

Toepassing van machine learning-technieken op voorspelling van regenwateroverlast in stedelijk gebied

Christie Bavelaar, Thijs Simons, Mitra Baratchi, Jan van Rijn (Universiteit Leiden), Ton Beenen (Stichting RIONED, STOWA)

Dit artikel beschrijft een onderzoek waarin machine learning-technieken worden gebruikt om regenwateroverlast (hinder, overlast en schade) te voorspellen. Dit gebeurt aan de hand van neerslagdata van het KNMI, de Actueel Hoogtebestand Nederland-data en P2000-meldingen. De gebruikte methodes uit het AutoML-vakgebied genereren automatische voorspellende modellen. AutoML-methoden zijn methoden die gebruikers ondersteunen in het bouwen van modellen voor complexe problemen. Zo kan het model voorspellen welke gebieden kwetsbaar zijn voor regenwateroverlast. Hier worden drie verschillende methodes beschreven om deze technieken te evalueren. De resultaten zijn in alle gevallen veelbelovend.

Recente overstromingen zoals in Limburg, Duitsland en België wijzen op het gevaar van extreme regenval. Als overstromingsgevoelige locaties bekend zouden zijn, kunnen overheden en inwoners preventieve maatregelen nemen. Dan kan men schade beperken of voorkomen. Het voorspellen van hinder- of schadegevoelige locaties bij hevige neerslag is echter moeilijk. Extreme buien komen nog weinig voor en gegevens van eerdere wateroverlast zijn nauwelijks toegankelijk geregistreerd.

Het meeste onderzoek in het domein stedelijk waterbeheer maakt gebruik van hydraulische modellen om de stroming van het water te voorspellen [1]. *Machine learning*-technieken zijn een nieuwe manier om regenwateroverlast te voorspellen [2]. Omdat daar de meeste schade ontstaat, richt dit onderzoek zich op stedelijke gebieden. De onderzoeksvraag is: hoe kunnen *machine learning*-technieken, toegepast op verschillende databronnen, worden gebruikt om regenwateroverlast te voorspellen?

Om deze vraag te beantwoorden worden verschillende databronnen samengevoegd en wordt hierop een machine learning-model toegepast. In een eerder onderzoek [1] zijn de neerslagdata van het KNMI, de Actueel Hoogtebestand Nederland-data en twitterberichten over regenwateroverlast met elkaar gecombineerd, om te voorspellen welke gebieden kwetsbaar zijn voor regenwateroverlast. Het model behaalde een nauwkeurigheid (accuracy) van 57,9% en liet daarmee de potentie van de methode zien. Het gebruik van twitterberichten in dit eerdere onderzoek is een belangrijke beperking die ook door de auteurs benoemd wordt. Twitterberichten over wateroverlast zijn slechts een symptoom van het echte probleem: schade door regenwater in de vorm van tijdverlies, geld en leed. Toch werd aan de hand van dit onderzoek duidelijk dat het gebruik van databronnen en de *Data Science*-methode veel potentie heeft voor stedelijk waterbeheer. Wij bouwen op dit onderzoek voort door deze te vervangen door meldingen uit het P2000-netwerk dat hulpdiensten gebruiken, en een meer rigoureuze evaluatiemethode.

Voorspellingsmodel

De waarde van dit onderzoek zit in het combineren van verschillende databronnen, waardoor een dataset van voorbeelden ontstaat waar het model op kan worden getraind. Het belangrijkste onderdeel hiervan is een databron die aangeeft wanneer er regenwateroverlast heeft opgetreden en

wanneer niet (de klasse in het model waar de situatie toe behoort). Uiteindelijk zijn voorbeelden nodig waar wel regenwateroverlast optrad en voorbeelden waar geen regenwateroverlast optrad. Dit zijn lastig te verkrijgen data. In dit onderzoek worden twitterberichten en meldingen van wateroverlast op het alarmsysteem P2000 gebruikt. Daarnaast zijn een aantal databronnen met beschrijvende kenmerken (invoer) nodig. Deze data zijn doorgaans makkelijker in grote hoeveelheid te verkrijgen. In dit onderzoek zijn daarvoor de AHN2- en KNMI-regenradardata gebruikt.

Voorbeelden waar regenwateroverlast optrad

Het is moeilijk vast te stellen waar en wanneer in Nederland gevallen van regenwateroverlast zijn geweest. Er is geen uniforme registratie, waardoor de gegevens niet openbaar toegankelijk zijn. Er moeten dus aannames worden gedaan, gebaseerd op wat in de data te zien is. Bij de P2000-meldingen komen hulpdiensten (zoals brandweer) ter plekke. Het is aannemelijk dat dit een accurate beschrijving van de werkelijkheid oplevert. Het onderzoek bevestigt dat de P2000-meldingen betrouwbaarder zijn voor gebruik dan de twitterberichten. De resultaten in dit artikel zijn primair op de P2000-meldingen gebaseerd. De situaties gekoppeld aan een P2000-melding worden de positieve klasse genoemd.

De twitterberichten zijn niet heel betrouwbaar bij het voorspellen van regenwateroverlast. Iemand kan bijvoorbeeld een twitterbericht sturen over een straat die een aantal dagen geleden is ondergelopen. De locatie en datum van het bericht kloppen dan niet met de daadwerkelijke gebeurtenis. De twitterdata bevatten dus meer 'ruis' dan de P2000-data. Ook hebben de P2000-data een hogere escalatiedrempel dan de twitterdata. Bij een klein schadegeval of beperkte hinder door water op straat wordt de brandweer niet betrokken. Deze gebeurtenissen komen dus niet voor in de P2000-dataset. Bij de twitterberichten noch de P2000-meldingen is het zeker of regen wel de oorzaak was van de wateroverlast. Het kan bijvoorbeeld ook een breuk in een waterleiding zijn. Wateroverlastmeldingen waar regen niet de oorzaak was zijn ook een vorm van ruis. Om deze vorm van ruis in beide datasets te verminderen wordt een filter toegepast. Er wordt een minimumdrempel ingesteld van 10mm neerslag op de dag van de melding. Een dag met minimaal 10mm neerslag wordt door het KNMI gezien als een 'natte dag' [4]. Als er minder dan 10mm is gevallen, is het niet waarschijnlijk dat regen de oorzaak is van de wateroverlast. Als een melding hier niet aan voldoet wordt deze verwijderd uit de dataset. Op deze manier werd 29 procent van de twitterberichten en 10 procent van de P2000-meldingen verwijderd. Dit laat zien dat de twitterberichten meer ruis bevatten. Modellen die zijn getraind op datasets met P2000-meldingen presteren beter dan modellen die zijn getraind op twitterberichten omdat de P2000-dataset minder ruis bevat dan de twitterdataset.

Voorbeelden van situaties waar geen regenwateroverlast optrad

Om het model te trainen zijn ook voorbeelden van de negatieve klasse (geen regenwateroverlast) nodig. Wanneer er geen P2000-melding van regenwateroverlast is, wordt aangenomen dat die overlast er niet is. Ook alle negatieve voorbeelden moeten aan de drempelwaarde van 10mm neerslag voldoen. Zonder toepassing van de drempelwaarde op de negatieve voorbeelden kunnen deze door het model om de verkeerde reden worden geïdentificeerd; het gaat erom het model te leren onderscheiden waarom de ene locatie neerslag boven de drempelwaarde wel kan verwerken en de andere locatie niet. Ondanks dit vereiste zijn er (gelukkig maar) veel meer natte dagen zónder meldingen van regenwateroverlast dan mét. Machine learning-modellen hebben zowel positieve als negatieve voorbeelden nodig om te kunnen leren. Alle regenmetingen zonder melding van wateroverlast in dat

gebied kunnen worden gezien als negatief voorbeeld. Dit resulteert in een dataset met veel meer negatieve dan positieve voorbeelden. Hierdoor kan het model altijd een negatieve voorspelling doen en toch hoog scoren op accuracy (nauwkeurigheid), (zoals beschreven in de vorige sectie). Dit kan natuurlijk deels worden opgelost door ook naar de prestatie-indicatoren precision (precisie) en recall (herroeping) te kijken. Om dit probleem verder te voorkomen is een gebalanceerde dataset nodig. Dat wil zeggen dat met ongeveer evenveel positieve als negatieve voorbeelden. Hiervoor wordt gebruik gemaakt van ‘*subsampling*’ (een deelverzameling van alle voorbeelden). De manier waarop deze *subsampling* plaatsvindt heeft implicaties voor zowel de prestaties als de voorkeuren van het model en daardoor ook voor de betrouwbaarheid ervan. Er worden drie verschillende *subsampling*-methoden gebruikt: 1) willekeurig, 2) gebaseerd op het adressenbestand en 3) regenafhankelijk.

Willekeurige subsamplingmethode

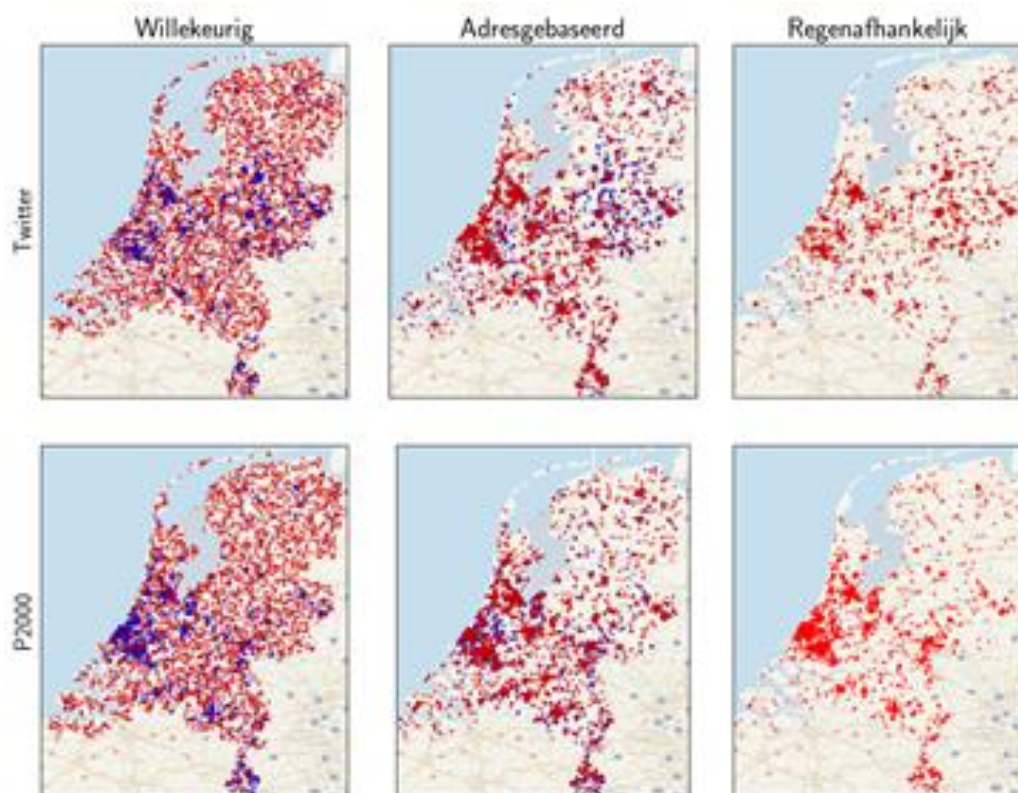
De willekeurige subsamplingmethode kiest volledig willekeurig een aantal locaties en tijdstippen waarop het in Nederland heeft geregend. Hiermee wordt het model getraind om in situaties van regenwateroverlast per ongeluk het verschil tussen landelijk (bv. agrarisch) en stedelijk gebied te leren herkennen. Dit is niet relevant. Regenwateroverlast komt namelijk vooral voor in bebouwd gebied, terwijl het meeste grondgebied in Nederland onbebouwd is.

Subsamplingmethode gebaseerd op een adressenbestand

Met de tweede subsamplingmethode wordt geprobeerd een gelijke mate van verstedelijking in de dataset te realiseren. De basisregistratie adressen en gebouwen (BAG) wordt gebruikt om willekeurige Nederlandse adressen te selecteren. Hier dient dan een willekeurig gekozen adres dat aan de eerdergenoemde voorwaarden voldoet als negatief voorbeeld.

Regenafhankelijke subsamplingmethode

Om het classificatieprobleem nog realistischer te maken wordt een derde, complexere, subsamplingmethode geïntroduceerd. De vraag die deze moet beantwoorden is: waarom heeft een gebied op het ene moment regenwateroverlast en op het andere moment niet? Voor elke melding van regenwateroverlast wordt dan een tegenvoorbeeld genomen, op een ander moment in hetzelfde gebied, wanneer geen regenwateroverlast gemeld is. De hoogtekkaart van het gebied is in dit geval gelijk voor de positieve en negatieve voorbeelden. Toch kan de informatie van de hoogtekkaart in combinatie met de regendata nog steeds informatief zijn. Laaggelegen gebieden kunnen overstromen bij hevige neerslag terwijl in hoger gelegen gebieden geen overlast is.



Afbeelding 1. Locatie van positieve (blauw) en negatieve voorbeelden (rood) voor de zes verschillende datasets

In afbeelding 1 zijn de locaties van elk voorbeeld in de datasets weergegeven. De blauwe stippen zijn voorbeelden waarbij een melding is van regenwateroverlast, de rode stippen zijn negatieve voorbeelden. Bij de regenafhankelijke subsamplingmethode zijn alleen rode stippen te zien. Doordat de positieve en negatieve voorbeelden op dezelfde locatie plaatsvinden, vallen de stippen dus over elkaar. De voorbeelden van regenwateroverlast concentreren zich in en rond stedelijk gebied. Bij de datasets die met willekeurige subsampling zijn gemaakt zijn de negatieve voorbeelden gelijkmatig over het land verspreid. Zo kan het model hoogtepatronen van niet-stedelijk gebied herkennen en op basis daarvan een voorspelling doen. Bij de andere twee subsamplingmethoden komen zowel de positieve als de negatieve voorbeelden uit stedelijk gebied, wat het classificeren moeilijker maakt. Gericht op het voorspellen van regenwateroverlast, passen de adres- en regenafhankelijke subsamplingmethoden beter bij het classificatieprobleem. De verwachting is dat datasets die op deze manier geconstrueerd zijn uitdagender zullen zijn voor het machine learning-model en daardoor een realistischer beeld geven via de prestatie-indicatoren (de eerdergenoemde accuracy, precision en recall).

Beschrijvende eigenschappen

Op basis van de voorgaande subsamplingmethoden kan nu (per subsamplingmethode) een dataset worden geconstrueerd met voorbeelden en tegenvoorbeelden van situaties waar regenwateroverlast optrad. Vervolgens zijn een aantal beschrijvende eigenschappen als invoer nodig, die het model kunnen helpen verklaren waarom er op een locatie op het ene moment wel regenwateroverlast optrad en op het andere niet. Hiervoor worden de terreinhoogte en de gemeten regendata gebruikt.

De terreinhoogtedata komen uit het Actueel Hoogtebestand Nederland (AHN2) [5]. Deze kaarten worden met laseraltimetrie gemaakt in opdracht van Rijkswaterstaat. De kaarten die voor deze experimenten zijn gebruikt hebben een resolutie van vijf meter. De data bestaan uit TIFF-afbeeldingen. Elke afbeelding beslaat een gebied van 5000 x 6250 m. Elke pixel in de afbeelding is een hoogtemeting. Dit is echter te veel voor het model om op te trainen. Voor de locatie van een melding zijn daarom alleen de hoogtewaarden in een gebied van 100m x 100 meter gebruikt. Dit geeft voor elk voorbeeld 400 hoogtewaarden om als kenmerken te gebruiken.

Er wordt één regenkenmerk gebruikt: de totale hoeveelheid regen in het gebied op de dag van de melding. Deze data komen van de rad-nl25-rac-mfbs-01h-dataset van het KNMI [6].

Het gebruik van verschillende databronnen met wateroverlastmeldingen en subsamplingmethoden biedt verschillende datasets om een model op te trainen. Met drie subsamplingmethoden zijn dit drie datasets waar een model op kan worden getraind, waarvan de prestaties empirisch kunnen worden vastgesteld.

Modellen

Auto-sklearn is een AutoML-methode die op een geautomatiseerde manier het beste model voor een machine learning-probleem probeert te vinden. Auto-sklearn kan kiezen uit een portfolio van verschillende modellen, zoals een *Random Forest*, *Support Vector Machine* of *Gaussian Process*. Daarnaast kiest het ook de juiste voorbewerkingsoperatoren en optimaliseert het de hyperparameters van het model (verschillende ontwerpcriteria die van invloed zijn op de prestaties ervan) [3]. In dit artikel worden de resultaten gepresenteerd van de modellen die worden gegenereerd door auto-sklearn.

Uit eerdere experimenten blijkt dat auto-sklearn betere resultaten oplevert dan een enkelvoudig model, maar dat dit model wel meer tijd nodig heeft om te trainen. Dit maakt auto-sklearn een goede keuze voor ingewikkelde problemen. In tegenstelling tot het eerdere onderzoek, waarin auto-encoders werden toegepast [2], wordt in dit onderzoek een lage actieradius (100 bij 100m) rond een voorbeeld gebruikt. Zo kunnen de data zonder tussenkomst van auto-encoders direct worden ingevoerd in het auto-sklearn-model.

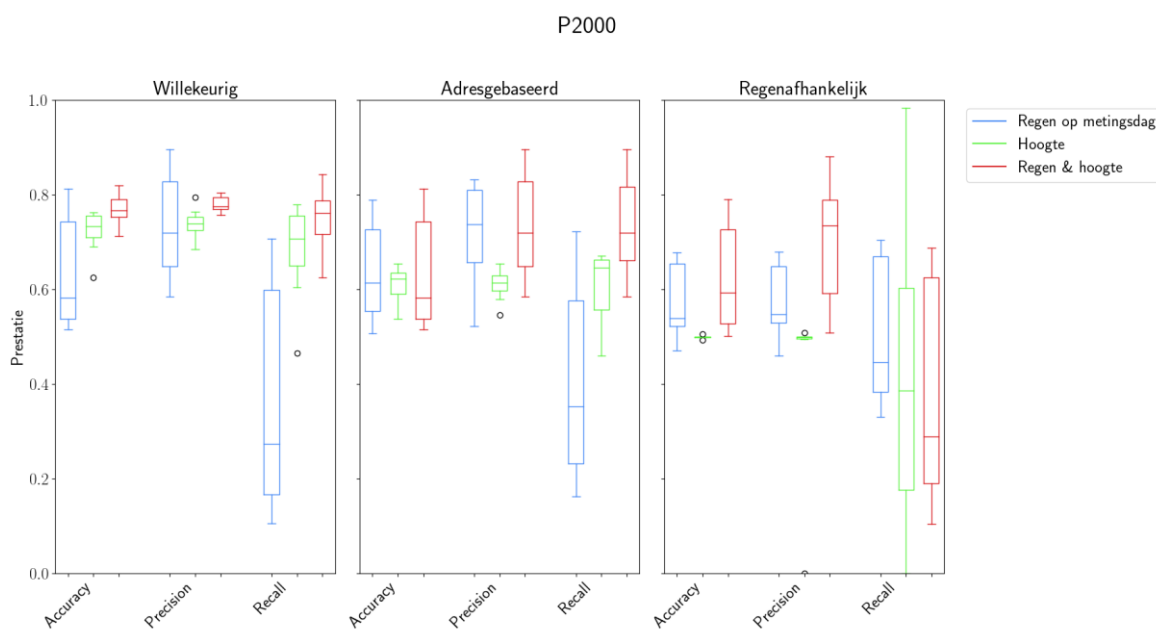
Resultaten

Aan de hand van 'accuracy', 'precision' en 'recall' worden de prestaties gemeten op basis van data die het model nog niet eerder heeft gezien. Aangezien de dataset gebalanceerd is (evenveel positieve als negatieve voorbeelden) zou een relevante baseline een accuracy-score van 50% halen. Om een relevant resultaat te behalen moeten de gepresenteerde modellen in ieder geval beter presteren dan deze baseline. Daarnaast wordt in dit onderzoek geprobeerd de invloed van de verschillende databronnen te kwantificeren.

Afbeelding 2 visualiseert de resultaten met een boxplot. De resultaten worden opgesplitst in de drie verschillende subsamplingmethoden. Voor iedere subsampling methode wordt het experiment tien maal herhaald (de exacte methode is 10-voudige cross-validation). De boxplots geven de distributie van de prestaties van het model aan. Daarnaast worden de kwartielen en uitschieters (bollen) getoond. In blauw zijn de resultaten wanneer alleen de hoeveelheid regen wordt gebruikt te zien. In het groen de resultaten wanneer alleen de hoogtkaart wordt gebruikt en in het rood de resultaten wanneer beide databronnen worden gecombineerd.

Tijdens de experimenten leverde een model getraind met een dataset met willekeurige subsampling de hoogste *accuracy* (gemiddeld 77%). Dit betekent dat het model in 77% van de gevallen de juiste voorspelling doet. Dit is hoger dan de gemiddelde *accuracy* van modellen die zijn getraind met datasets op basis van de adres- of regenafhankelijke subsamplingmethode. Deze modellen behalen een *accuracy* van respectievelijk 69% en 58%. Dit hoeft niet te betekenen dat deze modellen minder goed zijn. In dit geval wordt het probleem dat het model moet oplossen steeds moeilijker. Het is aannemelijk dat laatstgenoemde subsamplingmethoden een realistischer beeld van de prestaties, en daardoor toepasbaarheid in de praktijk, van het model geven.

Wanneer gekeken wordt naar de andere vormen van invoerdata, is een logisch patroon zichtbaar. Het model heeft duidelijk baat bij zowel regen- als hoogtedata. De rode boxplot ligt consequent boven de blauwe. Dit betekent dat de *accuracy* bij het gebruik van hoogtekenmerken hoger is. Dit suggereert dat regenwateroverlast vooral afhankelijk is van terreinhoogte. Toch houdt het model bij het voorspellen van regenwateroverlast een voorkeur voor de regenkenmerken. Hiervoor zijn een aantal verklaringen. Ten eerste kan de relatie tussen neerslag en regenwateroverlast sterker zijn dan de relatie tussen terreinhoogte en regenwateroverlast. Daarnaast krijgt het model 400 hoogtekenmerken en slechts één regenkenmerk. Dan is het moeilijker om bij de hoogtekenmerken te generaliseren. Als laatste is de keuze voor een gebied van 100m x 100m voor een melding niet gemaakt op basis van optimalisatie. Dit betekent dat deze kenmerken nog geoptimaliseerd kunnen worden.



Afbeelding 2. Boxplot van prestaties van auto-sklearn op datasets met P2000-meldingen als dataset met wateroverlast meldingen

Conclusies

Voor de databron met wateroverlastmeldingen zijn twee opties onderzocht. Een dataset met twitterberichten waarin regenwateroverlast genoemd is en een dataset met noodmeldingen uit het P2000-netwerk. De P2000-meldingen bevatten minder ruis dan de twitterberichten en zijn dus beter bruikbaar als databron. Als de prestaties van de modellen die de hoogtekenmerken wel en niet gebruiken met elkaar worden vergeleken, geeft het gebruik van hoogtekenmerken betere resultaten. Hieruit wordt geconcludeerd dat de hoogtedata een nuttig voorspellend kenmerk is.

Elke subsamplingmethode genereert andere negatieve voorbeelden en dus een andere dataset. De manier waarop deze worden gekozen heeft invloed op de moeilijkheid van het probleem. Als zodanig is het belangrijk de modellen gemaakt met de verschillende subsamplingmethoden individueel te evalueren. Het model getraind op de dataset die met de willekeurige methode is gegenereerd behaalt de beste *accuracy*. Deze methode geeft een dataset waarbij de negatieve voorbeelden vooral buiten stedelijk gebied liggen. Deze gebieden hebben andere hoogtekarakteristieken, waardoor het voor het model makkelijker is om een juiste voorspelling te doen. De negatieve voorbeelden die met methodes op basis van adressen en gelijke locaties worden gegenereerd liggen wel in dezelfde gebieden als de positieve voorbeelden. Dat maakt deze subsamplingmethoden beter geschikt voor het classificatieprobleem.

Tijdens dit onderzoek zijn *machine learning*-technieken toegepast op een relatief nieuw domein, namelijk stedelijk waterbeheer. Er is een dataset gekozen om gebeurtenissen van regenwateroverlast te benaderen, hoogtekarten en regenmetingen om te zetten in modelkenmerken en verschillende subsamplingmethoden voor negatieve voorbeelden te implementeren. Aan de hand van dit domein is gedemonstreerd dat machine learning-technieken kunnen worden toegepast om regenwateroverlast te voorspellen. Ook werd aangetoond dat terreinhoogtedata gebruikt kunnen worden bij het voorspellen van regenwateroverlast. Deze methode kan verder worden ontwikkeld en uiteindelijk worden gebruikt door gemeenten en waterschappen bij het identificeren van kwetsbare gebieden.

Verantwoording

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

Referenties

1. Tesema, D., Birhanu, B. (2020). 'A review of flood modeling methods for urban pluvial flood application'. *Modeling Earth Systems and Environment*, 6, 09.
2. Lamers, C., Rijn, J. van, Ton Beenen, T. (2020). 'Data science-technieken voor regenwateroverlast in stedelijk gebied'. *H2O-Online*, 2 oktober 2020. <https://www.h2owaternetwerk.nl/vakartikelen/data-science-technieken-voor-regenwateroverlast-in-stedelijk-gebied>
3. Feurer, M. et al. (2015). 'Efficient and robust automated machine learning'. *Advances in Neural Information Processing Systems* 28, pages 2962–2970.
4. KNMI (2021). *KNMI - Regenintensiteit, 2021*. <https://www.knmi.nl/kennis-en-datacentrum/uitleg/regenintensiteit>.
5. Rijkswaterstaat (2012). *Actueel Hoogtebestand Nederland 2 (AHN2)*
6. Overeem, A. (2021). *Precipitation - 1 hour precipitation accumulations from climatological gauge-adjusted radar dataset for The Netherlands (1 km) in KNMI HDF5 format - KNMI Data Platform*.