https://docs.ansible.com/> (bezocht op 03-06-2021).
 - [35] Pulumi. „Automation API.” (), adres: <https://www.pulumi.com/docs/guides/automation-api/> (bezocht op 03-06-2021).
 - [36] ——, „Cloud Providers.” (), adres: <https://www.pulumi.com/docs/intro/cloud-providers/> (bezocht op 03-06-2021).
 - [37] ——, „Documentation.” (), adres: <https://www.pulumi.com/docs/> (bezocht op 03-06-2021).

-
- [38] Terraform. „Terraform Language Documentation.” (), adres: <https://www.terraform.io/docs/language/index.html> (bezocht op 03-06-2021).
 - [39] ——, „Providers.” (), adres: <https://registry.terraform.io/browse/providers> (bezocht op 03-06-2021).
 - [40] ——, „Terraform Documentation.” (), adres: <https://www.terraform.io/docs/index.html> (bezocht op 03-06-2021).
 - [41] S. Brown. „The C4 model for visualising software architecture.” (), adres: <https://c4model.com/> (bezocht op 07-06-2021).
 - [42] S. Overflow. „2020 Developer Survey.” (), adres: <https://insights.stackoverflow.com/survey/2020#technology-web-frameworks> (bezocht op 07-06-2021).
 - [43] ——, „2020 Developer Survey.” (), adres: <https://insights.stackoverflow.com/survey/2020#technology-most-loved-dreaded-and-wanted-web-frameworks> (bezocht op 20-06-2021).
 - [44] Facebook. „Getting Started.” (), adres: <https://reactjs.org/docs/getting-started.html> (bezocht op 07-06-2021).
 - [45] Vue. „Introduction.” (), adres: <https://vuejs.org/v2/guide/> (bezocht op 20-06-2021).
 - [46] O. Foundation. „About documentation.” (), adres: <https://nodejs.org/en/docs/> (bezocht op 07-06-2021).
 - [47] Microsoft. „.NET documentation.” (), adres: <https://docs.microsoft.com/en-us/dotnet/> (bezocht op 20-06-2021).
 - [48] S. Overflow. „2020 Developer Survey.” (), adres: <https://insights.stackoverflow.com/survey/2020#technology-databases> (bezocht op 25-06-2021).
 - [49] M. Jones. „The Repository-Service Pattern with DI and ASP.NET 5.0.” (2019), adres: <https://exceptionnotfound.net/the-repository-service-pattern-with-dependency-injection-and-asp-net-core/> (bezocht op 12-06-2021).
 - [50] S. Brown. „Frequently asked questions.” (), adres: <https://c4model.com/#FAQ> (bezocht op 07-06-2021).
 - [51] A. Alliance. „User Story Template.” (), adres: <https://www.agilealliance.org/glossary/user-story-template/> (bezocht op 11-06-2021).
 - [52] M. Rehkopf. „What is a kanban board?” (), adres: <https://www.atlassian.com/agile/kanban/boards> (bezocht op 11-06-2021).
 - [53] I. Viter. „Getting Started With a Kanban Board: 6 Tips for Productive Work.” (), adres: <https://www.forecast.app/blog/kanban-board> (bezocht op 11-06-2021).
 - [54] Adobe. (), adres: <https://www.adobe.com/nl/products/xd.html> (bezocht op 11-06-2021).
 - [55] Amazon. (), adres: <https://docs.aws.amazon.com/AWSJavaScriptSDK/latest/index.html>.
 - [56] GeeksforGeeks. „Best Python libraries for Machine Learning.” (23 aug 2019), adres: <https://www.geeksforgeeks.org/best-python-libraries-for-machine-learning/> (bezocht op 11-06-2021).
-

-
- [57] R. C. Martin, *Clean Code*. Pearson Education, Inc., 2018, ISBN: 9780132350884. (bezocht op 12-06-2021).

Bijlagen

I Scope deelvraag 1

Tabel 1: Scope deelvraag 1

D1: Uit welke stappen bestaat een machine learning pipeline?	
Scope	<p>Binnen de scope:</p> <ul style="list-style-type: none">• In kaart brengen uit welke stappen een pipeline bestaat• Acties die in een stap worden uitgevoerd• Of het mogelijk is om stappen te versimpelen / abstracteren voor developers• Of het mogelijk is om stappen en acties te automatiseren <p>Buiten de scope:</p> <ul style="list-style-type: none">• Automatisering van stappen en acties• Een versimpeling van machine learning
Toelichting	Er wordt gekeken naar welke stappen er in een pipeline zitten. De theorie wordt vervolgens toegepast in een experiment. De nadruk ligt vooral op de mogelijkheid er is om stappen en acties te automatiseren en of machine learning versimpeld kan worden, niet dat er een uitwerking is.

II Scope deelvraag 2

Tabel 2: Scope deelvraag 2

D2: Hoe kan een orkestratietool verschillende cloud computing platformen beheren om een machine learning pipeline op te zetten?	
Scope	<p>Binnen de scope:</p> <ul style="list-style-type: none">• High-level uitleg over hoe een orkestratietool werkt• Inventarisatie met de "knock-out" methode• Criteria lijst voor orkestratietools• Opmerkelijke features die relevant zijn voor de PoC• Ervaring opdoen doormiddel van een pipeline te maken op twee cloud computing platformen <p>Buiten de scope:</p> <ul style="list-style-type: none">• Performance en snelheid
Toelichting	Frameworks dat cloud computing platformen beheert worden in kaart gebracht. Met knock-out criteria wordt de lijst verkort. Met een framework wordt een pipeline opgezet.

III Scope deelvraag 3

Tabel 3: Scope deelvraag 3

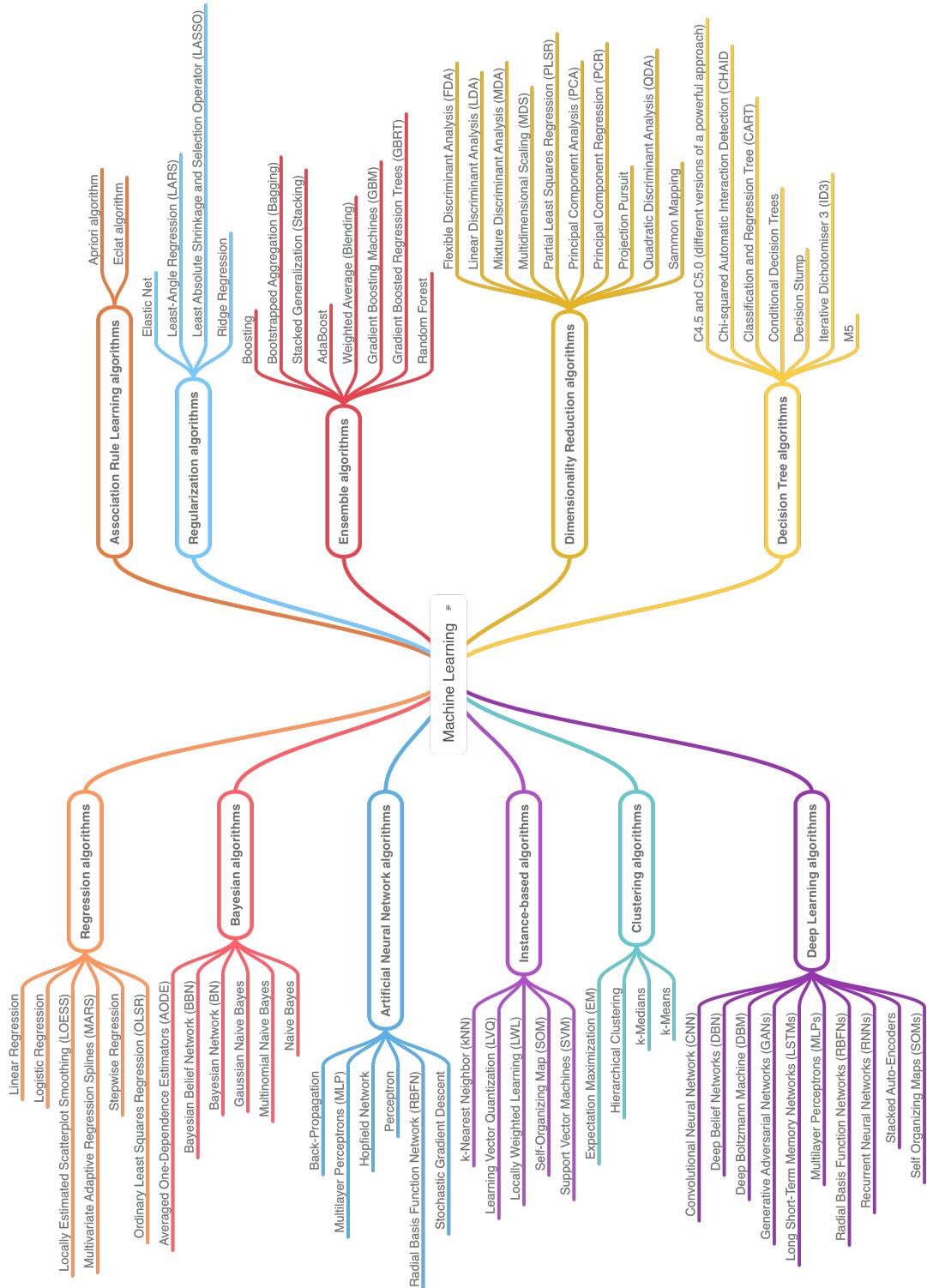
D3: Hoe ziet de architectuurblauwdruk van een applicatie, waarmee een platform-onafhankelijk machine learning pipeline opgezet kan worden, eruit?	
Scope	<p>Binnen de scope:</p> <ul style="list-style-type: none">• Technische tekeningen <p>Buiten de scope: -</p>
Toelichting	Het literatuuronderzoek slaat op of de technische tekeningen gemaakt zijn volgens een standaard zoals UML. Dit komt niet terug als theorie maar de bronnen worden wel vermeld.

IV Scope hoofdvraag

Tabel 4: Scope hoofdvraag

H: In welke mate kan een machine learning pipeline worden geautomatiseerd onafhankelijk van de onderliggende cloud computing platform?	
Scope	De scope wordt bepaald na de requirement analyse.
Toelichting	Onderzoek naar documentatie van gebruikte framework(s).

V Mindmap van machine learning algoritmes



Figuur 1: Mindmap van machine learning algoritmes

VI Machine learning pipeline experiment

Het experiment is uitgevoerd om te valideren of de stappen van de ML pipeline valide zijn en om de stappen beter te begrijpen. Het machine learning probleem dat wordt gebruikt om de ML pipeline te testen is het iris probleem. Bij dit probleem hoort een dataset dat bestaat uit 150 samples van drie iris soorten: Setosa, Virginica en Versicolor. Elk voorbeeld heeft de lengte en breedte van de kelk- en bloemblaadje met de bijbehorende soort iris. In Tabel 5 is een voorbeeld van de dataset te zien. Het model zal een van de drie soorten kunnen herkennen op basis van de lengte en breedte van een gegeven kelk- en bloemblaadje.

Kelkblaadje - lengte	Kelkblaadje - breedte	Bloemblaadje - lengte	Bloemblaadje - breedte	Soort
5.1	3.5	1.4	0.2	Iris-setosa

Tabel 5: Voorbeeld van de iris dataset

```
1 import pickle
2 import numpy as np
3 from pandas import read_csv
4 from pandas.plotting import scatter_matrix
5 from matplotlib import pyplot
6 from sklearn import svm, datasets
7 from sklearn.model_selection import train_test_split
8 from sklearn.metrics import classification_report
9 from sklearn.metrics import confusion_matrix
10 from sklearn.metrics import accuracy_score
11 from sklearn.svm import SVC
```

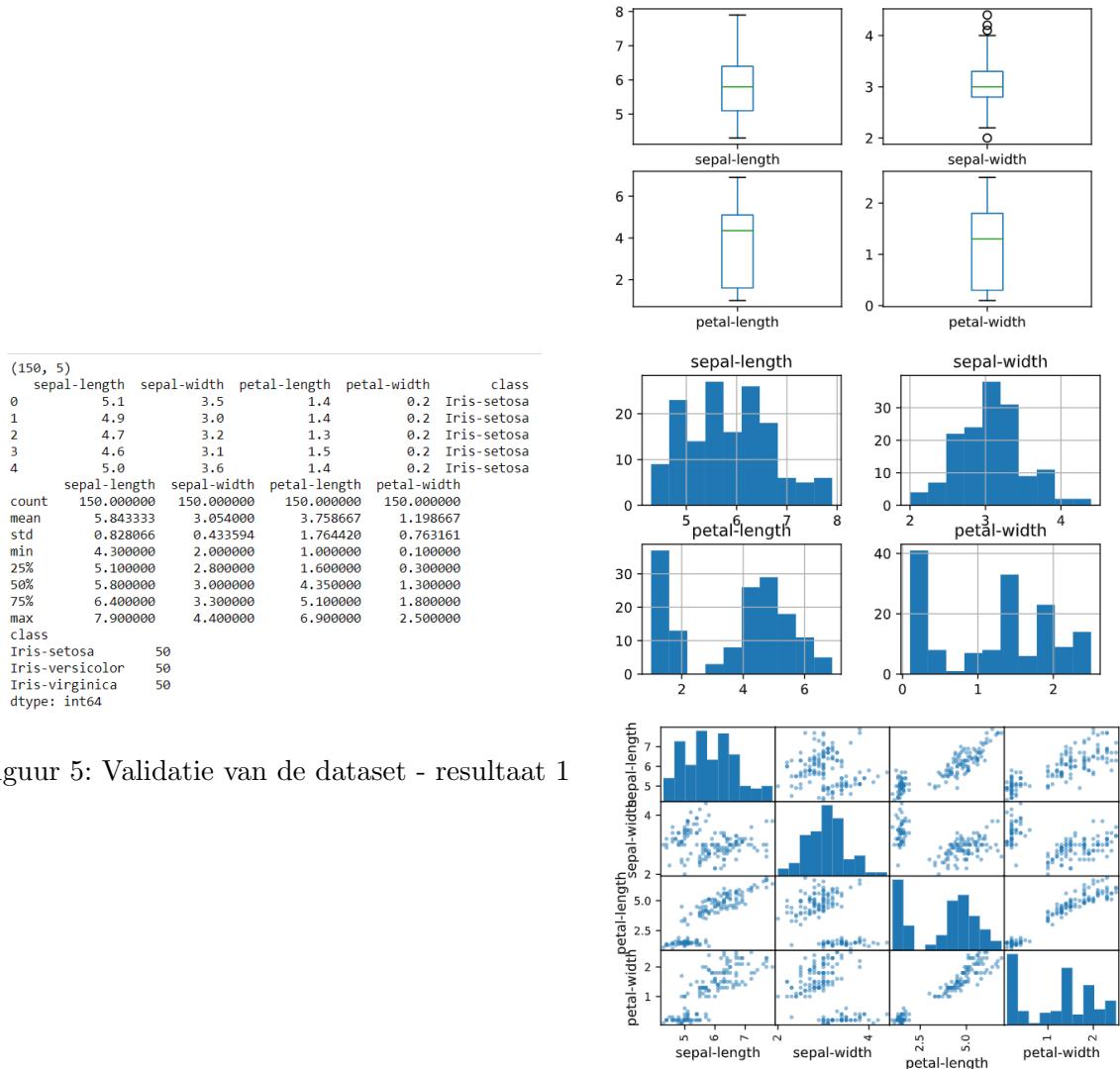
Figuur 2: Importeren van modules

```
1 # S1 - Data ingestion/versioning
2 url =
3     "https://gist.github.com/UNRULYEFON/7765e8bc37f928fc91aea3d10
4     9c337f5/raw/890f6e5c11f740150d57c578cf06d5bf9f35000d/iris.csv"
5
6 names = ["sepal-length", "sepal-width", "petal-length", "petal-width",
7           "class", ]
8 dataset = read_csv(url, names=names)
```

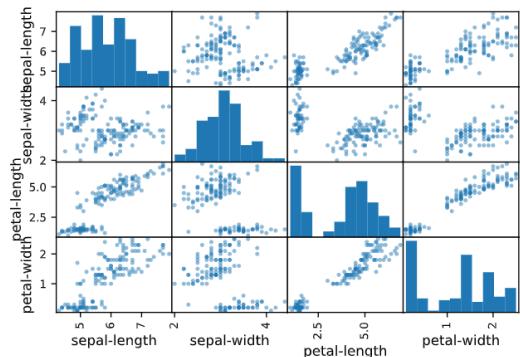
Figuur 3: Importeren van de dataset

```
1 # S2 - Data validation
2 # Statistical summary
3 print(dataset.shape)
4 print(dataset.head(5))
5 print(dataset.describe())
6 print(dataset.groupby('class').size())
7
8 # Visual summary
9 dataset.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)
10 pyplot.show()
11
12 dataset.hist()
13 pyplot.show()
14
15 scatter_matrix(dataset)
16 pyplot.show()
```

Figuur 4: Validatie van de dataset



Figuur 5: Validatie van de dataset - resultaat 1



Figuur 6: Validatie van de dataset - resultaat 2

```

1 # S3 - Data preprocessing
2 dataset['sepal-length'] = round(dataset['sepal-length'], 1)
3 dataset['sepal-width'] = round(dataset['sepal-width'], 1)
4 dataset['petal-length'] = round(dataset['petal-length'], 1)
5 dataset['petal-width'] = round(dataset['petal-width'], 1)
6
7 array = dataset.values
8 # Shallow copy (:) of the first four columns (0 → 3)
9 X = array[:,0:4]
10 # Shallow copy (:) of the last column (4)
11 y = array[:,4]
12 X_train, X_validation, Y_train, Y_validation = train_test_split(X, y,
   test_size=0.20, random_state=1)
```

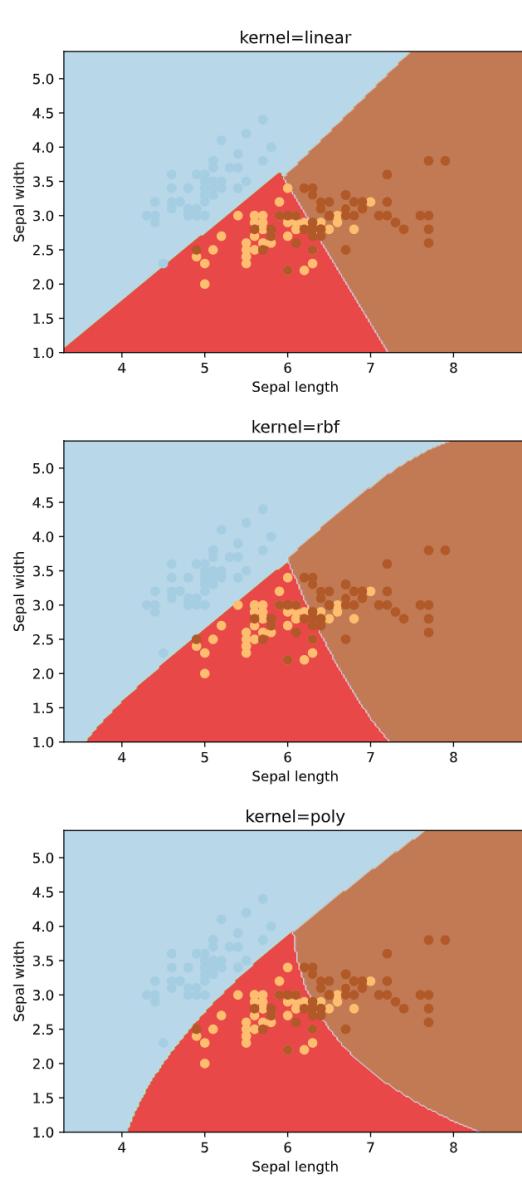
Figuur 7: Dataset voorbereiden

```

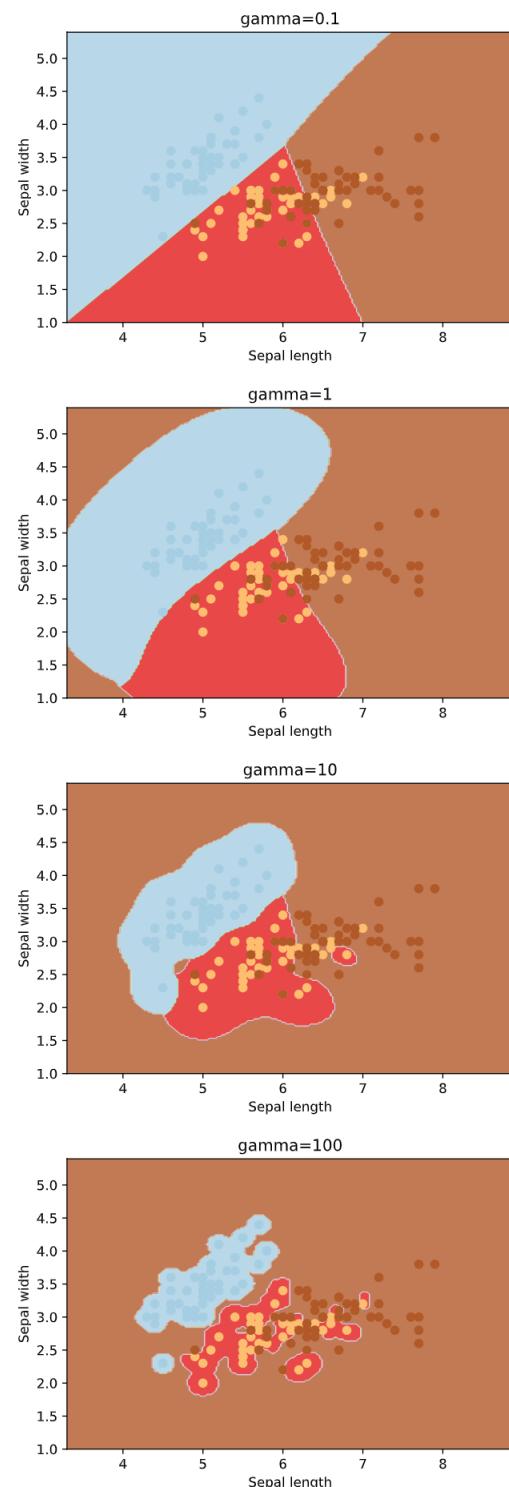
1 # S45 - Model Training - Model tuning
2 # Training
3 model = SVC(gamma='auto')
4 model.fit(X_train, Y_train)
5 predictions = model.predict(X_validation)
6
7 pickle.dump(model, open("model.sav", "wb"))
8
9 # Tuning
10 iris = datasets.load_iris()
11 X = iris.data[:, :2]
12 y = iris.target
13
14 def plotSVC(title):
15     x_min, x_max = X[:, 0].min() - 1, X[:, 0].max() + 1
16     y_min, y_max = X[:, 1].min() - 1, X[:, 1].max() + 1
17     h = (x_max / x_min)/100
18     xx, yy = np.meshgrid(np.arange(x_min, x_max, h),
19                           np.arange(y_min, y_max, h))
20     pyplot.subplot(1, 1, 1)
21     Z = svc.predict(np.c_[xx.ravel(), yy.ravel()])
22     Z = Z.reshape(xx.shape)
23     pyplot.contourf(xx, yy, Z, cmap=pyplot.cm.Paired, alpha=0.8)
24     pyplot.scatter(X[:, 0], X[:, 1], c=y, cmap=pyplot.cm.Paired)
25     pyplot.xlabel("Sepal length")
26     pyplot.ylabel("Sepal width")
27     pyplot.xlim(xx.min(), xx.max())
28     pyplot.title(title)
29     pyplot.show()
30
31 kernels = ["linear", "rbf", "poly"]
32 for kernel in kernels:
33     svc = svm.SVC(kernel=kernel).fit(X, y)
34     plotSVC("kernel=" + str(kernel))
35
36 gammas = [0.1, 1, 10, 100]
37 for gamma in gammas:
38     svc = svm.SVC(kernel="rbf", gamma=gamma).fit(X, y)
39     plotSVC("gamma=" + str(gamma))
40
41 cs = [0.1, 1, 10, 100, 1000]
42 for c in cs:
43     svc = svm.SVC(kernel="rbf", C=c).fit(X, y)
44     plotSVC("C=" + str(c))
45
46 degrees = [0, 1, 2, 3, 4, 5, 6]
47 for degree in degrees:
48     svc = svm.SVC(kernel="poly", degree=degree).fit(X, y)
49     plotSVC("degree=" + str(degree))

```

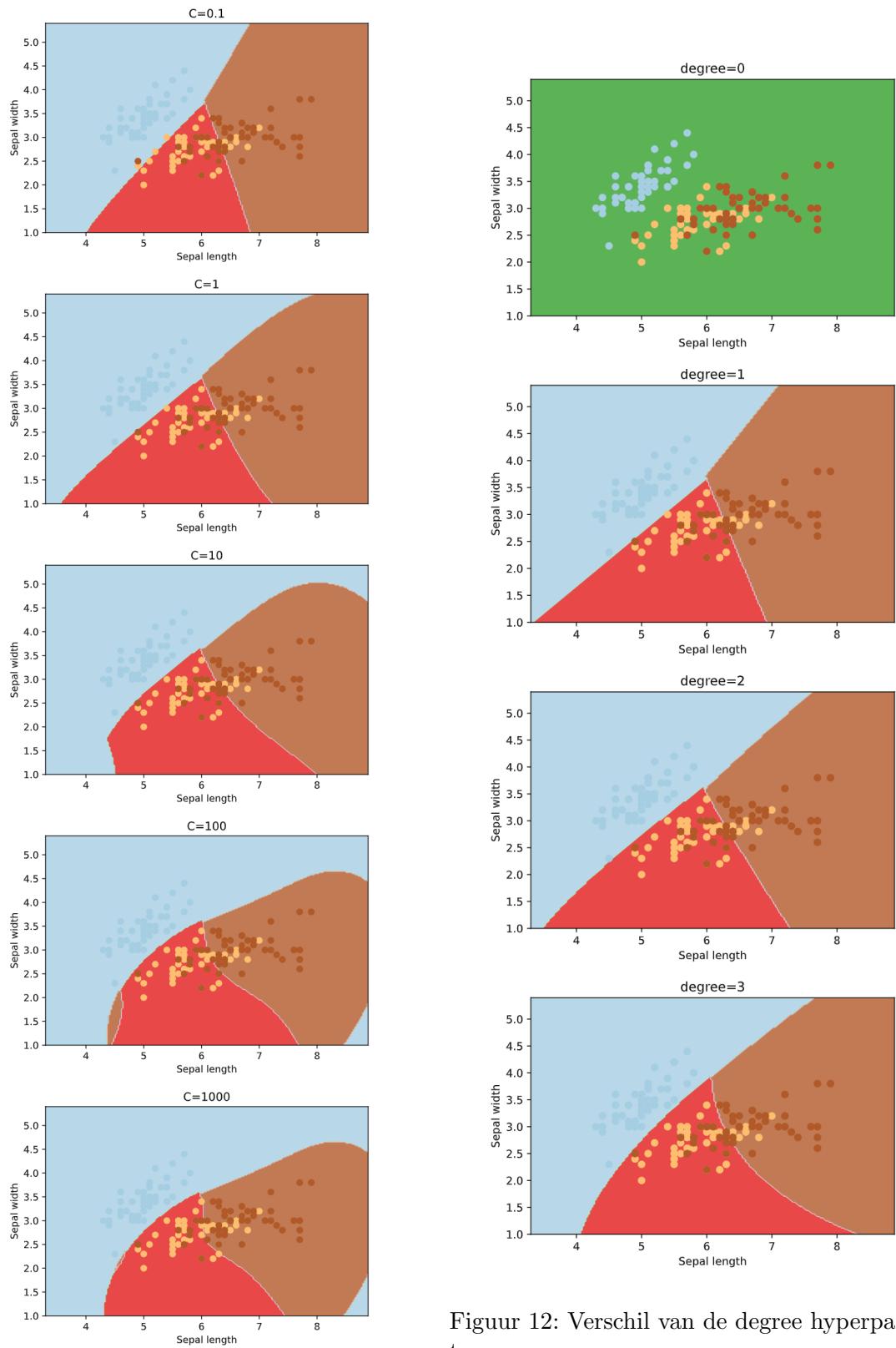
Figuur 8: Dataset voorbereiden



Figuur 9: Verschil van de kernel hyperparameter



Figuur 10: Verschil van de gamma hyperparameter



Figuur 11: Verschil van de c hyperparameter

Figuur 12: Verschil van de degree hyperparameter

```
1 # S67 - Model analysis - Model validation
2 print(accuracy_score(Y_validation, predictions))
3 print(confusion_matrix(Y_validation, predictions))
4 print(classification_report(Y_validation, predictions))
```

Figuur 13: Dataset voorbereiden

```
0.9666666666666667
[[11  0  0]
 [ 0 12  1]
 [ 0  0  6]]
      precision    recall   f1-score   support
Iris-setosa       1.00     1.00     1.00      11
Iris-versicolor   1.00     0.92     0.96      13
Iris-virginica    0.86     1.00     0.92       6
                                           accuracy         0.97      30
                                           macro avg     0.95      30
                                           weighted avg  0.97      30
```

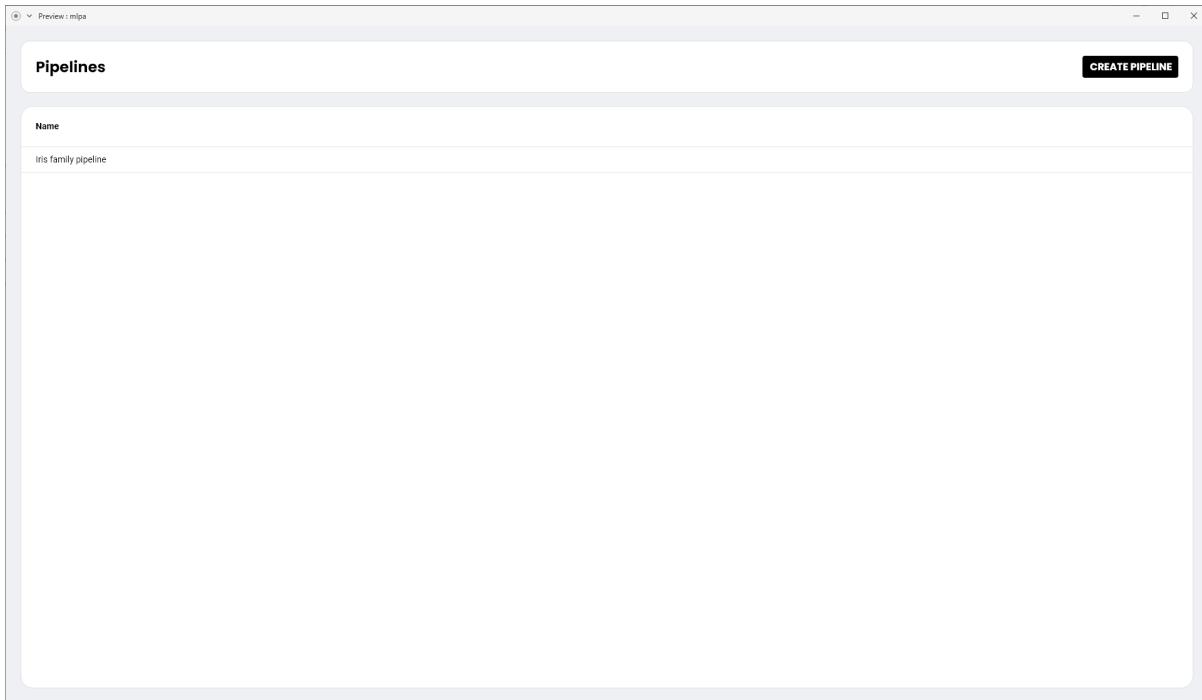
Figuur 14: Dataset voorbereiden

VII Stack configuratie van het experiment met Pulumi

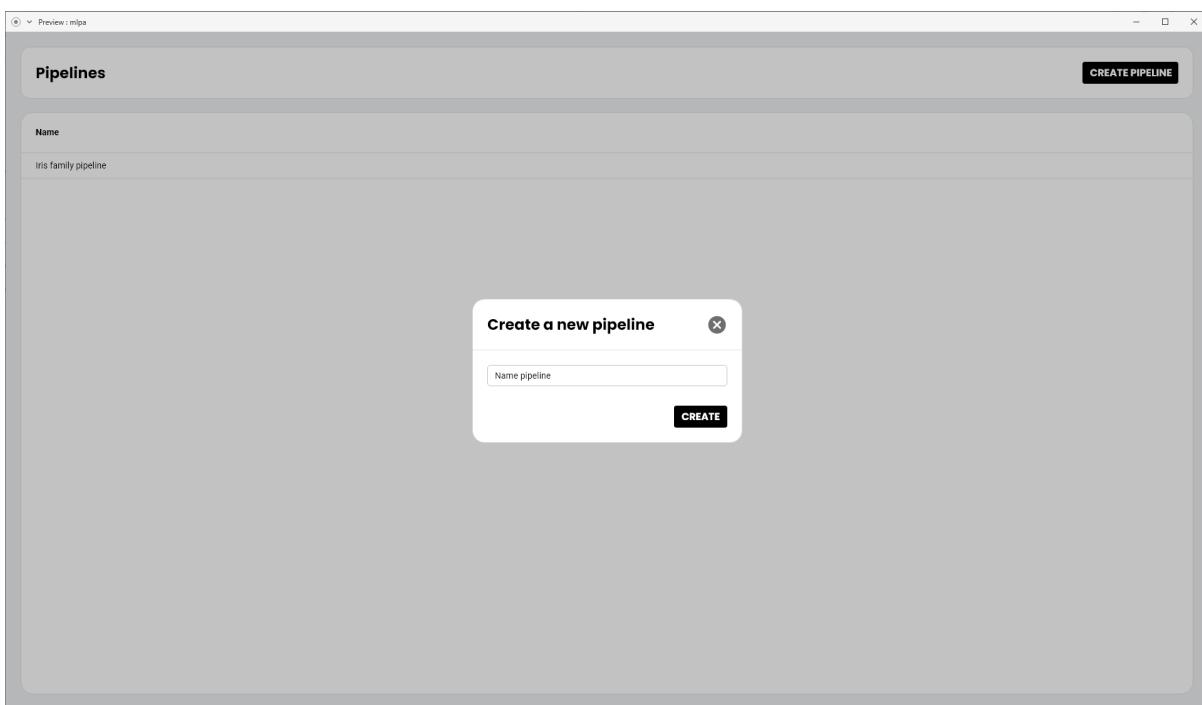
```
1 const CreateResource = (message: string) => async () => {
2   const computeNetwork = new compute.Network("network", {
3     autoCreateSubnetworks: true,
4     project: "disco-sky-312109"
5   })
6
7   const computeFirewall = new compute.Firewall("firewall", {
8     project: "disco-sky-312109",
9     network: computeNetwork.selfLink,
10    allows: [
11      {
12        protocol: "tcp",
13        ports: ["22", "80"],
14      },
15    ],
16  })
17
18  const startupScript = `#!/bin/bash
19 echo "${message}" > index.html
20 nohup python -m SimpleHTTPServer 80 &
21
22  const computeInstance = new compute.Instance("instance", {
23    project: "disco-sky-312109",
24    machineType: "f1-micro",
25    metadataStartupScript: startupScript,
26    bootDisk: {
27      initializeParams: {
28        image: "debian-cloud/debian-9-stretch-v20181210",
29      },
30    },
31    networkInterfaces: [
32      {
33        network: computeNetwork.id,
34        accessConfigs: [{}], // must be empty to request an ephemeral IP
35      },
36      serviceAccount: {
37        scopes: ["https://www.googleapis.com/auth/cloud-platform"],
38      },
39      zone: "europe-west1-c"
40    }, { dependsOn: [computeFirewall] })
41
42  const ip = computeInstance.networkInterfaces.apply(ni => ni[0].accessConfigs![0].natIp);
43
44  return { name: computeInstance.name, ip: ip }
45 }
```

Figuur 15: Stack configuratie van het experiment met Pulumi

VIII Mock up - Overzicht pipeline en pipeline aanmaken

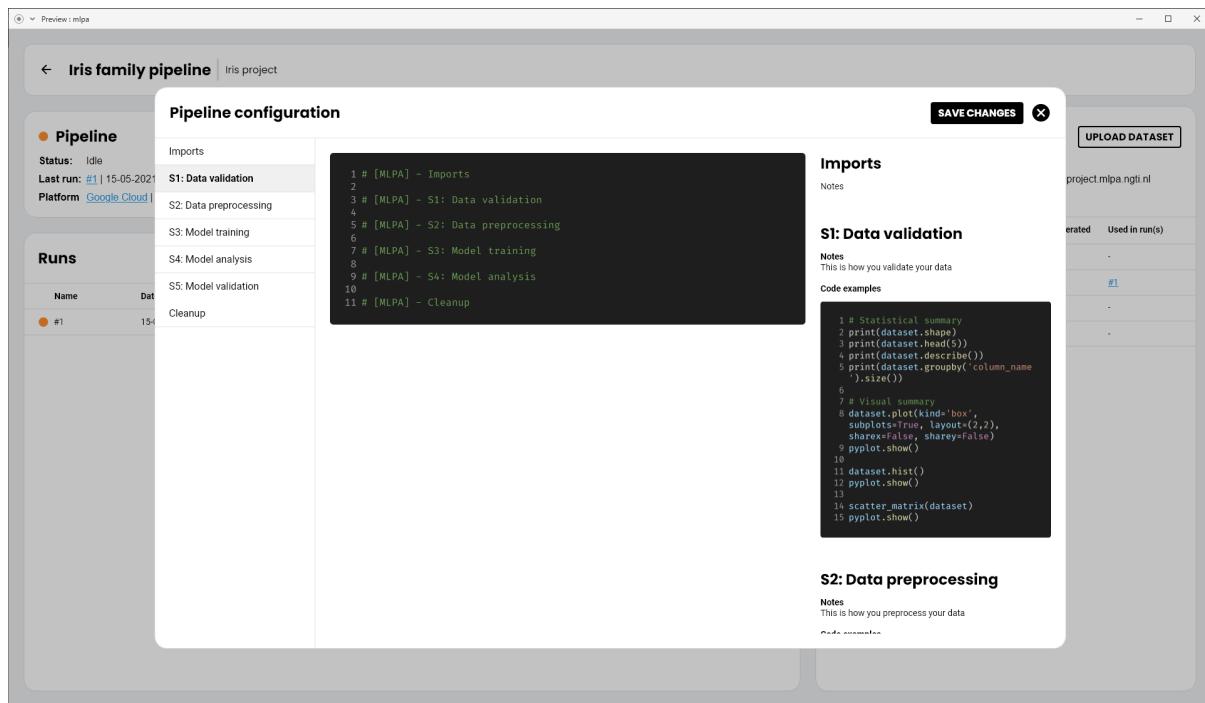


Figuur 16: Mock up van het pipeline overzicht



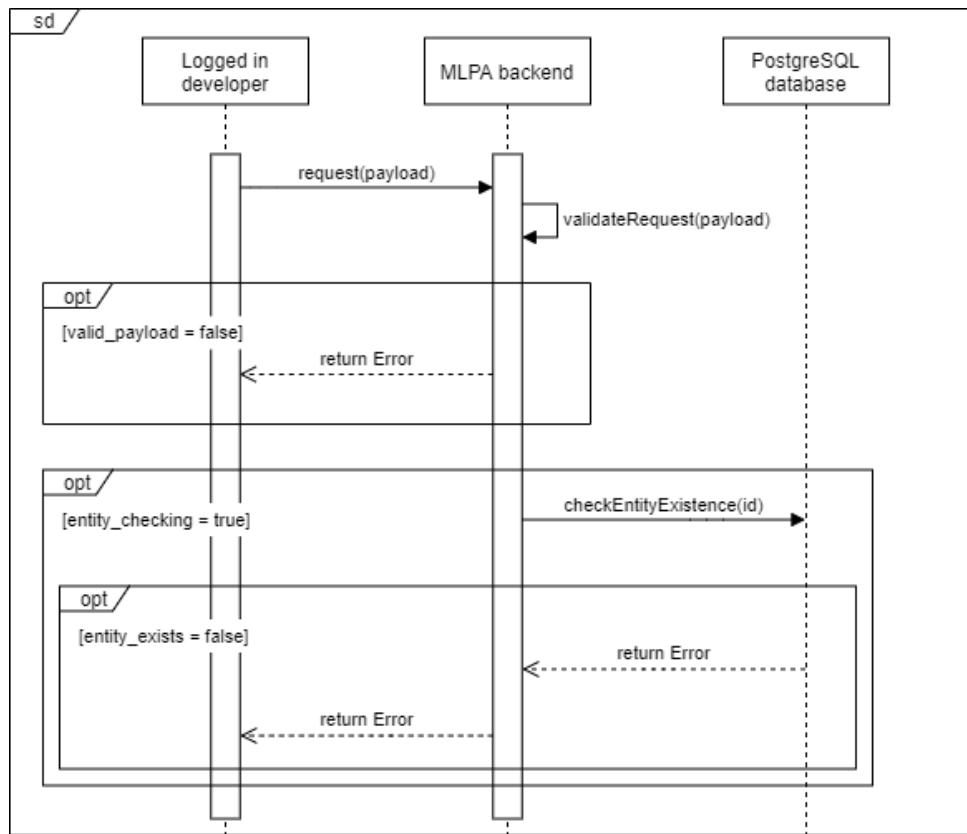
Figuur 17: Mock up van het aanmaken van een pipeline

IX Mock up - Configuration dialoog



Figuur 18: Mock up van het configuratie dialoog

X Validation sequence diagram



Figuur 19: Validation sequence-diagram

XI Proof of concept code - *gc_getRunStatus()*

```
1 export const gc_getRunStatus = async (name: string, run: number, run_id: number) => {
2   const compute = new Compute()
3   const zone = compute.zone('europe-west4-a')
4   const vm = zone.vm(`${name}-${run}`)
5
6   const metadata = await vm.getMetadata()
7   const output = await vm.getSerialPortOutput()
8
9   await updateRun(run_id, { status: metadata[0].status, output: output[0] })
10
11   return await output[0]
12 }
```

Figuur 20: Code om de status van een VM op te halen in Google Cloud

XII Proof of concept code - VM output in de front-end

Run

```
6.900000 2.500000
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script: class
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script: Iris-setosa 50
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script: Iris-versicolor 50
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script: Iris-virginica 50
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script: dtype: int64
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script: 0.9666666666666667
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script: [[1 0 0]
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script: [0 12 1]
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script: [0 0 6]]
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script: precision recall f1-score
support
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script:
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script: Iris-setosa 1.00 1.00 1.00
11
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script: Iris-versicolor 1.00 0.92 0.96
13
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script: Iris-virginica 0.86 1.00 0.92
6
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script:
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script: accuracy 0.97 30
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script: macro avg 0.95 0.97 0.96
30
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script: weighted avg 0.97 0.97 0.97
30
Jun 12 13:13:32 test-pipeline-27 GCEMetadataScripts[2498]: 2021/06/12 13:13:32 GCEMetadataScripts: startup-script:
Jun 12 13:13:33 test-pipeline-27 google_metadata_script_runner[2498]: 2021/06/12 13:13:33 logging client: rpc error: code = Unauthenticated desc = transport: metadata: GCE metadata "instance/service-accounts/default/token?scopes=https%3A%2F%2Fwww.googleapis.com%2Fauth%2Flogging.write" not defined
[ 174.666435] google_metadata_script_runner[2498]: 2021/06/12 13:13:33 logging client: rpc error: code = Unauthenticated desc = transport: metadata: GCE metadata "instance/service-accounts/default/token?scopes=https%3A%2F%2Fwww.googleapis.com%2Fauth%2Flogging.write" not defined
[0;32m OK [0m] Started snap.google-cloud-sdk.gcloud...5a0b-4ff3-ad34-ca3b15b1681b.scope.
Jun 12 13:13:34 test-pipeline-27 systemd[1]: Started snap.google-cloud-sdk.gcloud.11babaa84-5a0b-4ff3-ad34-ca3b15b1681b.scope.
```

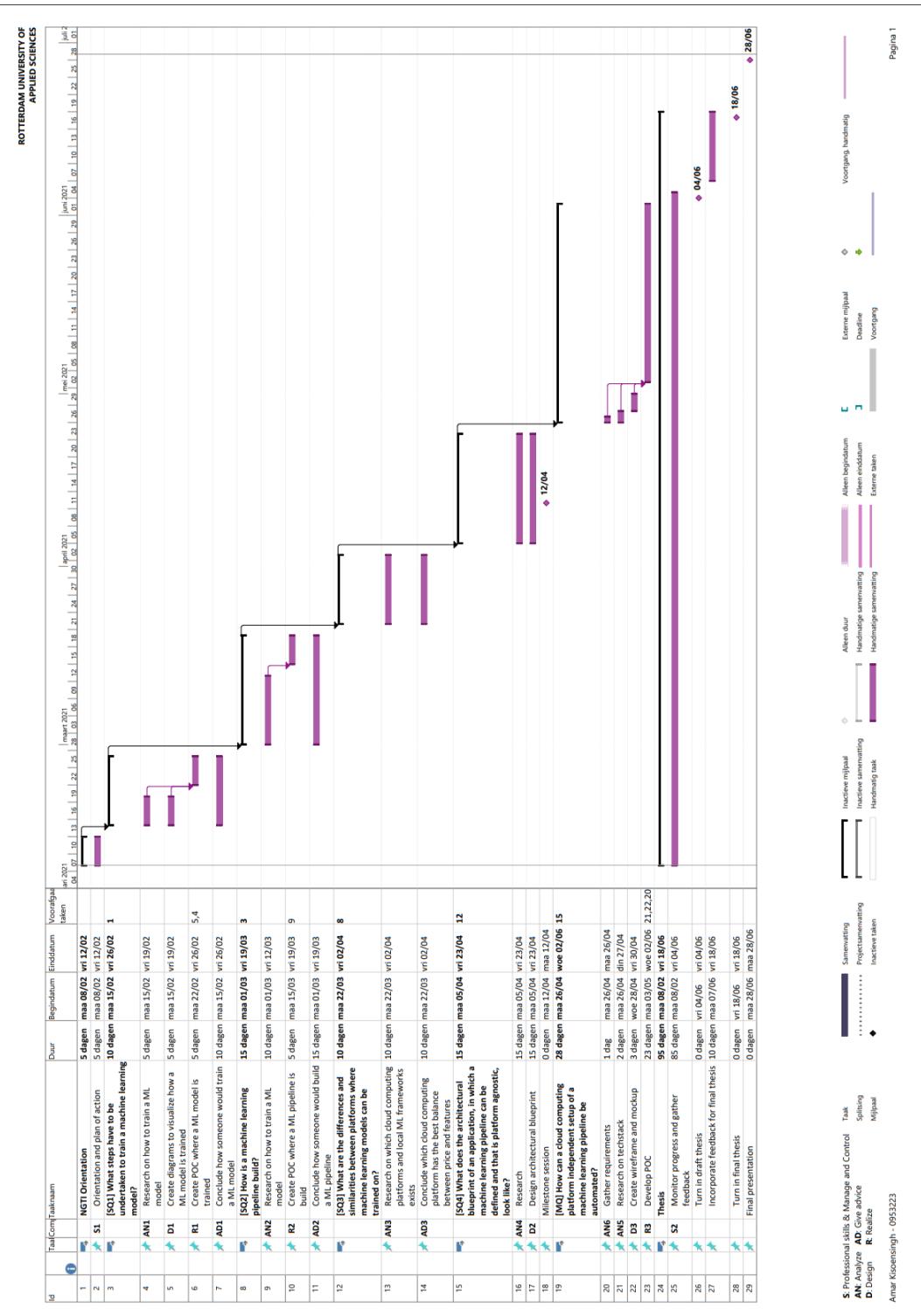
Figuur 21: Virtual machine (VM) output in de front-end

XIII Proof of concept code - *gc_stopRun*

```
1 export const gc_stopRun = async (name: string, run: number, run_id: number) => {
2   const compute = new Compute()
3   const zone = compute.zone('europe-west4-a')
4   const vm = zone.vm(`${name}-${run}`)
5
6   await vm.delete()
7
8   const _run = await fetchRun(run_id)
9
10  await updateRun(run_id, { status: 'TERMINATED' })
11  await updatePipeline(_run.pipeline.id, { status: 'IDLE' })
12 }
```

Figuur 22: Code om een virtual machine (VM) te verwijderen in Google Cloud

XIV Globale planning



Figuur 23: Globale planning

XV Notulen bij meetings

18-06 // Intervisie 7 18-6-21 Hoeveel tijd duurt de scrip...
10-06 // Update stage 10-6-21 Oefenpresentatie
04-06 // Intervisie 6 4-6-21 Cijfers uitschrijven
01-06 // Sync Kolja 1-6-21 Benoem dat je experimente...
27-05 // Update stage 27-5-21 Voortgang scriptie is een ...
25-05 // Sync Kolja 25-5-21 Waarom niet andere fram...
21-05 // Intervisie 5 21-5-21 Ontvankelijkheid
19-05 // Sync Kolja 19-5-21 Aanmaken kost tijd
12-05 // Update Stage 15-5-21 T shirt heeft P beïnvloed e...
11-05 // Sync Kolja 11-5-21 Navbar weghalen
06-05 // Update stage 6-5-21 Moet de resources van de ...
04-05 // Sync Kolja 4-5-21 Diagrammen

Figuur 24: Notities bij elke meeting

XVI Getuigenverklaring bedrijfsbegeleider

 Feedback van bedrijfsbegeleider over de afstudeerde	
Naam student:	Amar Kisoensingh
Studentnummer:	0953223
Opleiding:	Informatica
Titel eindverslag:	Machine Learning Pipeline Automation
Inleiding Met dit formulier wordt naar uw mening gevraagd over het functioneren van de afstudeerstudent. Hierbij komen aan bod: <ul style="list-style-type: none">• Personal skills en manage en control• De vijf ICT-competenties• Een vrij veld waar u zelf aanvullende observaties kan vermelden Het formulier hoeft niet door één persoon ingevuld te worden: meningen van opdrachtgever, begeleider, projectleider en collega's kunnen meegenomen worden (eventueel in bijlagen). Uiteindelijk moet het formulier door de begeleider ondertekend worden. Per onderdeel wordt gevraagd de sterke punten en de verbeterpunten aan te geven. Wees a.u.b. zo expliciet mogelijk, zodat een goed beeld over het functioneren van de student gevormd kan worden.	
Professional skills & Manage en control Aandachtspunten hierbij zijn: communicatie, samenwerking, inzet, zelfstandigheid, planning en voortgang Aandachtspunten hierbij zijn: planmatig werken, projectvoortgang bewaken, principes toepassen om een softwareontwikkelproces te managen en te bewaken.	
Sterke punten Amar kan goed zelfstandig werken en sterk schrijven & presentaties maken. Ook zijn planning heeft hij goed opgepakt en in de gaten gehouden na een aarzelende start. Verbeterpunten Geen	
Analyseren Aandachtspunten hierbij zijn: doelgerichtheid, keuze van de juiste deelvragen die beantwoord moeten worden om tot een goed beroepstechnische probleemstelling te komen, diepgang van de analyses, combinatie theorie en eisen uit de praktijk, kritisch gebruik bronnen, helder vaststellen van de problemen.	
Sterke punten Analyseren was erg sterk van Amar, zeker op de inhoud. Verbeterpunten Amar moet over zijn onzekerheid van zijn eigen keuzes stappen. Hij mag zelfverzekerder zijn, omdat hij laat zien dat hij het kan. Wel iets te lang in de theorie blijven hangen, eerder toepassen en dan terugkoppelen zou hem veel tijd bespaard hebben.	
Adviseren Aandachtspunten hierbij zijn: gemaakte keuzes en onderbouwing hiervan, volledigheid (bijvoorbeeld kwaliteitsaspecten, security, schaalbaarheid, performance, privacy), presentatie van de adviezen.	
Sterke punten Goede keuzes en sterke onderbouwingen. Overziet het spectrum, misschien iets te veel. Verbeterpunten	

Figuur 25: Pagina 1 van de getuigenverklaring bedrijfsbegeleider

<p>CURSUSHANDLEIDING</p> <p>Ook hier iets te veel in de theorie en mogelijkheden blijven hangen. Eerder naar toepasbaarheid en haalbaarheid kijken.</p> <p>Ontwerpen Aandachtspunten hierbij zijn: keuze van relevante ontwerptechnieken (FO, TO, UI, DB), kwaliteit van de ontwerpen, architectuur, teststrategie, toetsing van de ontwerpen (prototyping, voorleggen aan experts).</p> <p><u>Sterke punten</u> Zelfstandig alle aspecten goed doorlopen.</p> <p><u>Verbeterpunten</u> Voert het ook hier wat te ver door soms. Eerder laten zien en toetsen zou hem tijd hebben bespaard.</p> <p>Realiseren Aandachtspunten hierbij zijn: kwaliteit van het beroepstechnische en kwaliteit van het ontwikkelproces. Clean code, documenteren, testen, integreren, continuous integration & deployment, gebruik frameworks.</p> <p><u>Sterke punten</u> Complexe materie eigen gemaakt en gerealiseerd. Uitstekend gedaan.</p> <p><u>Verbeterpunten</u> Geen.</p> <p>Aanvullende observaties Amar is iets te perfectionistisch, deels gevoed door onzekerheid, maar er zit een kei van een ontwikkelaar in hem. Hij is een veelbelovend talent in mijn ogen, die nu al tegen high junior niveau aanzit. Als hij meer in zichzelf gaat geloven kan hij heel ver komen. Wij zijn erg onder de indruk van hoe hij zijn opdracht heeft volbracht, een van de beste die ik voorbij heb zien komen (van de tientallen in 25 jaar). Wij zouden hem dan ook graag inlijven voor onze organisatie.</p>	 <p>HOGESCHOOL ROTTERDAM/CMI</p>
---	---

Datum: 18-6-2021
Naam bedrijfsbegeleider: Kolja van der Vaart
Handtekening:


CURSUSHANDLEIDING AFSTUDEREN INFORMATICA 1 FEBRUARI 2021 2

Figuur 26: Pagina 2 van de getuigenverklaring bedrijfsbegeleider