2018

Scriptie

Atos Amstelveen

VISMIGRATIEROUTES RAPPORTEREN EN VOORSPELLEN MET MACHINE LEARNING

Brammerloo, Imro

Studentnummer: 0899670

Begeleider(s): Michel Metselaar

Versie 16 [2018-06-20]

SAMENVATTING

De visstand neemt af mede omdat vismigratieroutes worden afgekneld door waterwerken. Als antwoord hierop zijn er vispassages aangelegd, zodat vissen vrij baan hebbe. Het is thans niet gekwantificeerd of en hoezeer deze gebruikt worden. Dit onderzoek gaat over hoe met behulp van data van devices, gegevens over biologische viseigenschappen en open data een voorspelling gemaakt kan worden van een passerende vissoort door middel van machine learning.

Er wordt onderzocht welke inherente biologische eigenschappen en gedragingen van de vis en indirecte invloeden als leefomgeving ingekaderd kunnen worden tot kwantificeerbare parameters van data. Ook wordt er vanaf een beginpunt inzichten vergaard over machine learning, die gaandeweg steeds dichter naar het uiteindelijke doel van vismigratie voorspellen groeien.

Kennis over vissoorten, vismigratie en machine learning is vergaard door exploratief onderzoek en literatuuronderzoek. Gaandeweg worden hypotheses getoetst aan de hand van de verzamelde gegevens en de geselecteerde algoritmes. Uiteindelijk wordt er een ontwerp opgesteld waarin de data en het algoritme zo zijn opgesteld dat er voorspellingen gemaakt kunnen worden.

Voorspelling blijkt algeheel afhankelijk van de hoeveelheid data die beschikbaar is. De voorspelling die gedaan zijn, zijn op basis van de verzamelde gegevens. De hoeveelheid data en gewenste parameters waren in mindere mate gevonden en beschikbaar. Aangeraden wordt dan ook de nodige informatiestromen te garanderen waardoor voorspellingen een meerwaarde hebben met betrekking tot voorspelling van de vismigratie.

HOOFDVRAAG

Hoe kan vismigratie worden voorspeld met machine learning?

DEELVRAGEN

- Welke data is belangrijk bij het voorspellen en rapporteren van vismigratie?
- Wat zijn de huidige standaarden in machine learning als het gaat over voorspelling?
- Welke algoritmes zijn het beste geschikt voor het voorspellen van vismigratie?

RAAKVLAKKEN EINDCOMPETENTIES

Adviseren	Ongeacht of er een werkbaar Proof of Concept wordt geleverd, zal aan het eind van de onderzoeken aanbevelingen worden gemaakt waarop verder gebouwd kan worden. Indien uit het onderzoek blijkt dat voortbouwen niet mogelijk is, dan zullen alle argumenten en bevindingen worden genoemd waaruit lering getrokken kan worden. Ook kunnen nieuwe inzichten worden geleverd in een andere discipline (vooral faunabeheer) aan de hand van de onderzoeksbevindingen.
Analyseren	Uit onderzoek naar vismigratie en de van daarop toepasbare machine learning zullen geldende standaarden en toepassingsparameters (m.b.t. vismigratie) naar voren komen. Er zullen efficiëntietests worden uitgevoerd en geanalyseerd worden. Deze zullen in het kader van overdraagbaarheid toegepast worden op deze afstudeeropdracht.
Beheren	Door voortgang bij te houden in een logboek kan deze voortang tegen de planning houden om de algehele voortgang te bewaken. Als blijkt dat hierin achterstand oploopt door een verkeerde oplossingsrichting of andere deficiëntie kan er nog tijdig worden bijgestuurd. Zo zal getracht worden controle over het project te behouden.
Ontwerpen	Een deel van het ontwerp is al verwerkt in het mandaat. Daarnaast zal het bij het onderzoeken naar de meest efficiënte algoritmes worden gekeken naar diens gangbare ontwerpen, met name bestaande implementatiestandaarden, uitbreidbaarheid en toepasbaarheid in bredere context(en). De algoritmes die in het kader van dit onderzoek gemaakt zullen worden zullen voldoen aan deze standaarden, zover mogelijk. Het geheel, de visdata in combinatie met de machine learning, zullen in een geheel ontwerp worden opgenomen.
Realiseren	Aan het eind van het onderzoek zal er een ontwerp framework worden aangeleverd dat gemaakt is om met meerdere typen data verschillende rapporten en voorspellingen te maken. Als het mogelijk is, kan deze ook toegepast worden op meerdere platformen. Er zal dan ook rekening worden gehouden met mogelijk nieuwe of meer efficiënte sensoriek of omgevingsvariabelen, welke door simpele aanpassingen in variabelen kunnen worden geüpdatet in de algoritmes.

Inhoudsopgave

Figuren

Tabellen

Code

BEGRIPPENLIJST

Termen met betrekking tot vissen	
Anadroom	Een vissoort die in paaitijd van de zee naar de rivieren trekt om te paaien. Hierbij trekt de vis dus van zout water naar zoet(er) water
Eutrofiëring	Het stijgen van voedselrijke stoffen (nutriënten) voor bepaalde aquatische planten in het water. Oorzaken zijn het vergaan van organische stoffen, zoals afvalwater en mest. Gebrek aan doorstromend (schoon) water draagt hier ook aan bij. Bij een sterke eutrofiëring kunnen dominante planten (voornamelijk algen) explosief groeien en andere planten verdrukken, waardoor overige planten kunnen verdwijnen. Ook op vissen heeft dit een nadelig effect: door toename van blauwalgen wordt het water zuurder, toxischer, zuurstofarmer en troebeler. In het algemeen leidt eutrofiëring tot een algehele afname van diversiteit van aquatische flora en fauna.
Hardheid water	Waterhardheid geeft de concentratie van metaal-ionen aan. Water met een hoge hardheid zoals kraanwater laten kalkaanslag achter. De eenheid die in deze scriptie gehanteerd wordt is dH (deutsche Härte). 1 dH staat voor 17.8 gram kalk per m^3.
Katadroom	In tegenstelling tot anadroom is een katadrome vissoort een die van de rivieren (zoet) trekt naar de zee om daar te paaien.
Paaien	Het voorplantingsmechanisme van de meeste vissen, voornamelijk beenvissen. Hierbij worden de eieren afgezet in het water.
Paaitijd	Tijdsperiode waarin een vis paaigedrag vertoond. Deze tijdsperiode verschilt per vissoort. Ook watertemperatuur heeft hier een invloed op.

Termen met betrekking tot machine learning	
Dataset	In principe een data sheet met koppen en desbetreffende waardes. Waarden die niet bijdragen of zelfs negatieve weerslag hebben op het uiteindelijke getrainde model, moeten worden gemuteerd of verwijderd. Dit proces wordt ook wel een dataset 'cleanen'.
Overfitting	Verschijning waarin een algoritmisch model zodanig precies op een dataset is getraind, dat het uiteindelijke algoritme waarden die te veel afwijken van originele training set verkeerd worden voorspeld. Het algoritme let dan zo veel op de details dat initiële afwijkende waarden of 'ruis' in het model getraind worden, waarbij dus ook bij voorspellingen dezelfde soort afwijkende waarden aanwezig zijn.
Testset	Een gedeelte van een dataset dat gebruikt wordt om een algoritmisch model te testen op nauwkeurigheid. Dit gebeurd door te kijken of de voorspelde waarden van het algoritmisch model geheel of voldoende overeen komen met de betreffende waarden in de testset. Normaliter beslaat de testset rond de 30% van de originele dataset.
Trainingsset	Een gedeelte van een dataset dat gebruikt wordt om een algoritmisch model te 'trainen'. Het uiteindelijke model gebruikt dus in principe de trainingsset als referentiekader. Normaliter beslaat de trainingsset rond de 70% van de originele dataset.
Supervised	Supervised machine learning heeft meestal betrekking op classificatie en regressie. Het model wordt getraind op vooraf bekende termen (labels) of bepaalde numerieke waarden. Het contextuele verband is al meegegeven in het begin. Doel bij supervised learning is om een nieuwe voorspelling of classificatie te maken aan de hand van de desbetreffende input.
Unsupervised	Bij unsupervised machine learning zijn onderlinge verbanden tussen waarden vooraf niet bekend. Het model wordt getraind door tussen (schijnbaar) willekeurige waarden een bepaalde correlatie te vinden. Zodoende leert het model elementen van een trainingsset in te delen in deze bevonden verbanden (clustering). Doel kan zijn het verband an sich te achterhalen, of nieuwe instanties in te delen in de gecategoriseerde clusters.

INLEIDING

PROBLEEMSTELLING

Op de huidige markt is er weinig tot geen aanbod op geautomatiseerd bijhouden van vismigratie, volgens Atos. Hier wil Atos op inspringen om een mogelijke vraag naar deze technologie te verschaffen voor andere concurrenten de kans hebben. Zodoende zou Atos deze dienst mogelijk kunnen aanbieden bij geïnteresseerde partijen.

Hieruit is de volgende opdracht opgesteld om bovenstaande probleemstelling aan te pakken: Een onderzoek dat zich richt op het verzamelen van data over vismigratieroutes middels IoT en nadat er voldoende data over is verzameld het kunnen voorspellen van de migraties.

De opdracht van vismigratie is tweeledig:

1. Onderzoeken en realiseren van IoT componenten welke een vismigratie kunnen meten en doorgeven aan een IoT Platform waar de data voorbereid wordt voor opslag en een dashboard met inzichten over de ontvangen data.
2. Onderzoeken en realiseren van de meest geschikte algoritmen welke een vismigratie kunnen rapporteren en voorspellen.

Het onderzoek dat hier zal worden gedaan gaat in op opdracht 2, en is in deze de hoofdonderzoeksvraag en onderwerp van dit document.

Opdracht 1 hoort bij een ander onderzoek, 'Vismigratie met IoT', en wordt door een andere onderzoeker volbracht.

AFBAKENING

Er is een scheidingslijn tussen de onderzoeker naar 'Vismigratie met IoT' en mijn werkzaamheden en onderwerpen. Zo zal het onderzoek en prototype niet behelzen:

- Het maken van keuzes omtrent het device dat de metingen doet
- Bezighouden met connectiviteit device en backend
- Inhoudelijke keuzes maken over welke sensoren kwalitatief optimaal zijn
- Zoutwatervissen (inheems) en zoute wateren
- Rekening houden met of incalculeren van niet-vis objecten

Verder zijn er nog overige afbakeningen:

- Geen marktonderzoek naar soortgelijke producten.

DOELSTELLING/ ONDERZOEKSVRAAG

Zoals hierboven genoemd, is het doel om te onderzoeken en realiseren van de meest geschikte algoritmen welke een vismigratie kunnen rapporteren en voorspellen.

Hieruit is de volgende onderzoeksvraag opgesteld: **Hoe kan vismigratie worden voorspeld met machine learning?**

DEELVRAGEN

Dit zijn deelvragen die voortvloeien uit de hoofdvraag "Hoe kan vismigratie worden voorspeld met machine learning?"

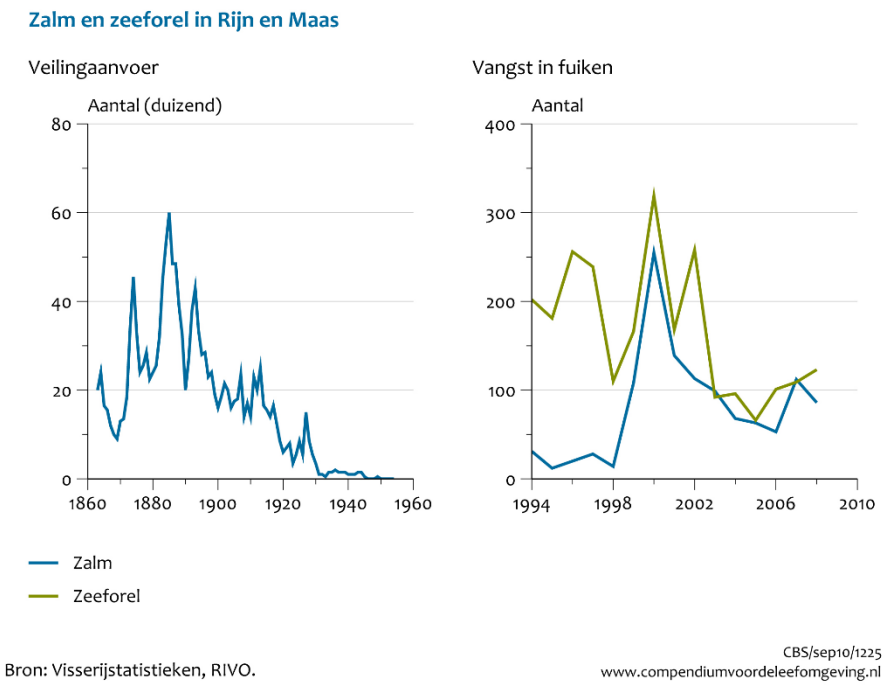
- Welke data is belangrijk bij het voorspellen en rapporteren van vismigratie?
- Wat zijn de huidige standaarden in machine learning als het gaat over voorspelling?
- Welke algoritmes zijn het beste geschikt voor het voorspellen van vismigratie?

RATIONALE

Het gaat niet goed met de visstand in Nederland. Vergeleken met het verleden zwemt er veel minder vis door Nederland [Figuur 1]. Een lage visstand in slecht voor de biodiversiteit. Een van de vele oorzaken hiervan is beschadiging of dood van de vis door installaties in watergangen, en het blokkeren van de vismigratieroutes [1].

Een van de oplossingen voor dit probleem is het aanleggen van vispassages of vistrappen. Dit zijn veilige omwegen voor vissen om deze obstakels te vermijden.

Het probleem is dat het onvoldoende inzichtelijk is in welke mate gebruik wordt gemaakt van deze vispassages. Hiertoe is de vraag ontstaan hoe dit gekwantificeerd kan worden.



Figuur 1: Zalm en zeeforel in Rijn en Maas

ONDERZOEKSMETHODIEK

Hier wordt beschreven hoe het onderzoek naar de hoofdvraag wordt gepleegd. Dit gebeurt door stapsgewijs antwoorden op de deelvragen te verkrijgen. De volgorde van het onderzoek valt veelal overeen met de gestelde volgorde van deelvragen.

WELKE DATA IS BELANGRIJK BIJ HET VOORSPELLEN EN RAPPORTEREN VAN VISMIGRATIE?

Onderzoek op deze deelvraag is tweeledig.

1. Exploratief en literair onderzoek vaar het verschijnsel vismigratie, en daarbij zoeken naar te kwantificeren parameters die later voor voorspelling van onder andere vissoorten van belang is.
2. Het onderzoeken van welke van bovenstaande parameters gemeten kunnen worden door sensoren en/ of andere databronnen. Hieruit zal een inventarisatie ontstaan uit data die meetbaar is, tegenover data dat nuttig of bruikbaar is. Ook zal er worden gezocht naar beschikbare relevante data, in de vorm van logs of datasets [Begrippenlijst, p. 6]. Deze bronnen moeten bestaan uit historische data en/ of live data. Hierbij is ook afstemming met de IoT-onderzoek uitvoerder van belang, om de leverbare parameters vast te leggen. Dit deel zal deels literatuurstudie en in inventarisatieonderzoek zijn.

WAT ZIJN DE HUIDIGE STANDAARDEN IN MACHINE LEARNING ALS HET GAAT OVER VOORSPELLING?

Hier zal dieper worden ingegaan over het fenomeen machine learning. Er zal worden gekeken naar gebruikte algoritmen en methodieken, die met elkaar zullen worden vergeleken. Er zal gekeken worden hoezeer huidige standaarden in voorspelling de resultaten van de vorige deelvraag tegemoetkomen. Hierin zullen de onderzoeksresultaten van de vorige deelvraag (alle soorten data die noodzakelijk zijn bij het zo goed mogelijk voorspellen van vismigratie) zwaarder wegen dan het meest efficiënte algoritme.

Na genoeg kennis te hebben genomen van de huidige standaarden, zal er een lijst worden opgesteld van meest gebruikte en aangeraden algoritmes. Deze lijst zal verder worden ingedeeld in subcategorieën geordend op functionaliteit en doelstelling.

In het begin zal er dus sprake zijn van fundamenteel onderzoek dat snel zal overgaan in literatuuronderzoek. Dit zal later overgaan in toetsend onderzoek.

WELKE ALGORITMES ZIJN HET BESTE GESCHIKT VOOR HET VOORSPELLEN VAN VISMIGRATIE?

Aan de opgestelde lijst van algoritmes van vorige deelvraag zullen deze worden getoetst op:

1. Theoretische functionaliteit analyse (analyse op basis van literatuuronderzoek)
2. Efficiëntie op snelheid en accuratesse (op basis van benchmark testen in R)
3. Testen op datasets met betrekking tot dit project

Deze toetsing wordt uitgevoerd in de volgende fasen. Deze worden in de koppen hieronder toegelicht.

THEORETISCHE FUNCTIONALITEIT ANALYSE

De top 10 van meest populaire algoritmes uit de vorige deelvraag worden verder ingedeeld op intrinsieke eigenschappen. Dit is om een beter idee te krijgen waar deze algoritmes het meest geschikt voor zijn, en om de resultaten te vergelijken met de eigenschappen die nuttig of noodzakelijk zijn in het kader van dit project. Deze eigenschappen worden hier ook opgesteld en geanalyseerd.

EFFICIËNTIE OP ACCURATESSE

Er worden benchmark tests gedaan op de top 10 van de meest populaire algoritmes in R op algemene datasets. Hierop wordt gelet op accuratesse.

TESTEN OP DATASETS

Dezelfde lijst met algoritmes wordt getest op datasets die in het verwachtingspatroon liggen van dit project. Hier wordt eveneens ook accuratesse in acht genomen. Deze datasets kunnen plaatselijk opgevuld worden met gegenereerde data als de volledigheid van de data niet toereikend is. Van deze dataset wordt dan een trainingset en testset [Begrippenlijst, p. 6] opgesteld.

ONDERZOEKSRESULTATEN

In dit hoofdstuk wordt de uitvoering en de resultaten van onderzoeken beschreven. Deze onderzoeken zijn per deelvraag ingedeeld. Delen die te uitvoerig zijn wordt naar gerefereerd naar bijlages. De conclusies van deze resultaten worden besproken in het volgende hoofdstuk.

WELKE DATA IS BELANGRIJK BIJ HET VOORSPELLEN EN RAPPORTEREN VAN VISMIGRATIE?

WAT IS VISMIGRATIE?

Om een idee te krijgen over het onderwerp vismigratie en mogelijke oplossingsrichtingen, is het belangrijk eerst te verdiepen in vissen en vismigratie. Er kunnen geïnformeerde beslissingen kunnen gemaakt.

BESTAANDE KENNIS

Nederland doet aan duurzaam waterbeheer. Dit is het winnen, gebruiken en teruggeven van water aan de natuur op een duurzame manier [2]. Er zijn verschillende beleidsvormen waarin dit wordt gehandhaafd. In dit project is de technische aspect van toepassing, namelijk waterwerken. Sommigen van deze waterwerken blokkeren de doorgang waar vissen door migreren.

Door de visstand van wateren te weten, kan er kennis worden genomen van de schoonheid en gezondheid van die wateren. Om een vissenpopulatie te laten floreren, is het belangrijk dat vissen zich kunnen bewegen tussen paai- en leefgebieden. Afsluiting van leefgebieden kan stagnatie of zelf afsterving van een (lokale) vispopulatie betekenen. Vooral voor de zalm, paling en driedoornige stekelbaars is vrije passage van belang [3].

INVENTARISATIE VISSEN IN NEDERLAND

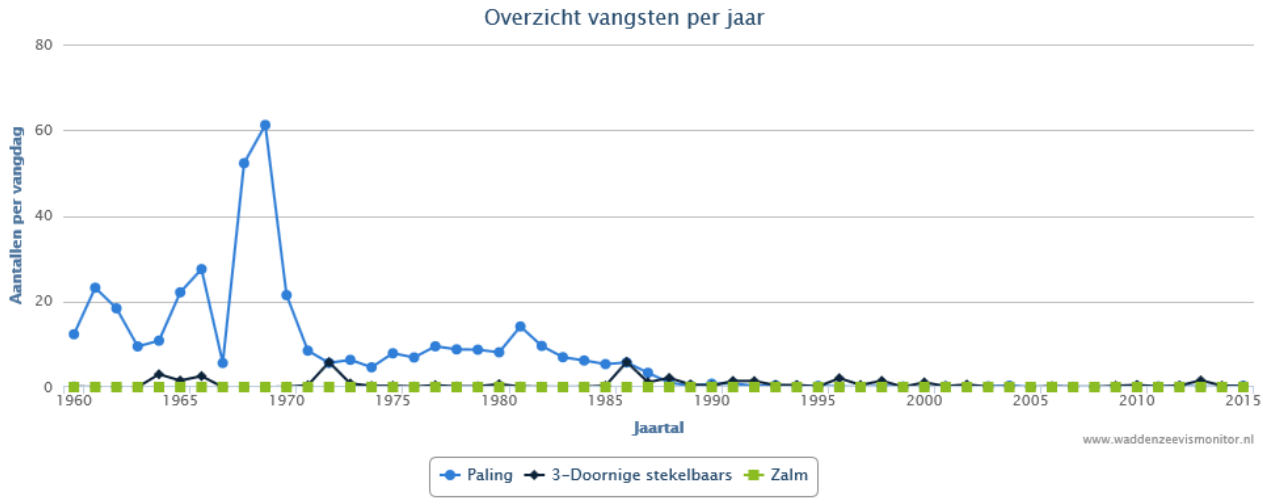
Met duidelijk genoeg geleverde informatie zou er in een gunstig geval een vissoort kunnen worden geclassificeerd. Denk aan beeldherkenning of hoge resolutie sonar of infrarood. Hieruit lijkt het logisch een lijst te maken met alle mogelijk te identificeren vissen in Nederland. Deze informatie is onder andere op het internet beschikbaar [4].

Hedendaags zijn er minsten 75 zoetwater vissoorten of inheemse vissoorten bekend in Nederland. Tijdens onderzoek bleek dat voor al elke specifieke vis alle informatie te verzamelen, niet realistisch was in de gegeven tijdspanne en andere onderzoeksstadia. Hierom is er gekozen om 3 vissen te kiezen, die een hogere prioriteit hebben.

De volgende 3 vissen zijn verkozen:

Aal/ Paling
Driedoornige stekelbaars
Zalm

Dit is omdat deze 3 vissen trekvissen zijn, en gebaat zijn bij een aaneengesloten migratieroute [3], [5]. Ook zijn de aantallen van paling en zalm in Nederland extreem gedaald als gevolg van afgesloten migratieroutes. Van paling werd zelf gedacht dat deze zou uitsterven, hoewel het recentelijk beter met deze vis gaat [6]. De visstand van de zalm lijkt nu ook te gaan verbeteren [7]. Desalniettemin zijn de visstanden nog voor veel verbetering vatbaar [Figuur 2, p. 12].



Figuur 2: Vangst vissen in de Waddenzee 1960-2015 © Waddenzee Vismonitor

MEETBARE PARAMETERS VISMIGRATIE

Om een voorspelling van een vismigratie te maken, is het van belang te weten welke vissoort deze migratie maakt. Deze vissoort moet te identificeren zijn en te onderscheiden van andere vissen. Hierbij is het van belang voldoende van elkaar variërende factoren te inventariseren. Hieronder staan deze factoren beschreven in een tabel. Dit zijn factoren die voldoende gedocumenteerd zijn om later betrouwbare modellen op te bouwen.

Categorie	Viseigenschap	Uitleg
	Lengte	
	Gewicht	
	Voedsel	
	Maximumleeftijd	
	Diepte	
Water	Temperatuur	
	Watersoort	Zoet, brak, zoet, etc.
	Hardheid	Zie begrippenlijst
Locatie gerelateerd	Leefgebied	
Paaigedrag	Paaitijd	Zie begrippenlijst
	Paaigebied	Soort water waar de vissoort doorgaans paait
	Paaigedrag	Katadroom/ anadroom
Overig	Voedseldiepte	Waar de vis zich bevindt om te voederen, zoals de bodem of aan de oppervlakte

Informatie voor deze eigenschappen is op verscheidene websites en online databases te vinden [4], [8]. Aan de hand van deze bronnen is een lijst op te stellen van gerapporteerde eigenschappen. Groot nadeel hierin is dat lang niet alle informatie even goed verdeeld is tussen vissoorten. De lijst van labels van de gerapporteerde viseigenschappen is te zien in [Bijlage A, p. 24].

De gestelde labels zijn aangepast om invoer in een database ter vergemakkelijken.

DATABRONNEN VISDATA

Het is moeilijk gebleken dataset te verkrijgen van vislogdata, in tegenstelling tot waterdata. Er zijn datasets gevonden van de 3 geprioriteerde vissen, maar deze logdata is buiten de geografische kaders van Nederland genomen [9]. De locaties liggen allen in zout oceaangebied. Ook is de data die verschaft is, deels afwijkend van de gestelde geprioriteerde eigenschappen waarop gefilterd zou worden zoals te zien in [Bijlage A, p. 24]. De gestelde eigenschappen die aanwezig zijn in deze dataset zijn beschreven in [Bijlage C, Tabel 5, p. 28-29]

De datasets bestaan elk in principe uit 2 delen.

Het 1e deel is deels gegenereerde data, gebaseerd op het latere 2e deel. Hierin is de 'legitieme' data enkel de vissoort met de locatie (lengtegraad, breedtegraad en C-Square code), waarnaast de resterende data gegenereerd is. De gegenereerde waarden zijn allen kansberekeningen (getal tussen 0 tot en met 1), waarbij in de kop van de dataset waarden staan waarop deze kansen betrekking hebben.

Het 2e deel verschaft de meeste informatie, wat reden is om dit gedeelte te gebruiken om later een algoritmisch model te trainen. Dit deel is minder omvangrijk (heeft minder records) dan de eerste, maar is in tegenstelling tot de eerste natuurgetrouw, en dus een betere basis.

DATABRONNEN VISUEEL

Er zijn ook datasets van visuele gegevens (foto en video) van vissen beschikbaar:

Dataset	Elementen	Soorten
Dataset QUT [10]	3960	468
Fish Recognition Ground-Truth data voor Fish4Knowledge Project [11]	27370	23

Er is geen diepgaand onderzoek gedaan naar visuele beeldverwerking. Reden is dat voor het device uit het onderzoek 'Vismigratie met IoT' energiezuinig moest zijn, en beeldverwerking dat niet is.

DATABRONNEN WATER

Ook andere factoren hebben een negatieve weerslag op vissen, en dus op de vismigratie. Het water zegt ook iets over de leefbaarheid voor vissen. Door goede waterkwaliteit kan de visstand beter gedijen. Vissen zijn intrinsiek verbonden met water. Daarmee is interessant waterkwaliteit en eigenschappen mee te nemen in het datamodel.

Daarnaast kan het water ook iets zeggen over het gedrag van de vis. Vissen zullen wateren eerder mijden als deze als onprettig of schadelijk worden gevonden. Daarnaast hebben vissen ook specifieke voorkeuren in welk water deze gedijen, zoals een bepaalde temperatuur, een bepaalde zuurtegraad, etc. Indien deze eigenschappen in water positief kunnen worden beïnvloed, kan de kans dat een specifieke vis voorkomt in dat water toenemen.

Om een machine learning framework te maken dat nieuwe inzichten zou kunnen verschaffen, is het wenselijk een framework te hebben dat met zo veel mogelijk sensoriële input kan werken. Hierbij is namelijk meer ruimte en kans om overeenkomsten over ogenschijnlijk niet-relevante parameters in te zien. Hieruit is het wenselijk een inventarisatie te maken met alle mogelijk meetbare parameters. Deze opstelling is tot stand gekomen middels literair onderzoek naar verschillende watersensoren. De watersensoren zijn opgedeeld in 2 categorieën: waterkwaliteit [12] en waterionen.

RIJKSWATERSTAAT

Op het internet is data beschikbaar over waterkwaliteit. Op de site van Rijkswaterstaat staat live data over verschillende categorieën waterdata [13]. De categorieën die het meeste raakvlak met dit project hebben zijn de volgende:

Waterafvoer en stroomsnelheid	Waterhoogte	Stroming	Watertemperatuur	Zoutgehalte	Waterkwaliteit

Waterafvoer	Waterhoogte t.o.v. NAP	Stroomrichting	Watertemperatuur	Zoutgehalte	Eijsden (Maas)
Stroomsnelheid		Stroomsnelheid			Lobith (Rijn)

Hierbij staan in Nederland verschillende sensors op een kaart die de behorende data weergeven en bijhouden. De aanbieders hiervan verschillen, dus er zijn locaties waar er slechts eenzijdig data beschikbaar is.

EUTROFIËRING

Zo zorgt eutrofiëring [Begrippenlijst, p. 6] voor een zeer negatieve leefomgeving voor vissen. Als de staat van het water kan worden vastgesteld, kan er een leefbaarheidsanalyse voor vissen (of bepaalde vissoorten) worden samengesteld. Dit kan van belang zijn om een leefgebied gunstiger te maken voor vissen en een hogere visstand te bespoedigen. In [Bijlage C, Tabel 4, p. 28] is een overzicht te zien van de water parameters waarmee eutrofiëring wordt ingekaderd.

MEETBARE PARAMETERS WATER

Als er verscheidene waterdata beschikbaar is, is er een kans dat er een verband tussen visgedrag of vismigratie en een bepaalde watereigenschap gevonden kan worden. Hoewel prioritering is gegeven aan bovenstaande gegevens, zullen overige waterparameters niet buiten beschouwing worden gelaten.

De totale opsomming van relevante waar te nemen eigenschappen van water is te zien in [Bijlage B, p. 26]. Deze zijn afkomstig van (externe) databronnen en sensoren die watereigenschappen kunnen meten. Referenties zijn in de koppen opgenomen.

Er is nu een opsomming gemaakt mogelijke data. Hierna volgen de resultaten van de volgende deelvraag, waarin gekeken wordt naar de basis van machine learning en langzaam naar relevante categorisering wordt gewerkt van algoritmes.

WAT ZIJN DE HUIDIGE STANDAARDEN IN MACHINE LEARNING ALS HET GAAT OVER VOORSPELLING?

Nu er een basis gelegd is van beschikbare data met betrekking tot vismigratie, moet er nu gekeken worden wat er met de data gedaan moet worden om een voorspelling te maken. Vastgesteld is dat machine learning hierin een essentieel onderdeel is. Om een idee te krijgen welk algoritme het meest geschikt is, moet er een overzicht zijn van huidige standaarden, methoden en algoritmes die vandaag de dag efficiënt worden bevonden.

Om te beginnen is er een lijst opgesteld van de meest voorkomende en aangeraden machine learning algoritmes genoteerd. Deze samenstelling van algoritmes is op basis van verschillende artikelen op sites [14], [15], [16], [17], [18]. Dit is om te peilen welke algoritmes het meest aangeraden worden. Dit is om een idee te krijgen over wat de huidige standaarden zijn. Sites waar de algoritmes genoemd zijn, zijn opgenomen in de referenties. Deze tabel is te vinden in [Bijlage C, Tabel 7, p. 29-30].

Voordat er een keuze gemaakt kan worden welke algoritmes het meest geschikt zijn, moet er worden vastgesteld welke type algoritmes nodig zijn in het kader van de dit project. Voor de volgende vraagstukken moet een type algoritme worden gevonden:

- Herkenning van vissoort
- Voorspelling van o.a. vismigratie

Hieronder staan de algemene types van machine learning algoritmes.

THEORETISCHE INKADERING DATA

Er zijn verschillende typen algoritmes, die elk gespecialiseerd zijn in het oplossen van bepaalde vraagstukken. Nu moet er worden gekeken welke specialisatie het beste past bij het bepaalde vraagstuk van herkenning van vissoorten en vismigratie. De typen algoritmes worden hieronder in de volgende categorieën ingedeeld. Ook het specifieke vraagstuk is erbij vernoemd. In deze context wordt een niet nader omschreven collectie van data aangeboden.

Type	Vraagstuk	Functie
Classificatie	Is dit A of B?	Onderscheid maken en indelen in opgegeven categorieën (Supervised)
Regressie	Wat is de nieuwe waarde?	De gegeven numerieke waarden opnemen om een bepaalde eindwaarde te berekenen (Supervised)
Clustering	Wat is het patroon?	In een onbekend gestructureerde collectie data iteratief elementen indelen op basis van correlatie met elkaar (Unsupervised)
Reinforcement	Wat moet de volgende stap zijn?	In een bepaalde omgeving een positieve of gewenste beslissing kunnen maken. Het is wenselijk dat dit met een mate van zelfstandigheid wordt volbracht

Op basis van het bovenstaande schema kunnen voor de vismigratie vraagstukken een type algoritme worden toegewezen:

Vraagstuk	Type algoritme	Redenering
Herkenning van vissoort	Classificatie	Bij elke waarneming van een vis is het van belang dat er een vissoort aan wordt toegekend. Classificatie leent zich hier het beste bij.
Voorspelling vismigratie	Regressie	De waarschijnlijke vissoort of migratieroute kan worden berekenen aan de hand van historische gegevens.

		Ook andere voorspellingen zijn mogelijk. Geprojecteerde leefbaarheid voor een vissoort in bepaalde wateren kan worden aan de hand van historisch log van watereigenschappen.

Nu worden de algoritmes van [Tabel 7, p. 29-30] er bij genomen om deze in te delen op bovenstaande categorieën. Deze tabel geeft dan meer duidelijkheid over welke algoritmes te prioriteren.

	Classification	Regression	Clustering	Reinforcement	Rule-Based Association
Random Forest	x	x			
Decision Trees	x	x			
K-means			x		
Naïve Bayes	x				
Apriori					x
K-Nearest Neighbour	x	x			
Regression (Logistic)	x				
Support Vector Machine	x	x			
Regression (Linear)		x			
Neural Network (Convolutional)	x	x		x	

In het volgende hoofdstuk zal deze lijst verder worden gebruikt om nuttige algoritmes voor te trekken on ongeschikte algoritmes aan te duiden of helemaal buiten beschouwing te laten.

WELKE ALGORITMES ZIJN HET BESTE GESCHIKT VOOR HET VOORSPELLEN VAN VISMIGRATIE?

Nu zullen deze algoritmes van de vorige deelvraag een selectieproces doorlopen. Dit is om dieper in te gaan op specifieke voordelen en om minder geschikte algoritmes uit te sluiten om de uiteindelijke vismigratie te voorspellen.

Om meer onderscheiding tussen de algoritmes in te brengen, zullen de algoritmes worden geanalyseerd op basis van eigenschappen die nuttig zijn in het kader van dit project. Dit wordt samengevat in [Tabel 1]. Daar onder staat een korte toelichting van de eigenschappen waarop getoetst is.

	Simpliciteit model	Begrijpelijkheid werking	Overfitting resistentie	Snelheid	Noise	Veel data	Complexe datasets	Missende data	Totaal
Random Forest [19]	3	2	3	2	3	3	2	3	**21**
Decision Trees [19], [20]	2	2	2	2	3	3	2	3	**19**
K-means	3	2	3	3	2	3	1	1	**18**
Naïve Bayes [19]	3	2	3	3	2	2	1	2	**18**
Apriori	3	3	3	2	1	2	2	1	**17**
K-Nearest Neighbour [20]	2	2	3	2	2	3	1	2	**17**
Regression (Logistic) [19]	3	3	2	3	3	2	0	1	**17**
Support Vector Machine [19], [20]	1	1	3	2	3	2	2	3	**17**
Regression (Linear) [20]	3	2	2	3	2	2	0	1	**15**
Neural Network (Convolutional) [19], [20]	1	1	1	1	2	1	3	2	**12**

Tabel 1: Algoritmes geprioriteerd op nuttige eigenschappen

- Simpliciteit model: hoe gemakkelijk is het een model op te stellen?
- Begrijpelijkheid werking: Hoe simpel is het de werking van het algoritme in te zien en de output te interpreteren?
- Overfitting resistentie: Hoe robuust is het algoritme tegen overfitting?
- Snelheid: Hoe snel is het algaritme in gebruik?
- Prestatie bij…
    - Noise: Hoe goed is het algoritme bestand tegen onjuiste of afwijkende (onbelangrijke) data?
    - Veel data: Hoe efficiënt is het algoritme qua rekenkracht en potentie als er grote hoeveelheden data geïnterpreteerd (moeten) worden?
    - Complexe datasets: Hoe efficiënt gaat het algoritme om met hoog-dimensionale of niet-eenduidige data?
- Missende data: Hoe goed bestand is het algoritme bij incomplete of missende data? Dit moet een zo min mogelijk negatieve weerslag hebben op het eindresultaat.

VERVALLEN ALGORITMES

Nu er een duidelijker beeld is welke algoritmes beter geschikt zijn, is het gemakkelijker de minder nuttige algoritmes buiten beschouwing te laten. Deze vervallen algoritmes en de toelichting zijn te zien in [Bijlage C, Tabel 8, p. 30].

De lijst van resterende algoritmes is als volgt:

Random Forests	*Naïve Bayes*	*Linear Regression*	*Support Vector Machine*
Decision Trees	*K-Nearest Neighbor*	*Logistic Regression*	*Neural Network*

TESTEN ALGORITMES

De volgende stap is om de lijst van resterende algoritmes functioneel te testen, zodat er een algoritme kan worden gekozen en geïmplementeerd. Deze tests worden uitgevoerd op voorbeeld-dataset *Pima Indians Diabetes* onder andere beschikbaar in R. De uitvoering van deze tests zijn vastgelegd in [Bijlage D, p. 31]. Uit deze test bleek het Random Forest algoritme het beste te presteren. De volgende stap is om te testen met een echte dataset met visgegevens.

TESTEN DATASET

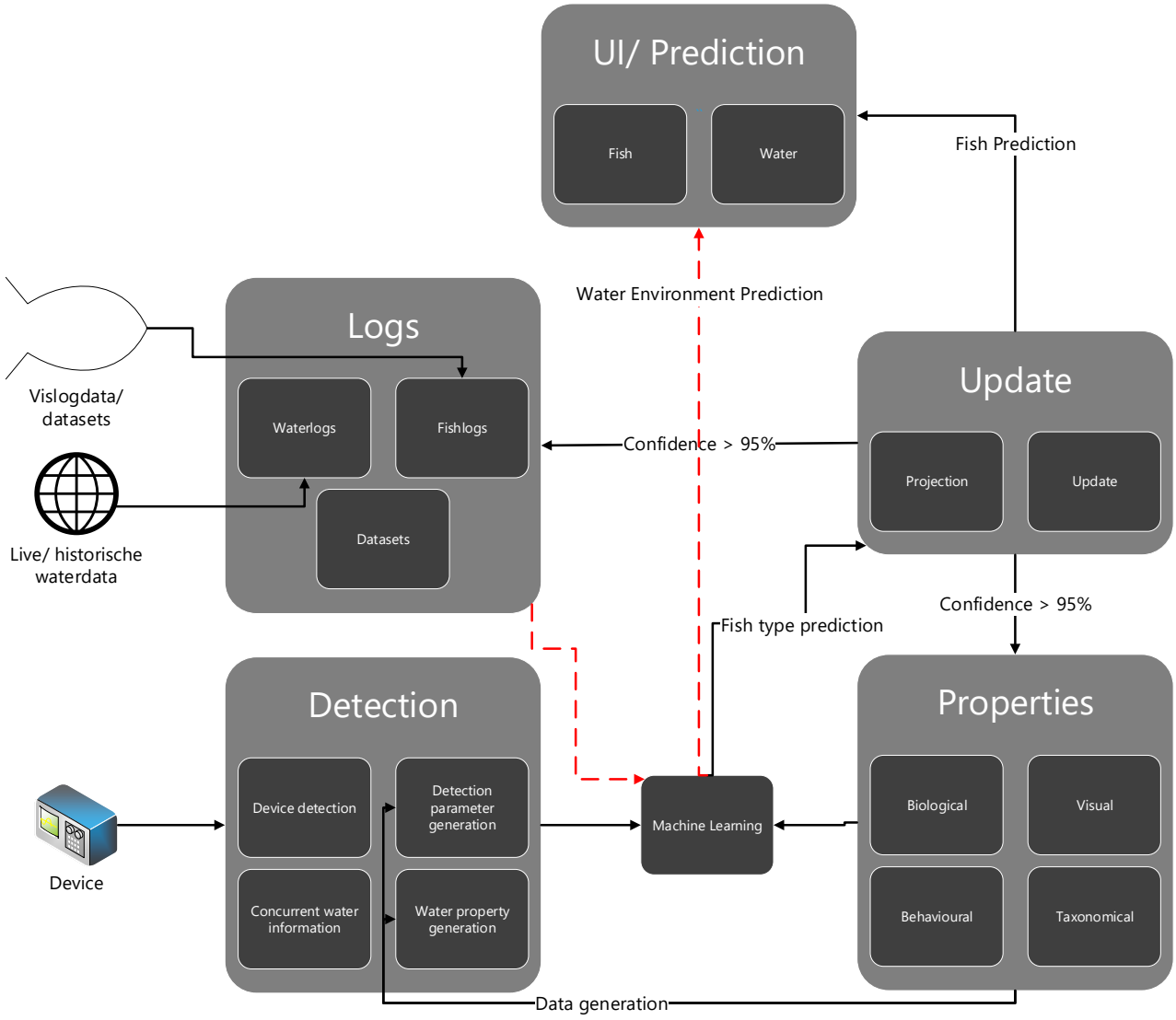
Nu er een algoritme verkozen is voor het classificeren voor vissoorten, moet er worden getest op de vis dataset. Dit is om een idee te krijgen welk algoritme het meest effectief is in een relevante dataset. De uitkomt van deze test weegt zwaarder dan die met de voorbeeld-dataset. De uitvoering van deze test is ook vastgelegd in [Bijlage D, Resultaten benchmark visdataset en Resultaten benchmark Visdataset 2, p. 38 - 39].

Hieruit kwam wederom de Random Forest uit als meest geschikte algoritme, gevolgd door de Support Vector Machine (SVM).

MODEL

Met de tot nu verzamelde gegevens van de drie deelvragen kan er een model worden gemaakt waarmee voorspelling gedaan kan worden. Zo zijn de gegevens die nuttig en noodzakelijk zijn opgesteld van vissen [Bijlage A, p. 24] en water [Bijlage B, p. 26]. Het model dat wordt geconstrueerd verbindt deze gegevens op een manier waar op lange termijn met grote influx van data kan worden gewerkt. Dit is een blauwdruk waarop gebouwd kan worden. Dit model is een antwoord op de hoofdvraag.

Dit model wordt basaal toegelicht, voor meer gedetailleerde velden wordt aangeraden het externe ontwerpdocument te raadplegen.



Figuur 3: Model visvoorspelling

Cluster	Bevat	Functie
Detection	• Detectie vis(gegevens) van device • Gelijktijdige mogelijkheid watereigenschap ten tijde van visdetectie • Ruimte genereren waarde in geval van testen en debuggen	Deze cluster is thans de 'trigger' waardoor het systeem een verwerking uitvoert, en kan zodoende beschouwd worden als de 'ingang'. Data wordt verstuurd naar een R-server om geïnterpreteerd te worden naar een vissoort
Properties	• Bevat viseigenschappen	Is in feite het informatieve referentiekader van het model. Bevat waarden en eigenschappen waarmee secundaire

	• Bevat visuele dataset voor visuele herkenning • Bevat gedragingen waaraan elementaire viseigenschappen ten grondslag liggen • Bevat taxonomische namen	eigenschappen of vraagstukken mee kunnen worden beantwoord. Bijvoorbeeld '*Hoezeer wijkt de zuurtegraad va dit water af tegen de geprefereerde of tolereerbare zuurtegraad van een paling? Hoezeer moet ik mijn water manipuleren om dit binnen tolereerbare grenzen te krijgen voor vissoort x?*'. De waarden kunnen ook geüpdatet worden zodra het algoritme zodanig vertrouwen heeft in een juiste classificatie. Wanneer een waarde van die geclassificeerde vis onjuist is of mist in deze cluster, wordt deze gemuteerd.

Logs	• Vastlegging van vissen:
	  ○ Juist geclassificeerde visdetecties
	  ○ Externe logs zoals van hobbyvissers of andere databronnen
	• Logs van watereigenschappen
	• Vis datasets
		Hier worden alle records in het model opgeslagen, mits betrouwbaar.
		Logs van watereigenschappen bevatten periodieke waarden van open data zoals Rijkswaterstaat.
		Vis datasets bestaan uit subsets van visdetectie in combinate met qua tijd en locatie gerelateerde watereigenschappen. Indien gewenst kan deze set gebruikt worden om het model nog beter te trainen.

Update	• Tijdelijke opslagplaats van veelal geclassificeerde detectiegegevens • Projectiewaarden op basis van verzamelde gegevens	Hier wordt elke afzonderlijke detectie vastgelegd in een tijdelijke database. Indien de confidence rate hoger is dan 0.95 (95%), dan kan dit record verplaatst worden naar de logs of properties. Tijdelijke projectiegegevens zijn hier ook mogelijk. Hier wordt naar trends gekeken en op basis hiervan naar de toekomst gecalculeerd.
UI/ Prediction	• Gedeelte waar gebruiker inzicht kan krijgen	Hier is het portaal waar een gebruiker of systeem inzicht kan krijgen in het model. Ook bepaalde queries kunnen worden geconstrueerd.

CONCLUSIE

In dit hoofdstuk worden de onderzoeksresultaten besproken en wordt er gekeken hoe er invulling aan de hoofd- en deelvragen is gegeven.

Hoe kan vismigratie worden voorspeld met machine learning?

De hoofdvraag wordt beantwoord door ook de deelvragen te beantwoorden. De deelvragen en conclusies:

WELKE DATA IS BELANGRIJK BIJ HET VOORSPELLEN EN RAPPORTEREN VAN VISMIGRATIE?

Om eerst een rapportage of een voorspelling te doen, is het van belang te onderzoeken welke informatie aan de basis hierover bestaat en nodig is. Hierbij zijn de volgende bronnen informatief gebleken:

Type data	Omschrijving	Bronnen	Uitwerking
Algemene viseigenschappen	Statische, vastgestelde biologische eigenschappen. Doel is om informatie te verschaffen om het toewijzen van vissen gemakkelijker verloopt.	[4], [8]	Bijlage A
Vislogdata	Logboeken of datasets van geïdentificeerde vissen waarbij plaats en tijdstip vermeld is.	[21]	Tabel 5
Waterdata	(Open) databronnen met historische en/of realtime gegevens over de staat van water in Nederland.	[13]	Bijlage B
Visuele visdata	Afbeeldingen of video's van vissen		

Deze data is noodzakelijk om een algoritme mee te kunnen trainen. Daarnaast is het van groot belang dat waarden zoveel mogelijk natuurgetrouw zijn en minimaal gesimuleerd.

WAT ZIJN DE HUIDIGE STANDAARDEN IN MACHINE LEARNING ALS HET GAAT OVER VOORSPELLING?

Deze deelvraag was vooral gericht op het onderzoeken van concept machine learning. Er zijn verschillende typen van machine learning, waarvan regressie en classificatie het meest nuttig lijken. Om een idee te krijgen welke algoritmes veel gebruikt of aangeraden worden, is er een lijst samengesteld van algoritmes op basis van mentions op verschillende websites die affiniteit hebben met machine learning [Tabel 7, p. 30].

WELKE ALGORITMES ZIJN HET BESTE GESCHIKT VOOR HET VOORSPELLEN VAN VISMIGRATIE?

Uit de verschillende test is gebleken dat het Random Forest algoritme het beste algoritme is om op voort te bouwen. Het Support Vector Machine algoritme kwam wel vaak in de buurt van de Random Forest. Deze heeft dus een tweede plaats verdiend.

HOE KAN VISMIGRATIE WORDEN VOORSPELD MET MACHINE LEARNING?

Aan de hand van het ontwerp [Figuur 3, p. 19] kan er een voorspelling worden gedaan van een mogelijk vissoort. Op basis van externe eigenschappen kunnen er secundaire voorspellingen kunnen gemaakt, waaronder hoe gunstig de leefomgeving kan zijn voor bepaalde vissen. Dit geeft dan een kans dat een vis langs een bepaald punt is gemigreerd. De betrouwbaarheid en veelzijdigheid van dit ontwerp hangt volledig af van de hoeveelheid en verscheidenheid van data.

DISCUSSIE

INNOVATIE

De innovatie van de oplossing zit hem vooral in de combinatie van machine learning, een relatief recent populair geworden tak, met de oude disciplines van visserij en waterwerken, welke al eeuwen aanwezig zijn in Nederland. Er zijn meer internationaal gerichte bewegingen die vispopulatie statistische berekenen, maar die zijn meer macro-georiënteerd en doen meer voorspellingen voor in de zee. Voor Nederland is er geen vergelijkbaar onderzoek gevonden waar met machine learning multidimensionaal informatie kan worden verschaft.

AFWIJKING RESULTAAT

Het uiteindelijke resultaat week wel af van de initiële planning. Vooral het concept voorspellen van vismigratie, die in de hoofdvraag is inbegrepen, is anders tot stand gekomen dan gedacht. Gezien de data waarmee gewerkt kon worden, en het niet was toegestaan intrusieve methodes te gebruiken om vissen te volgen, leek het al snel onmogelijk om te voorspellen waar een specifieke vis naar zou kunnen migreren. Het was hier dat de toenmalige secundaire methode, het met data onderbouwen van een specifiek gevolgde vis (primaire methode, voormalig), de primaire techniek werd waarmee voorspellingen gedaan moesten worden.

EXTERNE INVLOEDEN OP RESULTAAT

Daarnaast speelden gebrekkige kennis aan het begin over vissen en machine learning een rol in het niet volledig kunnen verdiepen in specifieke aspecten van vismigratie en machine learning. Langtijdig gebrek aan realistische vis dataset speelden ook parten in het experimenteren en analyseren van resultaten. De afhankelijkheid van de data werd onvoldoende gerealiseerd.

AANSLUITENDE LITERATUUR

Apart van de vermelde bibliografie waren er ook naslagwerken die nuttig werden bevonden bij het bestuderen van de hoofd- en deelvragen, en de daarbij aangrenzende aspecten. Overige aanbevelingsrichtingen zijn beschreven in het volgende hoofdstuk.

Onderwerp	Bron(nen)
Informatie over watertypen en bijbehorende vissen	https://www.sportvisserijnederland.nl/files/diepwater-coregonen_4599.pdf https://www.sportvisserijnederland.nl/files/viswater-baars-blankvoorn_4620.pdf https://www.sportvisserijnederland.nl/files/viswater-ruisvoorn-snoek_4623.pdf https://www.sportvisserijnederland.nl/files/viswater-blankvoorn-brasem_4621.pdf https://www.sportvisserijnederland.nl/files/viswater-brasem-snoekbaars_4622.pdf
Webpagina met groot aantal informatieve naslagwerken over vis en waterbeheer	https://www.sportvisserijnederland.nl/hsv-service/viswaterbeheer/

AANBEVELINGEN

Hoewel er getracht is een blauwdruk te maken in dit project, zijn er nog een aantal aandachtspunten. In dit hoofdstuk zal worden aangeduid wat de aangeraden richting is.

IMPLEMENTATIE

Aangeraden wordt om het geheel te implementeren op een Hadoop cluster. Deels was deze cluster al geconfigureerd, maar het geheel verwerken van het model in deze cluster zou de efficiënte en werkbaarheid van de hoofdvraag zeer ten goede komen. Het implementeren van een R distributie met parallel processing is ook aangeraden.

BEELDVERWERKING

Het aspect van beeldherkenning op vissen is niet volledig geëxploreerd, mede omdat het niet energiezuinig is. Er zou hier meer onderzoek gepleegd naar kunnen worden. Indien beeldverwerkingstechniek voor visherkenning doenbaar bevonden wordt, zou langs belangrijke waterroutes beeldverwerking kunnen worden toegepast. Energie en/ of computatie zou dan extern via een energiebron of een *edge device* gedaan kunnen worden.

EXPERTISE

Aangeraden wordt om een (marine) bioloog te betrekken bij dit project, incidenteel of stelselmatig. Dit is om visgedrag en overige kwantificeerbare, beschrijvende eigenschappen beter te kunnen aangeven en controleren. Bij het onderzoek naar vissen is gebleken dat de literatuur over marine fauna te diep reikend is om het te doorgronden.

Ook is het raadzaam een ontwerper te raadplegen, om onder andere user stories in acht te kunnen nemen. Deze kunnen dan gebruikt worden om bepaalde protocollen in het model aan te brengen, waarop gespitst kan worden op een specifieke informatiestroom. Ook het ontwerpen van een dashboard dat geïntegreerd is met het backend is voor een mogelijke eindgebruiker interessant.

DATA

Zoals al aangegeven werd bij de conclusie, komt meer data alleen maar ten goede van de nauwkeurigheid van het voorspellingsmodel. Daartoe zijn er een aantal aangeraden databronnen opgesteld. Het is mogelijk dat deze partijen de data niet kosteloos verschaffen.

- Sportvisserij Nederland [4]. Deze partij is een autoriteit als het gaat om kennis van vis, en hebben al vele informatieve publicaties uitgebracht, waarvan sommigen gebruikt werden voor dit document. De reden dat deze partij genoemd wordt is de mogelijkheid dat zij beschikken over visvangsten van vissers. Zodoende zouden er vislogs of visdatasets uit kunnen worden verhaald. Sinds vislogdata veel schaarser is dan waterdata, is elke gelegenheid om visdata te bemachtigen een goed idee. Site: www.sportvisserij.nl
- Visadvies.nl. Deze partij is gespecialiseerd in het aanleggen en monitoren van de visstand in Nederland. Ook doen zij onderzoek voor hun opdrachtgevers. Het is een mogelijkheid dat deze partij de devices kunnen verschaffen om detecties mee te kunnen maken. Omdat hun apparatuur betrouwbaar is, is de kans op betrouwbare data ook hoger. Site: www.visadvies.nl
- Vismetpiet.nl. Deze persoon is als een van de laatsten (of laatste) beroepsmatig werkzaam in Amsterdam als visser. Niet in eerste instantie relevant, maar deze persoon geeft excursies aan geïnteresseerden in Amsterdam. Wat hieraan relevant is, is de bewering dat hij alle vangstgegevens bijhoudt. Ook heeft hij kennis van vissen en is werkzaam in de buurt van Amstelveen. Contact met deze partij zou een goede eerste stap kunnen zijn voordat andere partijen gecontacteerd worden. Site: www.vismetpiet.nl

BIJLAGE A

Dit zijn alle velden die tot op heden beschikbaar zijn in de database van viseigenschappen. Voor elk van deze velden kan een (onderbouwde) waarden toegewezen. Elke vis waarop geclassificeerd kan worden, heeft een 'profiel' van alle hieronder gestelde velden.

Data label	Origine data	Omschrijving
Naam	Naam	Nederlandse benaming
Naam Latijn	Naam	Beschrijving in Latijn (Wetenschappelijke naam)
Naam Engel voorvoegsel	Naam	Beschrijvend voorvoegsel voor de Engelse benaming
Naam Engels	Naam	Engelse benaming
Naam Engels Volledig	Naam	Combinatie van bovenstaand voorvoegsel en Engelse benaming
Status	Status	Status van bedreiging
Stam	Naam	Taxonomische naam van stam (Engels: Phylum)
Klasse	Naam	Taxonomische naam van klasse (Engels: Class)
Orde	Naam	Taxonomische naam van orde (Engels: Order)
Familie	Naam	Taxonomische naam van familie (Engels: Family)
Geslacht	Naam	Taxonomische naam van geslacht (Engels: Genus)
Omgeving 1	Biologie	Omschrijving van gewenste leefomgeving
Omgeving 2	Biologie	Omschrijving van gewenste leefomgeving
Omgeving 3	Biologie	Omschrijving van gewenste leefomgeving
Omgeving 4	Biologie	Omschrijving van gewenste leefomgeving
Lmin	Biologie/ visueel	Minimumlengte (lage prioriteit)
Lmid	Biologie/ visueel	Gemiddelde lengte
Lmax	Biologie/ visueel	Maximale lengte
Gram	Biologie	Gewicht (lage prioriteit)
Diepte min	Gedrag	Gewoonlijke/ gewenste minimumdiepte
Diepte max	Gedrag	Gewoonlijke/ gewenste maximumdiepte
Diepte gewoonlijk min	Gedrag	Minimum leefdiepte (Absoluut)
Diepte gewoonlijk max	Gedrag	Maximum leefdiepte (Absoluut)
Temp min	Gedrag	Gewoonlijke minimumtemperatuur water
Temp mid	Gedrag	Gewoonlijke gemiddelde temperatuur water
Temp max	Gedrag	Gewoonlijke maximumtemperatuur water
Lowest temp	Biologie	Laagst te tolereren watertemperatuur
Highest temp	Biologie	Hoogst te tolereren watertemperatuur
Salinity min (psu)	Biologie	Minimum tolerantie zoutgehalte
Salinity mid (psu)	Biologie	Gemiddelde tolerantie zoutgehalte
Salinity max (psu)	Biologie	Maximum tolerantie zoutgehalte water
pH min	Biologie	Minimum tolerantie zuurtegraad water
pH max	Biologie	Maximum tolerantie zuurtegraad water
dH min	Biologie	Minimum tolerantie hardheid water
dH max	Biologie	Maximum tolerantie hardheid water
Paaitijd min	Biologie/ gedrag	Begin paaitijd (maand)
Paaitijd max	Biologie/ gedrag	Eind paaitijd (maand)
Paaitemp min	Biologie/ gedrag	Minimumtemperatuur gunstig voor paaitijd
Paaitemp max	Biologie/ gedrag	Maximumtemperatuur gunstig voor paaitijd
Voedsel1	Biologie/ gedrag	Soort voedsel
Voedsel2	Biologie/ gedrag	Soort voedsel
Voedsel3	Biologie/ gedrag	Soort voedsel
Dorsal spinesmin	Visueel	Minimum aantal ruggenwervels
Dorsal spinesmax	Visueel	Maximaal aantal ruggenwervels
Dorsal soft raysmin	Visueel	Minimum aantal zachte rugstralen
Dorsal soft raysmax	Visueel	Maximum aantal zachte rugstralen
Anal spinesmin	Visueel	Minimum aantal anale wervels
Anal spinesmax	Visueel	Maximum aantal anale wervels
Anal soft raysmin	Visueel	Minimum aantal zachte anale stralen
Anal soft raysmax	Visueel	Maximum aantal zachte anale stralen

Dorsal finmin	Visueel	Minimum aantal rugvinnen
Dorsal finmax	Visueel	Maximum aantal rugvinnen
Second dorsal finmin	Visueel	Minimum aantal tweede rugvinnen
Second dorsal finmax	Visueel	Maximum aantal tweede rugvinnen
Caudal finmin	Visueel	Minimum aantal staartvinnen
Caudal finmax	Visueel	Maximum aantal staartvinnen
Anal finmin	Visueel	Minimum aantal anaalvinnen
Anal finmax	Visueel	Maximum aantal anaalvinnen
Pelvic finmin	Visueel	Minimum aantal vinnen van bekken
Pelvic finmax	Visueel	Maximum aantal vinnen van bekken
Pectoral finmin	Visueel	Minimum aantal borstvinnen
Pectoral finmax	Visueel	Maximum aantal borstvinnen
Vertebrae min	Visueel	Minimum aantal wervels
Vertebrae max	Visueel	Maximum aantal wervels

Tabel 2: Velden database viseigenschappen

BIJLAGE B

Dit zijn de waterparameters die op het heden in het ontwerp opgenomen zijn. Hiernaast zijn de bronnen die informatie tot deze parameters kunnen verschaffen. De hardwarematige bronnen zijn zelf beschikbaar om aan te komen, en de databron kolom rechts is afkomstig van het open water dataplatform van Rijkswaterstaat [13].

Parameter	Hardware [12]	Databron [13]
Stroomsnelheid		Rijkswaterstaat
Waterhoogte t.o.v. NAP		Rijkswaterstaat
Stroomrichting		Rijkswaterstaat
Temperatuur	Libelium Smart Water Model Libelium Smart Water Ions Single Libelium Smart Water Ions Double Libelium Smart Water Ions Pro	Rijkswaterstaat
Zoutgehalte		Rijkswaterstaat
Geleidbaarheid	Libelium Smart Water Model	
Turbiditeit (Troebelheid)	Libelium Smart Water Model	
Zuurtegraad (pH)	Libelium Smart Water Model Libelium Smart Water Ions Single Libelium Smart Water Ions Pro	
Opgeloste zuurstof	Libelium Smart Water Model	
Oxidatiereductie potentie	Libelium Smart Water Model	
Calciumionen (Ca^{2+})	Libelium Smart Water Ions Single	
Fluorideionen (F^-)	Libelium Smart Water Ions Single Libelium Smart Water Ions Pro	
Fluoroboraationen ($BF4^-$)	Libelium Smart Water Ions Single	
Nitraationen (NO_3^-)	Libelium Smart Water Ions Single	
Bromideionen (Br^-)	Libelium Smart Water Ions Double Libelium Smart Water Ions Pro	
Chlorideionen (Cl^-)	Libelium Smart Water Ions Double Libelium Smart Water Ions Pro	
Koperionen (Cu^{2+})	Libelium Smart Water Ions Double Libelium Smart Water Ions Pro	
Iodide ionen (I^-)	Libelium Smart Water Ions Double Libelium Smart Water Ions Pro	
Zilverionen (Ag^+)	Libelium Smart Water Ions Double Libelium Smart Water Ions Pro	
Ammoniumionen (NH_4^+)	Libelium Smart Water Ions Pro	
Calciumionen (Ca^{2+})	Libelium Smart Water Ions Pro	
Lithiumionen (Li^+)	Libelium Smart Water Ions Pro	
Magnesiumionen (Mg^{2+})	Libelium Smart Water Ions Pro	
Nitraationen (NO_3^-)	Libelium Smart Water Ions Pro	
Nitrietionen (NO_2^-)	Libelium Smart Water Ions Pro	
Perchloraationen ($CIO_4^{-)}$	Libelium Smart Water Ions Pro	
Kaliumionen (K^+)	Libelium Smart Water Ions Pro	
Sodiumionen (Na^+)	Libelium Smart Water Ions Pro	
Stroomsnelheid m/s		Rijkswaterstaat (Expert\)

(massa)Concentratie cyanide in Oppervlaktewater ug/l		Rijkswaterstaat (Expert \ Algemene waterkwaliteit)
(massa)Concentratie zuurstof in Oppervlaktewater mg/l		Rijkswaterstaat (Expert \ Algemene waterkwaliteit)
Verzadigingsgraad zuurstof in Oppervlaktewater %		Rijkswaterstaat (Expert \ Algemene waterkwaliteit)
Zuurgraad Oppervlaktewater		Rijkswaterstaat (Expert \ Algemene waterkwaliteit)
(massa)Concentratie chloride in Oppervlaktewater mg/l		Rijkswaterstaat (Expert \ Zouten)
Geleidendheid Oppervlaktewater S/m		Rijkswaterstaat (Expert \ Zouten)
Saliniteit Oppervlaktewater		Rijkswaterstaat (Expert \ Zouten)
(massa)Concentratie ammonium in Oppervlaktewater uitgedrukt in stikstof / opgeloste fractie in mg/l		Rijkswaterstaat (Expert \ Nutriënten en eutrofiëringsparameters)
(massa)Concentratie Biochemisch zuurstofverbruik met allylthioureum in Oppervlaktewater mg/l		Rijkswaterstaat (Expert \ Nutriënten en eutrofiëringsparameters)
(massa)Concentratie Chemisch zuurstofverbruik in Oppervlaktewater mg/l		Rijkswaterstaat (Expert \ Nutriënten en eutrofiëringsparameters)
(massa)Concentratie chlorofyl-a in Oppervlaktewater ug/l		Rijkswaterstaat (Expert \ Nutriënten en eutrofiëringsparameters)
(massa)Concentratie fosfaat in Oppervlaktewater uitgedrukt in fosfor / opgeloste fractie in mg/l		Rijkswaterstaat (Expert \ Nutriënten en eutrofiëringsparameters)
(massa)Concentratie fosfor totaal in Oppervlaktewater particulair gebonden in mg/l		Rijkswaterstaat (Expert \ Nutriënten en eutrofiëringsparameters)
(massa)Concentratie fosfor totaal in Oppervlaktewater uitgedrukt in fosfor / opgeloste fractie in mg/l		Rijkswaterstaat (Expert \ Nutriënten en eutrofiëringsparameters)
(massa)Concentratie nitraat in Oppervlaktewater uitgedrukt in stikstof / opgeloste fractie in mg/l		Rijkswaterstaat (Expert \ Nutriënten en eutrofiëringsparameters)
(massa)Concentratie nitriet in Oppervlaktewater uitgedrukt in stikstof / opgeloste fractie in mg/l		Rijkswaterstaat (Expert \ Nutriënten en eutrofiëringsparameters)
(massa)Concentratie stikstof totaal in Oppervlaktewater uitgedrukt in Stikstof in mg/l		Rijkswaterstaat (Expert \ Nutriënten en eutrofiëringsparameters)
Waterhoogte Oppervlaktewater t.o.v. Mean Sea Leven in cm		Rijkswaterstaat (Expert \ Waterhoogten)

Tabel 3: Mogelijke velden water parameters

BIJLAGE C

In deze bijlage zijn overige tabellen die door plaatsgebrek of een te indirect verband met de hoofdvraag opgenomen.

De volgende parameters kaderen het fenomeen eutrofiëring in. Ook wordt ernaast genoteerd welke parameter afkomstig van Rijkswaterstaat de vraag tegemoetkomt [22].

Verschijnsel	Parameter	Bron
Nutriënten	Fosfor	- Concentratie fosfor totaal in oppervlaktewater particulair gebonden in mg/l - Concentratie fosfor totaal in oppervlaktewater uitgedrukt in fosfor/ opgeloste fractie in mg/l - Concentratie fosfor totaal in oppervlaktewater in fosfor/ opgeloste fractie in mg/l
	Stikstof	- Concentratie som nitraat en nitriet in oppervlaktewater uitgedrukt in stofstof/ opgeloste fractie in mg/l
Stikstof subelementen	Nitraat	- Concentratie nitraat in oppervlaktewater uitgedrukt in stikstof/ opgeloste fractie in mg/l
	Nitriet	- Concentratie nitriet in oppervlaktewater uitgedrukt in stikstof/ opgeloste fractie in mg/l
	Ammonium	- Concentratie ammonium in oppervlaktewater uitgedrukt in stikstof/ opgeloste fractie in mg/l
Gevolg algengroei	(Afname) zuurstof in water	- Concentratie zuurstof in oppervlaktewater mg/l - Concentratie biochemisch zuurstofverbruik met allylthioureum in oppervlaktewater mg/l - Concentratie chemisch zuurstofverbruik in oppervlaktewater mg/l - Concentratie chemisch zuurstofverbruik in oppervlaktewater opgeloste fracit (bijv. na filtratie) in mg/l
	Zuurtegraad	- Zuurgraad oppervlaktewater

Tabel 4: Parameters en bronnen eurofiëring

Veld	Deel 1	Deel 2	Toelichting
Genus	X	X	Taxonomische benoeming geslacht. Een orde hoger dan soort.
Species	X	X	Taxonomische benoeming soort.
Center Lat	X	X	Center Latitude; centrale breedtegraad
Center Long	X	X	Center Longitude; centrale lengtegraad
C-Square Code	X	X	Verkorting voor *Concise Spatial Query and Representation System*. Een bepaalde indexering van globale plots waarin lengte- en breedtegraad verwerkt zijn.
Overall Probability	X		Kans op rechtmatige melding van een vis gebaseerd gunstige omgevingsvariabelen [21]
Depth	X	X	Diepte onder de zeespiegel
Temperature	X		Temperatuur van water
Salinity	X	X	Mate van zoutheid van het water, gemeten in psu (Practical Salinity Unit). 1 psu staat voor 1 gram zout per kilogram.

Primary Production		De mate van organische koolstofproductie door autotrofen, specifiek foto-autrofen. Een foto-autrofe levensvorm gebruikt licht (fotosynthese) om zichzelf te voeden en om organische koolstof uit te scheiden.
Sea Ice Concentration.		IJsconcentratie in het zeewater
Distance to Land		Afstand van de kust
SST		Sea Surface Temperate; temperatuur van de oppervlakte op de locatie waar de vis is waargenomen, op de zee.

Tabel 5: Beschrijving velden datasets afkomstig van Aquamaps.org **[21]**

Veld	Actie	Toelichting
Genus	Verwijderd	De soort (species) zegt iets specifieker dan het geslacht (genus). Vissen kunnen dezelfde geslachtsbenaming hebben maar toch van een andere soort zijn. Daarnaast is deze informatie redundant, omdat een enkele naam even veel zegt over de soort. Vooral omdat er in dit geval 3 vissoorten zijn. Ook is het voor het model simpeler om te interpreteren en trainen als er één beschrijving is.
Species	Mutatie	De soort van de vis. Dit is het belangrijkste veld qua machine learning. Er moet op soort worden geclassificeerd. De soorten zijn in deze de *labels*. In R moet het type van dit veld van `character` naar `factor` worden geconverteerd, zodat de algoritmes deze labels kunnen interpreteren.
Center Lat	Verwijderd	De kale lengtegraad en breedtegraad zijn deze staat niet van belang. Tenzij de juiste context aan deze coördinaten wordt gegeven, zullen deze waarden het model alleen verwarren. Met de juiste context wordt bedoeld het kunnen interpreteren als een locatie op een kaart, waar rekening wordt gehouden met geologische liggingen van zee, land en bijbehorende factoren.
Center Long	Verwijderd	Zie Center Lat
C-Square Code	Verwijderd	Hiervoor geld in principe hetzelfde als voor de lengte- en breedtegraad. Om deze informatie te vertalen is nog een vertaalslag nodig om deze code in lengte- en breedtegraad te converteren.
Depth	Gehandhaafd	Relevant genoeg bevonden. Deze eigenschap zegt iets over vissoorten zelf, sinds het algemeen bekend is dat bepaalde vissen op bepaald dieptes in de oceaan leven. Daarnaast zijn deze gegevens ook opgenomen in de database van viseigenschappen.
Salinity	Gehandhaafd	Ook relevant genoeg bevonden. Hoewel het zoutgehalte uiteraard hoger is dan zoet water, zegt dit ook iets specifieks over de vissoort. Ook deze eigenschap is opgenomen in de database van viseigenschappen.
Primary Production	Verwijderd	Was eerst inbegrepen, maar bleek niet specifiek genoeg om de nauwkeurigheid van het model te verhogen. In sommige gevallen was deze zelfs verlaagd.
Sea Ice Concentration.	Verwijderd	Buiten beschouwing gelaten. Deze data is onder de 3 vissen niet goed gerepresenteerd, en daarnaast zijn de waarden summier.
Distance to Land	Gehandhaafd	In de dataset gehouden. Dit kan iets over de vissen vertellen.
SST	Gehandhaafd	*Sea surface temperature*. Ook is deze eigenschap opgenomen in de visdatabase, waardoor dit veld zich uitleent voor gebruik.

Tabel 6: Veranderingen originele dataset

Algoritme	Variant	Type	[14]	[15]	[16]	[17]	[18]	Totaal
Decision Trees	Linear	Supervised	1	1	1	1	1	**5**

Decision Trees	Logistic	Supervised	1	1	1	1	1	**5**
Naïve Bayes		Supervised	1	1	1	1	1	**5**
Random Forest		Unsupervised	1	1	1	1	1	**5**
Regression	Linear	Supervised	1	1	1	1	1	**5**
Regression	Logistic	Supervised	1	1	1	1	1	**5**
K-means		Unsupervised	1	1		1	1	**4**
K-Nearest Neighbour		Unsupervised	1	1	1		1	**4**
Support Vector Machine		Supervised	1	1	1	1		**4**
Adaboost		Metaheuristic	1		1		1	**3**
Neural Network	Convolutional	Supervised	1	1		1		**3**
Apriori		Unsupervised		1			1	**2**
Principal Component Analysis		Unsupervised				1	1	**2**
Independent Component Analysis		Unsupervised				1		**1**
Neural Network	Learning Vector Quantization	Unsupervised			1			**1**
Linear Discriminant Analysis		Unsupervised			1			**1**
Markov		Supervised	1					**1**
Neural Network	Recurrent	Supervised	1					**1**
Singular Vasue Decomposition		Supervised				1		**1**

Tabel 7: Populariteitstabel machine learning algoritmes

Vervallen algoritme	Toelichting
Adaboost	Sinds Adaboost geen losstaand algoritme is maar een booster algoritme, is deze uit deze lijst gefilterd, ondanks de populaire positie. Om deze reden was AdaBoost al uit de vergelijking in deze deelvraag verwijderd.
K-means	Bij clustering algoritmes zijn labels niet van tevoren bekend, zoals gewoonlijk met Unsupervised algoritmes. Het voordeel van clustering algoritmes is om juist in onbekende, gegeneraliseerde data structuur en patronen te herkennen. Dit is in het kader van vismetingen niet aan de orde, sinds elke variabele al onder een noemer wordt opgeslagen. Clustering algoritmes op de lijst worden vanaf dit punt niet meer meegenomen.
Apriori	Het functionele doel van dit algoritme valt buiten de scope van dit project. Het voordeel van dit algoritme is het minen van data, waarbij de hoeveelheid data niet bekend is. Dit is niet van toepassing in dit project. De argumentatie tegen dit algoritme is ook gedeeltelijk vergelijkbaar met clustering algoritmes.

Tabel 8: Vervallen algoritmes

BIJLAGE D

In deze bijlage worden verschillende algoritmes getest op verschillende datasets, om uiteindelijk een meest geschikte kandidaat te vinden. De eerste 2 testen (classificatie en regressie) worden uitgevoerd op een voorbeelddataset. Deze test werd gedaan om een praktisch inzicht te krijgen over de functionaliteit van de algoritmes, alvorens in een later teststadium deze algoritmes op relevante datasets te gebruiken.

Na deze 2 tests worden de algoritmes getest op een relevante dataset met visgegevens. De resultaten van deze test wegen vanzelfsprekend zwaarder dan de tests met de voorbeeld trainingsset. Uiteindelijk zal een algoritme functioneel het beste uit deze test komen. Dit algoritme is dan nog niet verkozen als algoritme waarop in het vismigratieproject wordt verder gebouwd. De uitslag van de test wordt simpelweg als een argument voor een algoritme gebruikt, naast een theoretisch argument onderbouwd in deelvraag 3: [Welke algoritmes zijn het beste geschikt voor het voorspellen van vismigratie?].

Indien deze argumenten verschillende algoritmes aanwijzen, zal tussen deze algoritmes een laatste test worden gedaan om het beste algoritme te verkiezen.

METHODIEK

Er wordt gebruik gemaakt van R libraries `mlbench` en `caret`. De methodiek van deze test is gebaseerd op een artikel van Jason Brownlee [23].

Er zal een statistische methodiek worden toegepast, genaamd *k-fold cross validation*, om de algoritmes te vergelijken. K-fold it houdt in dat een dataset in *k* delen wordt gesplitst. De delen bevatten een willekeurige selectie van de observaties van de dataset. Hierna wordt (*k - 1*) van de delen gebruikt om een model te trainen in een algortime, en het resterende deel wordt gebruikt om het model te testen. Dit wordt ook *k* keer herhaald (*k* folds), zodat elk deel een keer heeft gefungeerd als test data. Hierbij wordt de dataset maximaal benut in voorspellingspotentie. Ook zal een afzonderlijk k-fold proces 3 keer herhaald worden (3 *repeats*) waarbij steeds nieuwe willekeurige observaties verkozen zullen worden uit de dataset. Dit is om meer van elkaar onafhankelijke vergelijkingen te kunnen maken, waardoor de uitslag beter de ware nauwkeurigheid van een algoritme weergeeft.

RESULTATEN BENCHMARK VOORBEELDSET

De algoritmes zijn ingedeeld in 2 categorieën: regressie algoritmes en classificatie algoritmes. De meeste algoritmes komen in beide lijsten voor, maar worden functioneel op 2 vlakken getest (regressie en classificatie). De resultaten van de groepen verschillen en worden na de resultaat output toegelicht.

De uitleg van specifieke visuele weergaven is te lezen onder de Regressie kop, en niet bij de Classificatie kop, om redundantie te voorkomen. Wel wordt inhoudelijke informatie bij beiden koppen verschaft.

REGRESSIE

De volgende algoritmes zijn in R getest en vergeleken:

Random Forest	K-Nearest Neighbor	Linear Regression
Decision Trees	Support Vector Machine	Neural Network

SAMENVATTING

Hier volgen de resultaten van de test. De volgende samenvatting is in R tot stand gebracht met behulp van functie `Summary()`:

```
Call:
summary.resamples(object = reg.results)
```

```
Models: Rforest, Dtrees, KNN, SVM, RegLin, ANN
Number of resamples: 30

MAE
             Min.    1st Qu.     Median       Mean   3rd Qu.       Max. NA's
Rforest  5.297556  6.016100  6.318607  6.438938  6.757748  8.121544    0
Dtrees   6.222424  6.600610  7.018225  7.099034  7.495868  9.304296    0
KNN      6.950938  7.678239  7.921119  7.970069  8.184968  9.044733    0
SVM      4.708368  5.588907  5.954789  6.095354  6.453801  7.376463    0
RegLin   5.436650  6.444356  6.759726  6.803623  7.086321  8.200796    0
ANN     31.480519 32.032638 32.274522 32.241571 32.419686 32.946667    0

RMSE
             Min.    1st Qu.     Median       Mean   3rd Qu.      Max. NA's
Rforest  7.156189  8.099071  8.925807  8.885365  9.508734 10.92750    0
Dtrees   8.424398  9.093154  9.788228  9.828406 10.488783 12.17539    0
KNN      9.306298 10.096617 10.404766 10.544079 11.041364 12.03386    0
SVM      6.905333  8.300425  9.030034  9.059999  9.790336 10.81960    0
RegLin   7.527936  8.743679  9.593957  9.387384  9.846477 10.96112    0
ANN     33.282342 34.048462 34.427128 34.313811 34.646168 35.13915    0

Rsquared
               Min.    1st Qu.     Median       Mean   3rd Qu.       Max. NA's
Rforest 0.25058917 0.3834580 0.4376232 0.4359167 0.5057891 0.6311506    0
Dtrees  0.10810554 0.2669836 0.3151510 0.3102039 0.3756614 0.4863491    0
KNN     0.09267626 0.1600122 0.2152666 0.2087659 0.2506085 0.3413979    0
SVM     0.22768197 0.3763434 0.4320084 0.4282497 0.4929996 0.5807106    0
RegLin  0.19948550 0.3086912 0.3597976 0.3692375 0.4291044 0.5960603    0
ANN             NA         NA         NA       NaN         NA         NA   30
```

INTERPRETATIE GEMIDDELDEN

In deze waarden kan worden achterhaald hoe nauwkeurig elk algoritme te werk gaat op eenzelfde dataset. Hoewel er 3 testfuncties (MAE, RMSE en Rsquared, toegelicht in [*Toelichting regressie* meetmethode]) zijn gebruikt, geld voor de cijfers van MAE en RMSE dat een lage waarde beter is. Bij Rsquared is een hogere waarde positief.

De numerieke waarden zijn gemiddelden van de tests. Hierin in de Min het laagst genomen error rate uit de k-fold sessie. Daarop aansluiten ook de 1st Quarter, Median (middelste cijfer) en 3rd Quarter en ten slotte max. Verder is de Mean het gemiddelde van alle k-fold gemiddelden.

Van belang zijn vooral het gemiddelde en de het minimum. Het gemiddelde zegt wat over de betrouwbaarheid, vooral op langere termijn. Het minimum zegt iets over de potentie van een algoritme, in het geval van MAE en RMSE. Dit geeft weer hoe dichtbij elk algoritme geweest bij de laagste error rate. De relatie tussen het gemiddelde en het minimum zegt ook iets over de spreiding. Een positief minimum is gunstig maar als deze ver afstaat van het gemiddelde betekent dit dat het algoritme bijvoorbeeld 1 op de 10 keer heel dichtbij zit. Hetzelfde geld ook voor Rsquared, alleen is hier het maximum de beste poging van een algoritme in plaats van het minimum.

TOELICHTING REGRESSIE MEETMETHODES

Error rate meetmethode	Toelichting

MAE	Mean Absolute Error. Het gemiddelde van de absolute waarden van de error rate tussen de gemodelleerde waarde ($\hat{y}$) en de werkelijke waarde (y). Dit is de volgende formule: $MAE = \frac{\sum_{t}^{n}=1\|\hat{y}_t - y_t\|}{n}$
RMSE	Root Mean Squared Error. Het gemiddelde van een opsomming van de afstand tussen de gemodelleerde waarde en de werkelijke waarde. Hierover wordt de wortel berekend. $RMSE = \sqrt{\frac{\sum_{t}^{n}=1(\hat{y}_t - y_t)^2}{n}}$
Rsquared	Het gemiddelde van een regressie iteratiecijfer gedeeld door de totale error opsomming

INTERPRETATIE SAMENVATTING

De volgende kernachtige informatie kan uit de samenvatting worden geconcludeerd:

Methodiek	Beste gemiddelde	Beste uiterste
MAE	SVM	Random Forest (Minimum)
RMSE	SVM	Random Forest (Minimum)
Rsquared	Random Forest	Random Forest (Maximum)

Op te merken is dat SVM uit twee error test methodieken gemiddeld het hoogst scoort. Tegelijkertijd is het de Random Forest die eenmalig heeft gemaakt. Hieruit lijken de SVM en RF algoritme een gedeelde eerste plaats te hebben.

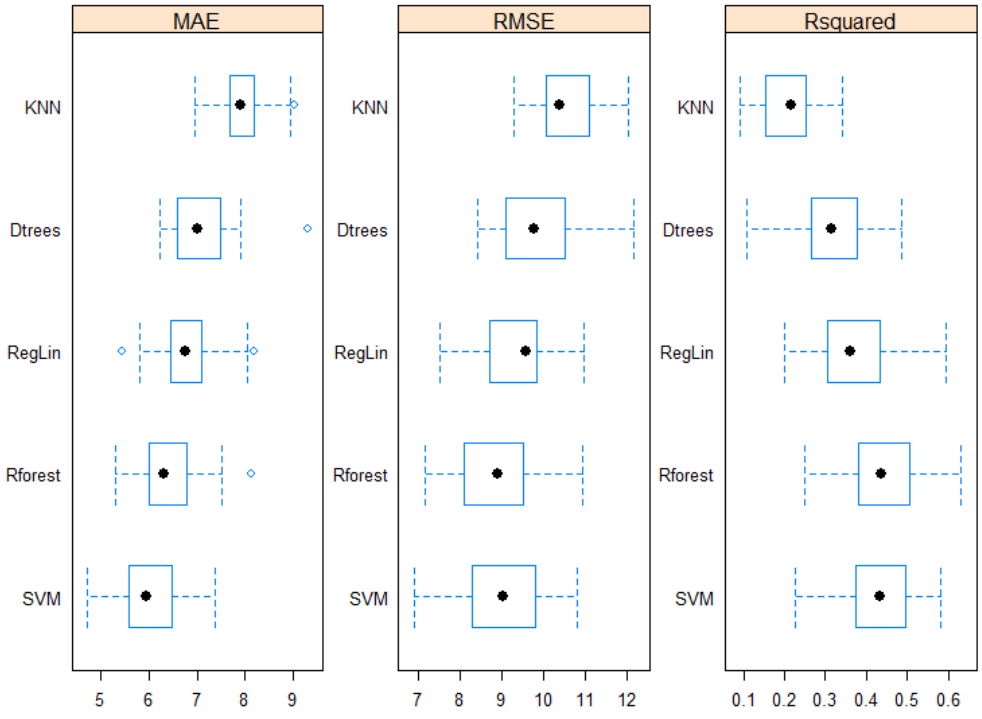
De volgende afbeeldingen geven bovenstaande resultaten visueel weer. Deze zijn gemaakt om de uitkomsten met elkaar te vergelijken en visueel weer te geven.

Wel moet opgemerkt worden dat het Neural Network algoritme niet inbegrepen is in de volgende afbeeldingen. Dit is omdat deze zodanig afweek dat dit een negatieve weerslag had op de leesbaarheid van andere algoritmeresultaten. Dit geeft ook aan dat het Neural Network onvoldoende effectief is om op verder te bouwen.

BOX AND WHISKERS TABEL

De volgende afbeelding is een zogenoemd *Box and Whisker* weergavetabel van de resultaten. Deze functie in te gebruiken in `R` onder package `lattice` als `bwplot()`. Deze tabel8heeft een automatische functie op de meest geschikte resultaten (beste gemiddelden) bovenaan te plaatsen.

Per error categorie (zoals bij de tekstuele resultaten) zijn de error rates van de geteste algoritmes



Figuur 4: Box and Whiskers resultaat regressie op voorbeeldset

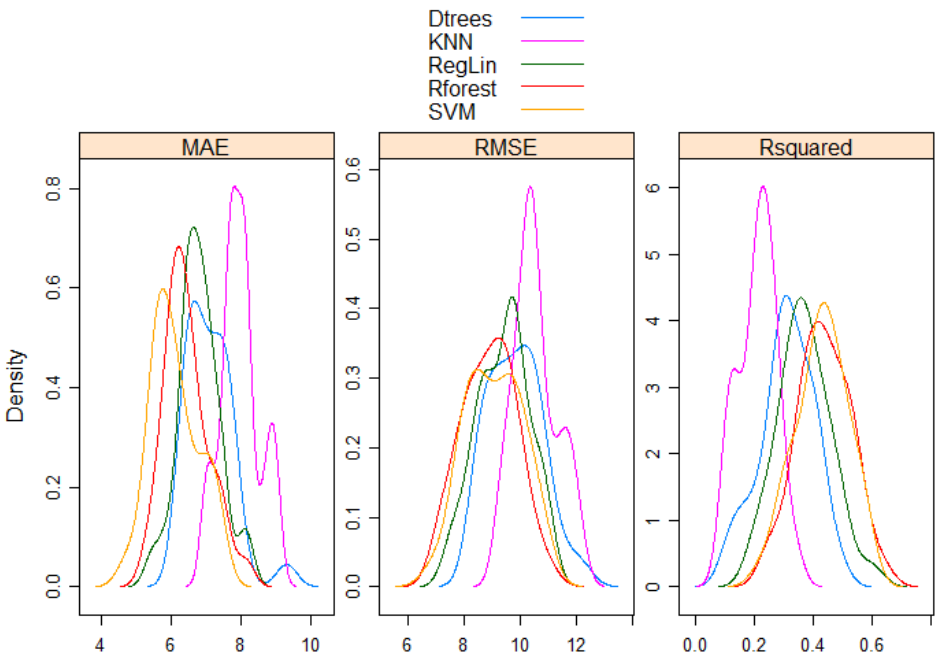
weergegeven. De stippellijn met begin en eindpunt representeren het minimum en maximum van de gemiddelden. De uitersten van de vierkanten (boxes) zijn het ¼ van het gemiddelde en de ¾ gemiddelde van de error rate, zoals bij bovenstaande tekstuele resultaat. De stip is het gemiddelde. Hierdoor is ook de spreiding goed te zien.

DICHTHEIDSVERDELING

Hierna volgt de dichtheid verdeling van de gemiddelden van de algoritmes. Deze functie in `R` kan worden aangeroepen met `densityplot`, en maakt deel uit van de `lattice` library.

De dichtheid is gerepresenteerd op de Y-as en de resultaten van de error rates op de X-as.

Hier kan informatie worden gehaald uit onder andere de pieken en de spreiding van de lijn. Hierin is de spreiding de minimum en maximum error rate van de algoritmes. Algoritmes die een grotere spreiding hebben, hebben vaker een lagere piek. Dit is omdat bij een kleinere spreiding de gemiddelden minder van elkaar afwijken (lagere variantie) en dus juist hoger pieken.



Figuur 5: Dichtheidsverdeling gemiddelden resultaat regressie op voorbeeldset

Te zien is dat het K-nearest neighbor (magenta) consistent de kleinste spreidingsfactor heeft qua error rates. Bevestigd wordt dat dit algoritme ook slechts scoort op MAE (laatste piek) maar het best bij Rsquared (eerste piek)

CLASSIFICATION

Hier worden classificatie algoritmes op de voorbeelddataset toegepast. De meeste algoritmes hier genoteerd zijn in staat zowel regressie als classificatie toe te kunnen passen; zodoende worden de meeste algoritmes hier hergebruikt. Alleen Naive Bayes en Logistic Regression zijn pure classificatie-algoritmes.

De volgende algoritmes zijn met classification getest:

Random Forest	K-Nearest Neigbor	Artificial Neural Network
Decision Trees	Logistic Regression	
Naïve Bayes	Support Vector Machine	

Nu volgende de resultaten van deze algoritmes. Genoteerd moet worden dat bij classificatie wordt er gekeken naar *Accuracy* (nauwkeurigheid) en *Kappa*. In tegenstelling tot regressie is een hogere waarde positiever dan een lagere waarde. Deze termen worden verder toegelicht onder [Toelichting classificatie meetmethodes].

```
Call:
summary.resamples(object = class.results)

Models: Rforest, Dtrees, NBayes, KNN, LogReg, SVM, ANN
Number of resamples: 30

Accuracy
             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Rforest 0.6883117 0.7402597 0.7712748 0.7691957 0.7922078 0.8571429    0
Dtrees  0.6623377 0.7272727 0.7532468 0.7465596 0.7662338 0.8289474    0
NBayes  0.6753247 0.7175325 0.7532468 0.7552005 0.7869105 0.8441558    0
KNN     0.6493506 0.7142857 0.7467532 0.7400832 0.7662338 0.8421053    0
LogReg  0.6883117 0.7272727 0.7646958 0.7543347 0.7784945 0.8961039    0
SVM     0.6973684 0.7305195 0.7662338 0.7665243 0.7922078 0.8441558    0
ANN     0.6623377 0.7142857 0.7532468 0.7495899 0.7869105 0.8311688    0

Kappa
             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Rforest 0.3036925 0.4120678 0.4732597 0.4765512 0.5494365 0.6602487    0
Dtrees  0.1894737 0.3495477 0.4340426 0.4184446 0.4827744 0.6302395    0
NBayes  0.2655008 0.3591963 0.4487152 0.4435648 0.5088825 0.6457055    0
KNN     0.2098062 0.3388961 0.4058531 0.4047018 0.4733369 0.6492308    0
LogReg  0.2786885 0.3543217 0.4335421 0.4284613 0.4854794 0.7638037    0
SVM     0.2517123 0.3670435 0.4590164 0.4500126 0.5211405 0.6457055    0
ANN     0.2042925 0.3105658 0.4184729 0.4076191 0.4913234 0.6260740    0
```

TOELICHTING CLASSIFICATIE MEETMETHODES

Term	Toelichting
Accuracy	Mate waarin een toegewezen label op een testset overeenkomt met het model. $$Accuracy = \frac{TP + TN}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

TP	True Positive	Een positief ingedeeld element

FP	False Positive	Een positief ingedeeld element die eigenlijk negatief moest zijn
TN	True Negative	Een negatief ingedeeld element
FN	False Negative	Een negatief ingedeeld element die eigenlijk positief moet zijn

Kappa

In de basis vergelijkbaar met *Accuracy*, alleen wordt bij deze methode getracht rekening te houden met de kans op toevallige positieve classificaties (*false positives*). Dit wordt door sommigen beschouwd als een beter betrouwbare methodiek dan *Accuracy*. Deze techniek is voorgesteld door Jacob Cohen, en heet zodoende *Cohen's kappa coefficient* [24].

$$p_e = \frac{1}{N^2}\sum_k n_{k1}n_{k2}$$

p_0	Vernomen accuracy
p_e	Hypothetische kans op false positive
k	Categorieën
N	Aantal elementen

INTERPRETATIE SAMENVATTING

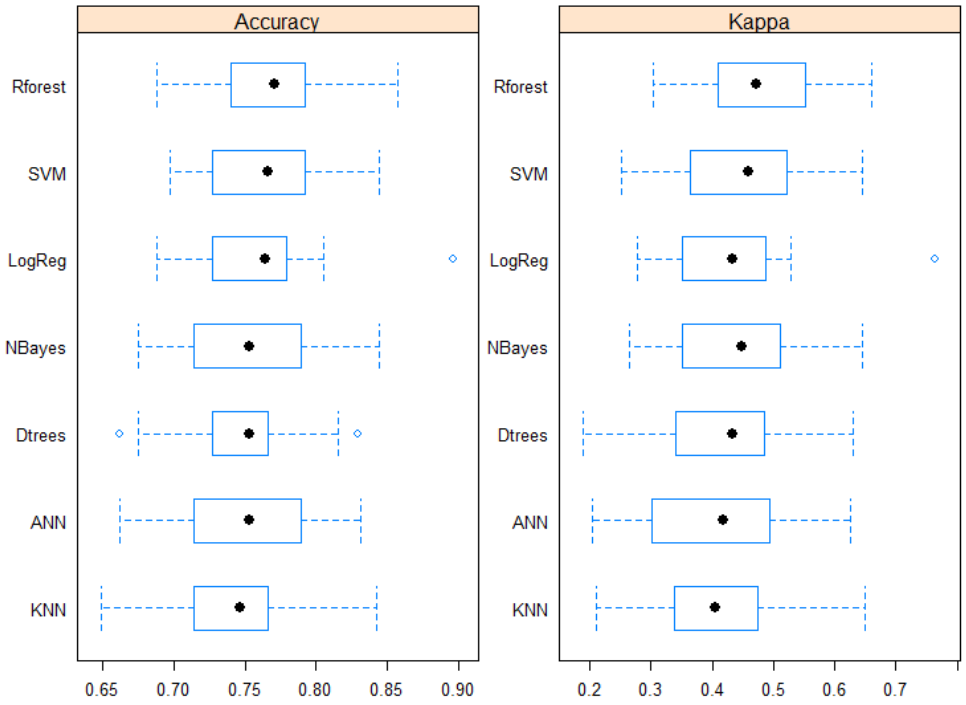
Methodiek	Beste gemiddelde	Beste uiterste
Accuracy	Random Forest	Logistische Regressie
Kappa	Random Forest	Logistische Regressie

In tegenstelling tot de resultaten van regressie liggen de gemiddelden van classificatie veel dichter bij elkaar. Vooralsnog blijkt in Accuracy en Kappa het Random Forest het meest te scoren. SVM lijkt hierin tweede te zijn. Wat opvalt is dat logistische regressie op Accuracy en Kappa het beste eenmalige resultaat behaald hebben. Deze lijkt dan ook een te grote spreiding te hebben in de toekomst betrouwbare voorspellingen mee te doen.

BOX AND WHISKER TABEL

Zoals bij de Box and Whisker bij regressie is beschreven, deelt deze functie automatisch de beste functie bovenaan. Hierdoor wordt bevestigd dat Random Forest het meest betrouwbaar blijkt.

Ook valt op dat de spreiding van gemiddelden wat groter lijkt bij de Kappa.

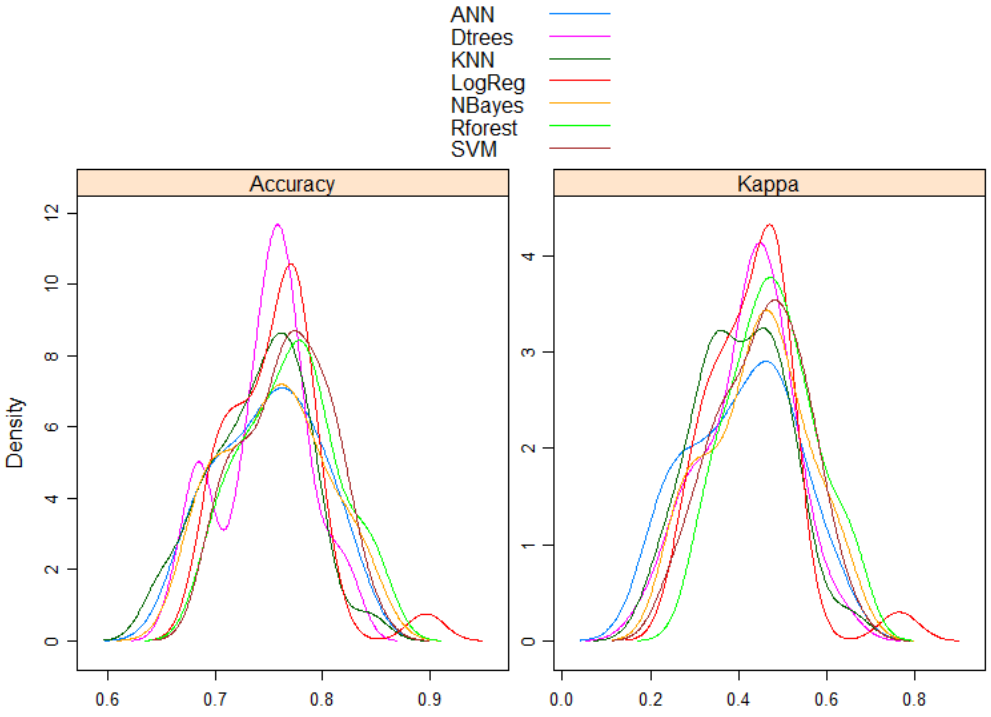


Figuur 6: Box and Whiskers tabel resultaten classificatie op voorbeeldset

DICHTHEIDSVERDELING

Hier wordt op eerste oogopslag bevestigd dat de gemiddelden dicht bij elkaar liggen. Hier geldt dat hoe dichter de piek (gemiddelde van gemiddelden) naar rechts is geplaatst over de X-as (dichter bij de 1), hoe beter een algoritme gepresteerd heeft.

Te zien is dat logistische regressie inderdaad bij Accuracy en Kappa wel een keer het dichts bij een correcte classificatie zaten, te zien aan de bobbel aan de rechterzijde.



Figuur 7: Dichtheidsverdeling gemiddelden van resultaat classificatie op voorbeeldset

BESTE ALGORITME CLASSIFICATIE OP VOORBEELDSET

Na deze test blijkt het Random Forest het beste resultaat te hebben. Wel is SVM daarop dicht in de buurt.

RESULTATEN BENCHMARK VISDATASET

Tijdens het testen bleek de uiteindelijke nauwkeurigheid te laag om in de toekomst op te kunnen bouwen. Om tot deze conclusie te komen is gebruik gemaakt van dezelfde methodiek beschreven in [Methodiek, p. 31]. De samenvatting die hieruit voortvloeit is hieronder beschreven.

```
Call:
summary.resamples(object = class.results)

Models: Rforest, Dtrees, NBayes, KNN, LogReg, SVM, ANN
Number of resamples: 30

Accuracy
             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Rforest 0.6410256 0.7339744 0.7692308 0.7645137 0.7968192 0.8461538    0
Dtrees  0.6025641 0.6635427 0.7051282 0.7037217 0.7435897 0.8333333    0
NBayes  0.6282051 0.7307692 0.7594937 0.7606892 0.7968192 0.8589744    0
KNN     0.5256410 0.5801282 0.6518987 0.6436276 0.6923077 0.7820513    0
LogReg  0.4743590 0.5527020 0.5897436 0.5916586 0.6317348 0.7051282    0
SVM     0.6538462 0.7692308 0.7820513 0.7866331 0.8205128 0.9102564    0
ANN     0.5641026 0.6314103 0.6835443 0.6772909 0.7206264 0.7820513    0

Kappa
             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Rforest 0.4615385 0.6009615 0.6538462 0.6468113 0.6951923 0.7692308    0
Dtrees  0.4038462 0.4967258 0.5576923 0.5556354 0.6153846 0.7500000    0
NBayes  0.4423077 0.5961538 0.6397024 0.6410764 0.6947532 0.7884615    0
KNN     0.2884615 0.3701923 0.4773363 0.4654662 0.5384615 0.6730769    0
LogReg  0.2115385 0.3290664 0.3846154 0.3874406 0.4467402 0.5576923    0
SVM     0.4807692 0.6538462 0.6730769 0.6799628 0.7307692 0.8653846    0
ANN     0.3461538 0.4471154 0.5250680 0.5159115 0.5811152 0.6730769    0
```

Termen die beschreven zijn in deze samenvatting zijn beschreven in [Classification, p. 35]. De vraag ontstaat of het mogelijk is de nauwkeurigheid nog hoger te krijgen (richting de 1).

Er is besloten één specifieke eigenschap toe te voegen aan de dataset. De waarden hiervan zullen gegenereerd worden op basis van viseigenschappen zoals deze gesteld zijn in de database van viseigenschappen. Om de data voor elke vis anders is, is het voor het model simpeler om classificatie te doen.

De verkozen eigenschap die toegevoegd zal worden is lengte. Hiervoor is gekozen omdat lengte van vissen een van de meest gerapporteerde eigenschappen is van vis, waardoor er zekerheid is dat elke vis gedekt is. Daarnaast wordt de eigenschap lengte ook verschaft door het device die in het IoT-vismigratieonderzoek is geproduceerd.

De vissen krijgen elk lengtewaarden gegenereerd die specifiek zijn voor de vissoort. Om deze reden wordt de samengevoegde dataset verwijderd. Deze zal weer worden aangemaakt zodra de aparte visdatasets lengtes bevatten.

Het proces waarin de lengtewaarden gegenereerd worden is beschreven in [Bijlage E]. De volgende test is met de dataset met toegevoegde lengtewaarden.

RESULTATEN BENCHMARK VISDATASET 2

Er wordt nu dezelfde test gedaan als die voordat de lengte aan de vis dataset werd toegevoegd. Ook wordt dezelfde testmethodiek aangehouden. Ter verduidelijking worden de algoritmes opnieuw genoemd:

Random Forest	K-Nearest Neigbor	Artificial Neural Network
Decision Trees	Logistic Regression	
Naïve Bayes	Support Vector Machine	

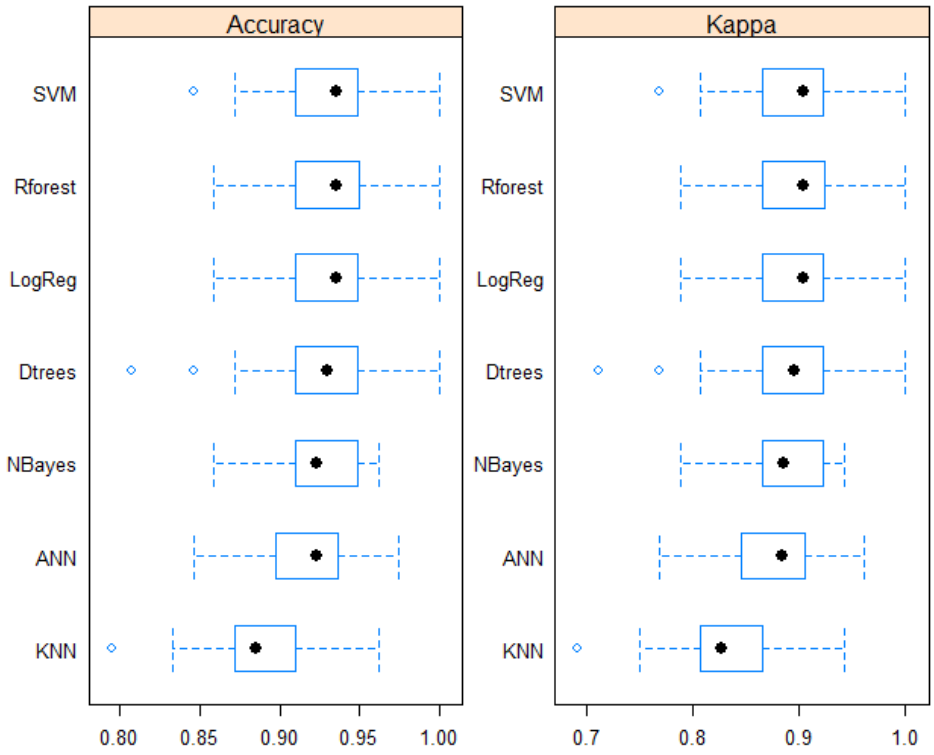
Indien de resultaten positiever dan de resultaten van de dataset zonder lengte, zal het resultaat verder geanalyseerd worden. Hieronder is de samenvatting van het resultaat:

```
Models: Rforest, Dtrees, NBayes, KNN, LogReg, SVM, ANN
Number of resamples: 30

Accuracy
             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Rforest 0.8589744 0.9105404 0.9358974 0.9297414 0.9492048 1.0000000    0
Dtrees  0.8076923 0.9134615 0.9299740 0.9237531 0.9457157 1.0000000    0
NBayes  0.8589744 0.9105404 0.9235638 0.9212106 0.9457157 0.9620253    0
KNN     0.7948718 0.8717949 0.8853457 0.8837715 0.9102564 0.9620253    0
LogReg  0.8589744 0.9105404 0.9358974 0.9293249 0.9487179 1.0000000    0
SVM     0.8461538 0.9102564 0.9358974 0.9284702 0.9487179 1.0000000    0
ANN     0.8461538 0.8977605 0.9230769 0.9165314 0.9365060 0.9743590    0

Kappa
             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Rforest 0.7884615 0.8657493 0.9038462 0.8946084 0.9237432 1.0000000    0
Dtrees  0.7115385 0.8701923 0.8949793 0.8856290 0.9185868 1.0000000    0
NBayes  0.7884615 0.8657493 0.8854186 0.8818092 0.9185411 0.9430288    0
KNN     0.6923077 0.8076923 0.8278195 0.8256256 0.8653846 0.9430288    0
LogReg  0.7884615 0.8657493 0.9038462 0.8939831 0.9230769 1.0000000    0
SVM     0.7692308 0.8653846 0.9038462 0.8927013 0.9230769 1.0000000    0
ANN     0.7692308 0.8465614 0.8846154 0.8747901 0.9046619 0.9615385    0
```

Te zien is dat in vergelijking met resultaten van de dataset zonder lengte, het toevoegen van de lengte veel heeft bijgedragen aan de nauwkeurigheid van het getrainde model. Dit bewijst dat het toevoegen van specifieke eigenschappen van vissen bijdraagt aan de nauwkeurigheid van de voorspelling.

BOX AND WHISKER TABEL

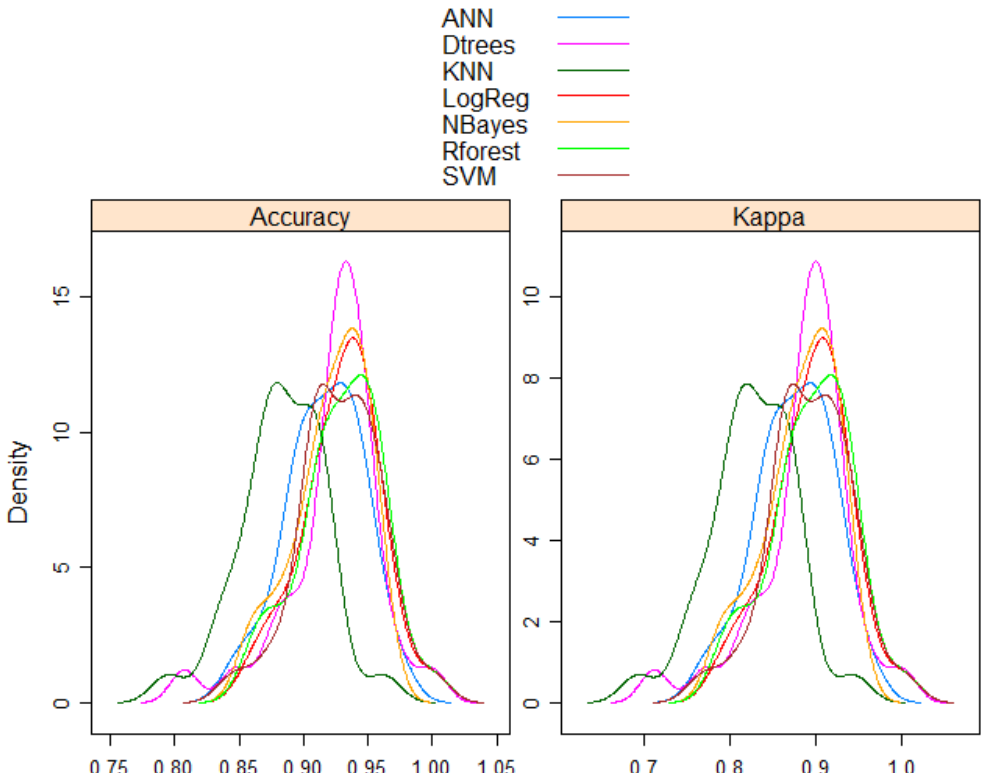
Hier is goed te zien dat de gemiddelden vooral van de eerste 3 algoritmes heel dicht bij elkaar liggen. Ook al staat SVM boven aan, het resultaat geeft aan dat gemiddeld het Random Forest algoritme marginaal beter presteerd.



Figuur 8: Box and Whiskers tabel resultaten classificatie op visdataset

DICHTHEIDSVERDELING

De gemiddelden liggen zo dicht op elkaar dat onderscheiden van algoritmes moeilijk is. Te zien is wel dat de Decision Trees in allebei de gevallen de laagste variantie had. Ook valt op dat het K-Nearest Neighbour algoritme bijzonder slecht presteert.



Figuur 9: Dichtheidsverdeling gemiddelden resultaat regressie op visdataset

CONCLUSIE

Deze test heeft uitgewezen dat het Random Forest algoritme nipt het beste gepresteerd heeft, respectievelijk opgevolgd door SVM en logistische regressie. De gemiddelden liggen zo dicht bij elkaar dat bij het kiezen van een ander algoritme functioneel niet veel zou worden ingeboet.

BIJLAGE E

Tijdens de eerste test met de visdataset in [Bijlage D, Resultaten benchmark visdataset, p. 38] was de nauwkeurigheid van de resultaten niet toereikend. Gekozen was om lengtewaarden toe te voegen aan de dataset. In deze bijlage worden de stappen beschreven die genomen zijn deze waarden te berekenen en genereren.

VOORBEREIDEN DATASET

De in de eerste deelvraag verzamelde dataset zal nu worden gebruikt om de algoritmes te trainen. Voordat dit kan gebeuren, moet de dataset gereed worden gemaakt.

De dataset is zoals het heet in R een `data frame`. Bepaalde data worden gemuteerd of buiten beschouwing worden gelaten, immers is niet alle soort data geschikt om een model mee te trainen.

Zoals bij de eerste deelvraag, kop [Databronnen visdata, p. 13] is gesteld, bestaan de verzamelde dataset uit 2 delen. Het tweede deel zal gebruikt worden als dataset. De gerapporteerde vissen zijn afkomstig van de zee (zout water).

Hierop worden de lengtes van de visrecords in de datasets worden verlaagd naar de vis met de minste records, zodat deze even lang zijn. Hierdoor zal het trainingsmodel geen voorkeur (bias) bevatten door specifieker op een vis te trainen die vaker voorkomt. De zalm dataset bleek de minste elementen te bevatten: 261 records.

Dan worden deze delen van de 3 vissen samengevoegd tot één enkele set. Ten slotte wordt de volgorde van de records in deze set willekeurig gemaakt. Anders wordt de lijst afgelopen in de orde waarin de vissen zijn toegevoegd. Als er een deel gebruikt zal worden om het model te trainen, is er een grote kans dat het model verhoudingsgewijs ongelijkmatig over de vissoorten leert. Hierdoor kan een model naderhand een voorkeur hebben voor een bepaalde vis of vissen (bias), waardoor de uitvoer minder betrouwbaar is.

Tijdens het invoeren van deze dataset blijkt sommige data niet goed te kunnen worden geïnterpreteerd door R. Hierdoor ontstaat de noodzaak om aanpassingen te maken aan de dataset. Deze veranderingen worden toegelicht in [Bijlage C, Tabel 6, p. 36].

ONVOLDOENDE NAUWKEURIGHEID

Tijdens het testen bleek de uiteindelijke nauwkeurigheid te laag om in de toekomst op te kunnen bouwen. Om tot deze conclusie te komen is gebruik gemaakt van dezelfde methodiek beschreven in [Bijlage D, Methodiek, p. 31]. De samenvatting die hieruit voortvloeit is te zien in [Bijlage , Resultaten benchmark visdataset, p. 38].

GENEREREN LENGTEWAARDEN

In [Tabel 9] staan de lengtewaarden voor deze vissen. Genoteerd is het minimum, gemiddeld en maximale lengte. De lengtes zijn afkomstig van verschillende sites [4], [8], [21].

Vissoort	Lengte minimaal (mm)	Lengte gemiddeld (mm)	Lengte maximaal (mm)
Aal	10	700	1300
Driedoornige stekelbaars	10	55	110
Zalm	10	731	1500

Tabel 9: Lengtewaarden vissen

Na een gunstige standaarddeviatie te hebben verkregen, kan er een set worden gegenereerd. Dit wordt in R gedaan met de functie `rnorm()`. Deze functie heeft respectievelijk de volgende waarden nodig:

Parameter rnorm()	Waarde
N (aantal elementen te genereren)	261

Gemiddelde	Gemiddelde van vis uit database viseigenschappen
Standaarddeviatie	Standaarddeviatie verkregen uit sd-berekeningsfunctie [Code 1: SD-calculatiefunctie in R]

- Aantal gewenste te genereren elementen

- Gemiddelde van de set
- Standaarddeviatie van de set
- Op basis van deze gegevens kunnen er normaalverdelingen gemaakt van de te gebruiken lengtes. In R is voor dit onderzoek een functie geschreven die een standaarddeviatie berekend over een set numerieke waarden, met combinatie met minimum- en maximumlimieten. De minimum- en maximumlimieten en het gemiddelde zijn respectievelijk de minimumlengte, maximumlengte en gemiddelde lengte. De volgende waarden zijn gebruik om de uiteindelijke waardes te generen:

Vis	n	Gemiddelde	SD
Aal	261	700	153
Driedoornige stekelbaars	261	55	13
Zalm	261	731	193

SD CALCULATIE

De volgende functie is geschreven om in R een standaarddeviatie te extraheren uit een reeks cijfers binnen een gestelde minimum- en maximumgrens. Als de gegenereerde set de limiet overschrijdt, word de standaarddeviatie voor de volgende iteratie verlaagd. Dit wordt herhaald totdat de gegenereerde set aaneensluitend binnen limiet blijft. De standaarddeviatie die gebruikt werd voor die serie juiste iteraties, wordt returned.

```
#Zijn de waarden binnen bereik?
isOutOfBound <-function(minVal, maxVal, MIN, MAX){
  return((maxVal > MAX) || (minVal < MIN))
}



#element_n    Aantal te genereren elementen
#element_mean Gemiddelde van de te genereren elementen
#element_sd   Start standaarddeviatie, deze moet altijd te hoog worden meegegeven
#MIN          Minimale grenswaarde, globale variabele
#MAX          Maximale grenswaarde, globale variabele
#precision    Aantal checks om te kijken of de elementen binnen de grenswaarden blijven, indien aaneenvolgend te gegeneren met precision getal

sd_app <- function(element_n, element_mean, element_sd, MIN, MAX, precision){
  element_draws <- rnorm(element_n, element_mean, element_sd) #genereer een set met de meegegeven waarden
  maxVal <- round(max(element_draws),2) #haal uit de gegenereerde set het maximale getal en sla op onder deze variabele
  minVal <- round(min(element_draws),2) #haal uit de gegenereerde set het minimale getal en sla op onder deze variabele
  outOfBound <- isOutOfBound(minVal, maxVal, MIN, MAX) #valt lokale minimum en maximum binnen opgegeven limieten?

  green_light <- 0 #Variabele die na een correcte iteratie (alle waarden binnen limieten) met 1 wordt geïncrementeerd. Anders, terug naaar 0

  while(green_light < precision){ #Uit de loop wanneer aaneensluitend juiste iteraties zijn gemaakt meegegeven in precision

    element_draws <- rnorm(element_n, element_mean, element_sd) #genereer een set met mogelijk verkleinde standaarddeviatie
    maxVal <- round(max(element_draws),2) #haal uit deze set opnieuw maximale getal
    minVal <- round(min(element_draws),2) #haal uit deze set opnieuw minimale getal
```

```
    outOfBound <- isOutOfBound(minVal, maxVal, MIN, MAX) #ligt de min en max binnen het limiet?

    if(outOfBound) { #buiten limiet
      green_light <- 0
      element_sd <- element_sd - 1 #verlaag standaarddeviatie met 1 om spreiding volgende set te dempen
      #print("else")

    } else { #binnen limiet
      green_light <- green_light + 1#incrementeer waarde voor correcte iteratie
    }

  }

  return(element_sd)
}
```

Code 1: SD-calculatiefunctie in R

BIBLIOGRAFIE

[1] S. nederland, „Vissterfte," [Online]. Available: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0ahUKEwil7NjWjt7bAhWDjqQKHXJkAT8QFggvMAE&url=http%3A%2F%2Fwww.sportvisserijnederland.nl%2Ffiles%2Fbasisboek-visstandbeheer-h5_5830.pdf&usg=AOvVaw0CzF9Cgi8H1Lu043QA_Yk4. [Geopend 2018 juni 2].

[2] S. D. Online, „Duurzaam waterbeheer," Stichting Deltawerken Online, 2004. [Online]. Available: http://www.deltawerken.com/duurzaam-waterbeheer/64.html.

[3] P. (. a. W. Philipsen en H. (. W. C. Wageningen, „Routekaart voor vismigratie in de Nederlandse Delta".

[4] „Vissoorten," Sportvisserij Nederland, 2018. [Online]. Available: https://www.sportvisserijnederland.nl/vis-water/vissoorten/.

[5] „Ruim baan voor trekvissen," Magazine Rijkswaterstaat, 14 mei 2018. [Online]. Available: https://www.magazinesrijkswaterstaat.nl/programmakrw/2018/01/ruim-baan-voor-trekvissen. [Geopend 2018 mei 20].

[6] M. Verschoor, „Het gaat weer beter met de paling in Nederland," Algemeen Dagblad, 27 februari 2017. [Online]. Available: https://www.ad.nl/economie/het-gaat-weer-beter-met-de-paling-in-nederland~a8f7ef1b/. [Geopend 2018 mei 10].

[7] J. Reumer, „Heel langzaam kan straks de zalm terugkeren in Nederland," Trouw, 2 december 2017. [Online]. Available: https://www.trouw.nl/groen/heel-langzaam-kan-straks-de-zalm-terugkeren-in-nederland~a5904e7d/. [Geopend 2018 mei 10].

[8] „Fishbase Main Page," Fishbase.org, [Online]. Available: https://www.fishbase.org/. [Geopend 8 april 2018].

[9] K. Kaschner, K. Kesner-Reyes, C. Garilao, J. Rius-Barile, T. Rees en R. Froese, „Computer generated distribution maps for Salmo salar (Atlantic salmon), with modelled year 2100 native range map based on IPCC A2 emissions scenario," www.aquamaps.org, [Online]. Available: www.aquamaps.org. [Geopend 2 juni 2018].

[10] „Fish Dataset," Qut Robotics, 6 juni 2016. [Online]. Available: https://wiki.qut.edu.au/display/cyphy/Fish+Dataset. [Geopend 2018 mei 20].

[11] P. X. Huang, B. B. Boom en R. B. Fisher, „Fish Recognition Ground-Truth data," Fish4Knowledge Project, 23 september 2013. [Online]. Available: http://groups.inf.ed.ac.uk/f4k/GROUNDTRUTH/RECOG/. [Geopend 2018 juni 18].

[12] L. C. D. S.L., „Smart Water Technichal Guide," Augustus 2017. [Online]. Available: http://www.libelium.com/development/waspmote/documentation/smart-water-board-technical-guide.

[13] „Waterdata," Rijkswaterstaat, [Online]. Available: https://www.rijkswaterstaat.nl/water/waterdata-en-waterberichtgeving/waterdata/index.aspx. [Geopend 2018 april 20].

[14] „10 Machine Learning Algorithms You Should Know in 2018," Octoparse, 1 januari 2018. [Online]. Available: https://www.octoparse.com/blog/10-machine-learning-algorithms-you-should-know-in-2018/. [Geopend 2018 mei 24].

[15] „Top 10 Machine Learning Algorithms," www.dezyre.com, 2018 mei 11. [Online]. Available: https://www.dezyre.com/article/top-10-machine-learning-algorithms/202. [Geopend 2018 mei 24].

[16] J. Le, „A Tour of The Top 10 Algorithms for Machine Learning Newbies," Towards Data Science, 20 januari 2018. [Online]. Available: https://towardsdatascience.com/a-tour-of-the-top-10-algorithms-for-machine-learning-newbies-dde4edffae11. [Geopend 2018 mei 24].

[17] J. Le, „https://gab41.lab41.org/the-10-algorithms-machine-learning-engineers-need-to-know-f4bb63f5b2fa," Lab41, 11 juli 2016. [Online]. Available: https://gab41.lab41.org/the-10-algorithms-machine-learning-engineers-need-to-know-f4bb63f5b2fa. [Geopend 2018 mei 24].

[18] R. Shaw, „Top 10 Machine Learning Algorithms for Beginners," KDnuggets, oktober 2017. [Online]. Available: https://www.dataquest.io/blog/top-10-machine-learning-algorithms-for-beginners/. [Geopend 2018 mei 24].

[19] „Machine Learning Algorithms Pros and Cons," HackingNote, [Online]. Available: https://www.hackingnote.com/en/machine-learning/algorithms-pros-and-cons/. [Geopend 2018 mei 23].

[20] L. D. Hamilton, „Machine Learning Algorithm Cheat Sheet," 2014 september 2014. [Online]. Available: http://www.lauradhamilton.com/machine-learning-algorithm-cheat-sheet. [Geopend 2018 mei 25].

[21] E. E. A. R. F. Kathleen Kesner-Reyes, J. Ready, M. J. France, R. Barile en K. Kaschner, „AquaMaps: An Overview," 2008. [Online]. Available: https://www.aquamaps.org/main/presentations/AquaMaps_General0908.ppt. [Geopend 2018 juni 1].

[22] „Nutriënten en eutrofiëringsparameters," Rijkswaterstaat Waterinfo, 18 juni 2018. [Online]. Available: https://waterinfo.rws.nl/#!/nav/expert/parameters/Nutri___C3___ABnten___20en___20eutrofi___C3___ABringsparameters/. [Geopend 18 juni 2018].

[23] J. Brownlee, „Compare The Performance of Machine Learning Algorithms in R," Machine Learning Mastery, 26 februari 2016. [Online]. Available: https://machinelearningmastery.com/compare-the-performance-of-machine-learning-algorithms-in-r/. [Geopend 2018 mei 12].

[24] J. Cohen, „A Coefficient of Agreement for Nominal Scales," 1 april 1960. [Online]. Available: http://journals.sagepub.com/doi/10.1177/001316446002000104. [Geopend 2018 juni 1].

[25] „Belang van vismigratie," Ruim baan voor Vissen, [Online]. Available: http://www.ruimbaanvoorvissen.nl/page.aspx?id=2.

[26] „Smart Water Board Guide," Libelium, augustus 2017. [Online]. Available: http://www.libelium.com/development/waspmote/documentation/smart-water-board-technical-guide/?action=download. [Geopend 2018 mei 20].

[27] „Pima Indians Diabetes Database," Kaggle, 2016. [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database.