Data science-technieken voor regenwateroverlast in stedelijk gebied

Christiaan Lamers (voorheen Universiteit Leiden), Jan van Rijn (Universiteit Leiden), Ton Beenen (STOWA, RIONED)

Dit artikel beschrijft een casestudy waarin data science-technieken zijn toegepast op de voorspelling van regenwateroverlast in stedelijk gebied. Gebaseerd op de hoogtekaart en radarmetingen wordt een 'random forest'-model geleerd te voorspellen of in een bepaald gebied, bij een bepaalde regenval, meldingen van regenwateroverlast zullen ontstaan. Dit model heeft een nauwkeurigheid van 57,9%, een precisie van 62,6% en een recall van 26,4%. Hoewel dit model slechts een abstractie is van het echte probleem, biedt deze procedure veel potentie voor de toekomst, in het bijzonder wanneer toegang tot extra databronnen, zoals data over verzekeringsclaims, beschikbaar komt.

Data zijn alomtegenwoordig in onze maatschappij. Denk daarbij aan de sensornetwerken die de waterstanden en de waterkwaliteit in kanalen en plassen monitoren, de telescopen die het universum en de aarde observeren, de ziekenhuizen die het aantal patiënten en de impact van bepaalde medicijnen bijhouden. Data zijn waardevol door de informatie die ze geven over het verleden en de trends die ze daaruit kunnen voorspellen voor de toekomst. Ook de waterschappen en gemeenten verzamelen veel data, zoals hoeveelheid en kwaliteit van het influent bij een rwzi, waterpeilen en pompschakelingen in gemaalkelders of meldingen van wateroverlast in zowel stedelijk als landelijk gebied.
Wanneer in korte tijd veel regenwater valt, kan de riolering de afvoer soms niet bijhouden en staat het water enige tijd, als het ware in de file, op straat. Dit water op straat is geen probleem, zolang weggebruikers hier geen hinder van ondervinden en zolang het water geen huizen of andere gebouwen instroomt. In de praktijk blijkt het redelijk lastig om de lokale bergingscapaciteit van een gebied te bepalen om daaruit te voorspellen bij welke combinatie van neerslagintensiteit en -hoeveelheid hinder of schade zal ontstaan.
Dit artikel beschrijft een toepassing van data science op deze problematiek. Het artikel gaat in op de volgende vragen: 1) Welke databronnen zijn beschikbaar voor voorspelling van regenwateroverlast in stedelijk gebied? 2) Welke data science-methoden kunnen deze data klaar maken voor gebruik door voorspellingsmodellen? en 3) hoe accuraat zijn de voorspellingen die uit het resulterende model komen? Dit onderzoek is een multidisciplinaire samenwerking tussen stakeholders uit data science en waterbeheer.

Achtergrond

Het vakgebied data science, en in het bijzonder Machine Learning, ontwikkelt technieken die trends in data kunnen ontdekken en hiervoor automatisch modellen kunnen genereren. Twee noodzakelijke voorwaarden zijn dat:

- er een correlatie bestaat tussen de invoerdata en de gezochte trend voor de voorspelling;
- er voldoende data beschikbaar zijn om de patronen automatisch te ontdekken.

Data science-technieken werken typisch op een matrix van getallen (bijvoorbeeld geordend in een

Excelsheet). Dit noemen we de dataset. Iedere kolom van de matrix representeert een specifiek attribuut of kenmerk van de waarneming, iedere rij representeert een observatie. Grofweg zijn er twee soorten attributen, 1) de voorspellende attributen en 2) het attribuut dat voorspeld moet worden (ook wel het label of de klasse genoemd). Als we bijvoorbeeld wateroverlast willen voorspellen, zijn terreinkenmerken en neerslagmetingen de voorspellende attributen. Wel of geen wateroverlast is het label of de klasse. Elke rij in de matrix bevat dan de waarden van de attributen op een locatie per gekozen tijdvak.

Het is vaak vrij gemakkelijk om aan veel observaties met voorspellende attributen te komen. De uitdaging zit erin om tot veel observaties met het label of klasse te komen. Dit vereist vaak menselijk inzicht of zelfs de inzet van domeinexperts. Wanneer er genoeg observaties van hoge kwaliteit met labels beschikbaar zijn, is het mogelijk een model te bouwen dat de labels voorspelt voor alle nieuwe waarnemingen zonder labels [1].

Voor deze specifieke casus zou dat het volgende betekenen. We zijn op zoek naar gedocumenteerde gevallen van regenwateroverlast en tegenvoorbeelden. Hier moeten bepaalde attributen uit worden geëxtraheerd, bijvoorbeeld de hoeveelheid regenwater die in de voorafgaande periode is gevallen, of de data van de terreinkenmerken (hoeveel procent gras, tegels, etc). Was er inderdaad regenwateroverlast, dan is het label positief, bij geen wateroverlast is het label negatief. Soms moeten meerdere databronnen worden samengevoegd om tot de juiste attributen te komen.

Geautomatiseerde routines verkennen de dataset en maken daar keuzes uit voor de bouw van het voorspellingsmodel. De precieze werking van de modelontwikkeling valt buiten de scope van dit artikel, maar is bijvoorbeeld te vinden in Hastie et al. [1]. De data waarop een model wordt getraind, wordt de trainingsset genoemd. Om een model te evalueren, moet getest worden op data waarop het model niet is getraind, maar waarvan de labels wel bekend zijn. Een deel van de observaties uit de dataset worden daarom apart gezet en gereserveerd voor dit evaluatieproces. Dit deel van de dataset heet de testset.

Gebaseerd op de voorspellingen in de testset zijn bepaalde statistieken te berekenen over de kwaliteit van de voorspelling. Bijvoorbeeld het percentage goede voorspellingen (*accuracy*). Dit geeft soms een vertekend beeld, aangezien het moeilijk te zeggen is wat een goede nauwkeurigheid is in een ongebalanceerde dataset. Daarom wordt vaak ook gekeken naar de *precision* (het percentage gevallen waarin regenwateroverlast wordt voorspeld, waarbij dit ook daadwerkelijk het geval was) en de *recall* (welk percentage van de positieve klasse werd als zodanig voorspeld). Er is een afweging tussen precision en recall; het spreekt voor zich dat zelfs een simplistisch model op een van deze criteria altijd een perfecte score kan behalen door altijd of nooit de positieve klasse te voorspellen. In deze casestudy zal de voorspelling van ‘nooit wateroverlast’ een goede score geven op precision. Er is immers maar zelden wateroverlast. Op recall is dit echter een slechte score omdat het model geen enkele keer goed voorspelde als er wel wateroverlast was.

Databronnen

In dit project zijn de volgende databronnen betrokken: KNMI-neerslagradarmetingen, de Algemene Hoogtekaart van Nederland (AHN2) en meldingen van regenwateroverlast op Twitter. Deze bronnen met hun eigenschappen zijn weergegeven in tabel 1.

Tabel 1. Overzicht van databronnen

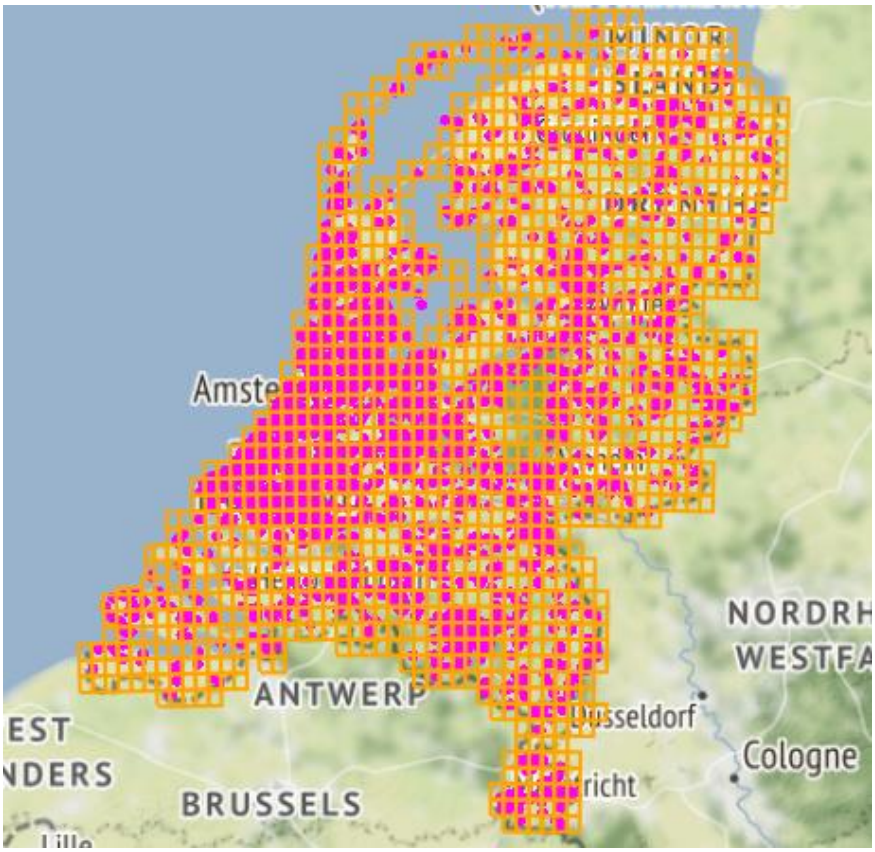
Naam	**Resolutie**	**Kwantiteit**	**Coördinaten**
KNMI-neerslagradar	Ca. 1000x1000 m	Hoog	KNMI-radarvakken
AHN2-terreindata	5x5 m	Hoog	Longitude / Latitude
Twitterberichten regenwateroverlast	Puntprecisie	7000	longitude / Latitude

Radardata: De KNMI-radardata zijn aangeleverd door STOWA en Stichting RIONED. Hoewel deze data ook beschikbaar zijn op de website van het KNMI, is de versie die hier is gebruikt voorbewerkt met een aantal procedures met het programma RadarTools. Daardoor zijn makkelijk bepaalde neerslagintensiteiten te selecteren. Deze data zijn beschikbaar voor geheel Nederland, derhalve is de kwantiteit hoog. Daarnaast is er momenteel geen reden om aan te nemen dat de resolutie niet groot genoeg is.
Terreindata: Als terreindata hebben we gebruik gemaakt van de Algemene Hoogtekaart van Nederland 2 (AHN2). Hoewel de AHN3 een hogere resolutie heeft, dekt deze op het moment van uitvoering nog niet de regio Enschede, waar veel datapunten voor dit onderzoek beschikbaar waren. Ook voor deze databron was een vrij hoge kwantiteit data beschikbaar.
Twitterberichten regenwateroverlast: Zoals eerder aangegeven is het voor data science-projecten van belang om voldoende labels te hebben. Dit betekent dat er een goede bron van voorbeelden van regenwateroverlast nodig is. Omdat deze niet beschikbaar was, is naar socialmediadata gekeken, in het bijzonder van Twitter. Door gebruik te maken van zoektermen die een grote correlatie leken te hebben met regenwateroverlast zijn ongeveer 7000 relevante tweets met locatie gevonden. Hoewel er stappen zijn ondernomen om deze dataset op te schonen, zijn er op basis van een willekeurige steekproef nog steeds Twitterberichten in deze dataset die wel claimen regenwateroverlast te representeren, maar dit in werkelijkheid niet doen. Dit staat bekend als ‘label ruis’. De meeste data science-modellen kunnen hier tot op zekere hoogte mee omgaan, maar wanneer er veel label ruis is gaat dat ten koste van de kwaliteit van de resultaten.

Uniformiteit

Een uitdaging bij dit project is de verschillende schaal en de verschillende coördinatenstelsels van de databronnen. Aan de ene kant zijn er de KNMI-radarvakken, die in een stereografische projectie zijn weergeven. Een stereografische projectie is een techniek om radarmetingen over de Aardebol op een 2-dimensionaal vlak weer geven [2]. De resolutie is relatief laag. Aan de andere kant zijn de terreindata in een relatief hoge resolutie, weergegeven in longitude (lengtegraad) en latitude (breedtegraad). Dit wordt geïllustreerd in afbeelding 1. Ieder oranje vlak is hier een radarvak en ieder roze punt representeert een gevonden Twitterbericht met regenwateroverlast. Ieder oranje vlak correspondeert ook ongeveer met

40.000 AHN2-datapunten, al gebruikt het KNMI ook een ander coördinatenstelsel.



Afbeelding 1. De schaal van verschillende datapunten. Hoogtekaartvakken in oranje, Twitterberichten in paars

Hier moet op een juiste manier mee om worden gegaan. De volgende sectie beschrijft hoe het resolutieverschil kan worden aangepakt.

Auto-encoders

Een andere uitdaging is het verschil in schaal tussen de radardata en de terreindata. In zijn algemeenheid betekent meer datapunten altijd betere kwaliteit, en meer attributen betekenen meer valkuilen. Een informele stelregel is dat er altijd minstens zoveel datapunten moeten zijn als attributen.

Aangezien de kaart met Twitterberichten op dezelfde schaal staat als de radardata, zou dit betekenen dat ieder Twitterbericht tegenover 40.000 (factor 200 in beide richtingen) attributen uit de terreindata zou moeten worden gemodelleerd. Hoewel dat op zichzelf zou kunnen, is het risico dat er willekeurige, niet op de realiteit gebaseerde, patronen worden gevonden.

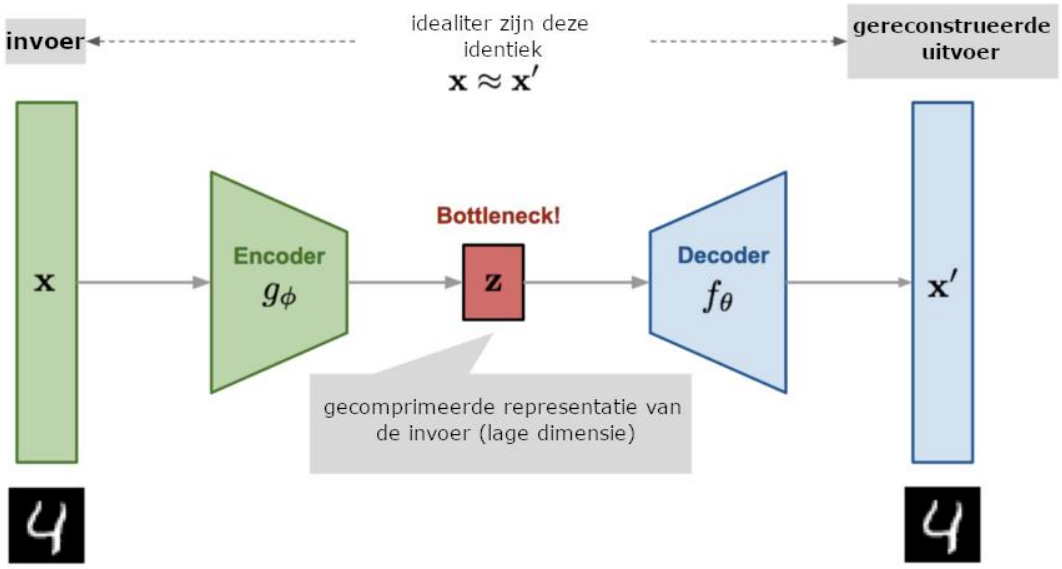
Er zijn verschillende mogelijkheden om dit op te lossen:

- Modelleren met alle attributen. Het risico hierbij is dat het model 'overfit' op willekeurige patronen, en er geen op realiteit gebaseerde patronen worden ontdekt.

- Modelleren met een lagere resolutiekaart. Er zullen bepaalde eigenschappen wegvallen.

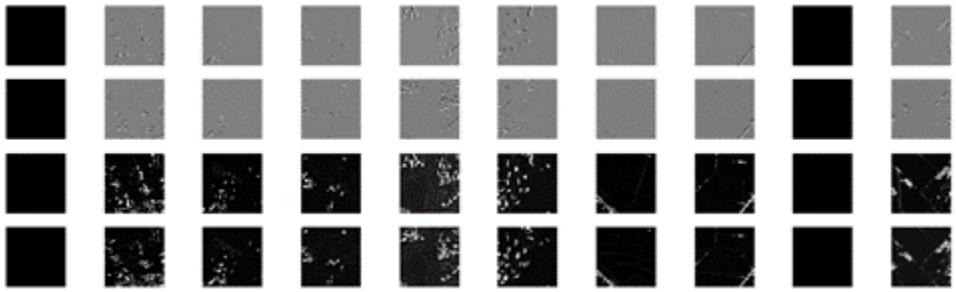
- Het gebruik van een auto-encoder [3], die door middel van compressie de data in een lagere dimensie encodeert.

Auto-encoders hebben een grote hoeveelheid data nodig, maar zijn typisch niet afhankelijk van labels. Zodoende is de auto-encoder te trainen op basis van de terreindata, en later toe te passen in de context van andere databronnen. Afbeelding 2 geeft schematisch weer hoe een auto-encoder werkt.



Afbeelding 2. Schematische weergave van een Auto-encoder [4]

Aan de linkerkant staat de invoer, weergegeven met een **x**. In dit geval het handgeschreven getal 4. Door een encoder wordt dit gecomprimeerd tot een kleinere dimensie, hier weergegeven met **z**, die vervolgens weer wordt gedecodeerd naar de originele dimensie.
Aan de rechterkant zien we de uitvoer van de Auto-Encoder, weergegeven met **x'**. Idealiter is dit hetzelfde als (of zo gelijk mogelijk aan) de invoer.
Standaard Machine Learning-technieken (zoals Stochastic Gradient Descent) [3] kunnen worden gebruikt om dit te bewerkstelligen. Uiteindelijk draait het om de waarden bij **z**. Hier zijn de data het meest gecomprimeerd. Dit is een compacte representatie van de eerdere invoer. Deze waarden zullen in het uiteindelijke model (het random forest) worden gebruikt.



Afbeelding 3. Invoer en uitvoer van de auto-encoder

Afbeelding 3 illustreert de werking van een auto-encoder die de hellingshoek van het landschap modelleert. De bovenste twee rijen van deze afbeelding tonen de invoer. De onderste twee rijen tonen de uitvoer. Als invoer zijn de x- en y-componenten van de gradiënt van het landschap gekozen, omdat de hellingshoek als relevanter voor wateroverlast wordt beschouwd dan de absolute hoogte. Een auto-encoder zou ook de directe hoogtekaart kunnen modelleren, maar experimentele evaluatie leert dat dit minder goed werkt.
Een visuele vergelijking toont dat de uitvoer van de auto-encoder veel accurater is dan het simpelweg neerschalen van de data. Dit valt te kwantificeren met de zogenaamde 'Mean Squared Error' [1].
De verwachting is dat het model op basis van de data van de Auto-Encoder een beter resultaat geeft dan het model op basis van eenvoudig neergeschaalde data.

Procedure en evaluatie

De van Twitter verzamelde dataset bevat slechts positieve gevallen (voorbeelden van vermeende regenwateroverlast). Om een model te kunnen leren, zullen ook voorbeelden waar geen regenwateroverlast heeft plaatsgevonden nodig zijn. Gelukkig zijn deze er in overvloed, onder de sterke aanname dat een zware bui in dichtbevolkt gebied in combinatie met een gebrek aan Twitterberichten een goed voorbeeld van de negatieve klasse is. Er worden net zoveel negatieve als positieve gevallen verzameld, zodat de dataset gebalanceerd is.
Zodra de dataset verzameld en samengevoegd is, wordt deze eerst opgesplitst in een trainingsset en testset. De trainingset is om het model op te leren, de testset is om het model te evalueren. De data wordt eerst voorbewerkt in de auto-encoder, zodat deze de gecomprimeerde attributenset kan leren. Naar aanleiding van deze gecomprimeerde attributenset kunnen we weer andere modellen trainen, in dit geval een random forest. Random forests staan erom bekend dat ze met minimale optimaliatieprocedures een uiterst goed resultaat opleveren [5], [6]. Dit model is dan in staat om, gegeven een positie op de hoogtekaart en een hoeveelheid neerslag, te voorspellen of dit tot Twitterberichten over regenwateroverlast leidt.
Zoals eerder genoemd moet een machine learning-model worden getest op data die het niet eerder heeft gezien. Wanneer het model de testdata al eens gezien heeft, is het niet moeilijk een perfecte score te behalen.
In de context van dit project was het nodig om het model verder te testen. Omdat de AHN2-data in

verschillende vakken is opgedeeld, was het belangrijk dat alle waarnemingen uit een bepaald vak ofwel in de trainingsset zaten, ofwel in de testset. Op die manier is te garanderen dat het model niet eerder getoonde terreindata onthoudt, maar echt generaliseert naar nieuwe, ongeziene terreindata. Dit maakt het natuurlijk wel lastiger. Tabel 2 presenteert het resultaat na tienmaal uitvoeren van het experiment met verschillende random states. Gepresenteerd zijn de gemiddelde resultaten en de standaarddeviatie.

Tabel 2. Resultaten van het model

Accuracy	57,9% ± 1,6
Precision	62,6% ± 3,4
Recall	26,4% ± 1,3

Het model wordt geanalyseerd aan de drie eerder genoemde criteria: accuracy, precision en recall. Aangezien het een binair classificatieprobleem betreft (er zijn maar twee mogelijke uitkomsten), zou een model dat willekeurige keuzes maakt een accuracy van 50% behalen. Het feit dat het gepresenteerde model hier ruim boven zit, toont dat het waardevolle correlaties in de data heeft ontdekt, en daardoor potentie heeft.

De waarden bij precision en recall zijn ingewikkelder te analyseren. Een hypothetisch model dat altijd dezelfde waarde voorspelt (altijd regenwateroverlast, of nooit regenwateroverlast) zou op een van deze twee waarden de perfecte score behalen, en op de andere 0% scoren. Het getoonde model heeft blijkbaar een neiging naar voorspellingen die geen regenwateroverlast aangeven: wanneer het model regenwateroverlast voorspelt, zit het in 62,6% van de gevallen goed. Het weet echter slechts 26,4% van deze gevallen correct te identificeren.

Het feit dat zowel precision als recall boven de 0 zitten, in combinatie met een accuracyscore significant boven de 50%, toont de potentie van het model aan.

Conclusies en toekomstvisie

Het was de bedoeling om door middel van de KNMI-regendata en de AHN2-terreindata te voorspellen of er op een zeker moment regenwateroverlast zou optreden, gemeten met Twitterberichten. Het getrainde model behaalt een accuracy van 57,9%, een precision van 62,6% en een recall van 26,4%, wat een hele redelijke score lijkt. Hoewel Twitterberichten niet het directe probleem representeren, ruis in de labels bevatten en fout kunnen worden geïnterpreteerd, is dit voor nu de beste abstractie van het probleem waar data voor beschikbaar zijn. Het feit dat er zelfs op een dataset met deze mate van ruis op de labels een goed resultaat wordt gevonden, betekent dat de methode potentie heeft.

Hoewel de resultaten veel perspectief bieden, is er een aantal potentiële verbeteringen die in de toekomst kunnen worden doorgevoerd.

Hogere kwantiteit van de data: machine learning-technieken werken beter wanneer er veel data beschikbaar zijn. Het verkrijgen van meer data is vrijwel kosteloos: ieder jaar verschijnen er meer Twitterberichten online, die met de huidige scripts automatisch kunnen worden gedownload.

Hogere kwaliteit van de labels: zoals gezegd zijn de Twitterberichten slechts een abstractie van het werkelijke probleem dat moet worden voorspeld, namelijk of er daadwerkelijk regenwateroverlast op straat heeft plaatsgevonden. Er zijn betere bronnen die dit kunnen verifiëren (bijvoorbeeld data over schadeclaims bij verzekeraars, satellietdata), al zijn deze momenteel nog niet toegankelijk.
Hogere resolutie: Door gebruik te maken van AHN3, een hoogtekaart met een hogere resolutie, of van plaatselijke regenstations (zoals bijvoorbeeld de Netatmo-weerstations voor thuis), zou het model met meer informatie gevoed kunnen worden.
Werken met buikenmerken: in de huidige opzet zijn neerslagdata geaggregeerd tot dagsommen. Het maakt uiteraard een verschil of er op een dag gespreid over de gehele dag 20 millimeter in een kilometervak valt, of dezelfde kwantiteit binnen een uur. Het model heeft hier momenteel geen kennis van. In een volgende versie zal het model wellicht regenbuikenmerken krijgen, in plaats van dagsommen. De aanname is dat deze extra informatie een accurater model zal opleveren.
Gebruik maken van een getuned model: goede machine learning-technieken werken met vele zogeheten hyperparameters, parameters van het model die precies goed moeten worden afgesteld om goede resultaten te kunnen verwachten. Dit zijn complexe technieken om toe te passen, maar ze leveren standaard betere resultaten op, doordat er langer en efficiënter naar een goed model wordt gezocht [7]. Wanneer het model zou worden geüpgraded met een betrouwbare databron, zou het potentieel een belangrijk maatschappelijk doel kunnen dienen bij het voorkomen van waterschade in stedelijk gebied en het identificeren van frauduleuze verzekeringsclaims.

Verantwoording
Dit onderzoek is uitgevoerd door het Leiden Institute of Advanced Computer Science (Universiteit Leiden) in opdracht van de Stichting Toegepast Onderzoek Waterbeheer (STOWA) en Stichting RIONED.

Referenties

1. Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning*. New York: Springer series in statistics
2. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
3. https://en.wikipedia.org/wiki/Stereographic_projection
4. https://mc.ai/auto-encoders-and-the-battle-of-generations/
5. Rijn, J. N. van, & Hutter, F. (2018). 'Hyperparameter importance across datasets'. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2367-2376).
6. Probst, P., Boulesteix, A. L., & Bischl, B. (2019). 'Tunability: Importance of Hyperparameters of Machine Learning Algorithms'. *J. Mach. Learn. Res.*, *20*(53), 1-32.
7. Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). *Automated machine learning: methods, systems, challenges* (p. 219). Springer Nature.