BTO 2016.007 | Januari 2016

BTO rapport

Datamining voor assetmanagement – inventarisatie en voorbeelden uit de watersector

KWR Watercycle Research Institute

Voorwoord

Binnen het BTO Thema Assetmanagement is het project Datamining voor Assetmanagement uitgevoerd. Het voorliggende rapport beschrijft de voornaamste kenmerken, mogelijkheden en beperkingen van datamining. Tevens geeft dit rapport de kennisbehoefte weer op het gebied van assetmanagement en datamining bij Nederlandse drinkwaterbedrijven.

Parallel aan dit BTO-project zijn twee TKI-projecten uitgevoerd bij Brabant Water en Vitens, respectievelijk DiAMANT-Water (DataMANagement van de Toekomst-Water) en BWD2SWG (from Big Water Data Towards Smart Water Grids). Deze TKI-projecten hebben zich gericht op de ontwikkeling van bedrijfsspecifieke data-analyses en tools op basis van datamining. Ervaringen die in de TKI-projecten zijn opgedaan zijn als casussen in voorliggend rapport uitgewerkt.

De auteurs willen graag Arno Knobbe (Leiden Institute of Advanced Computer Science) bedanken voor zijn tussentijdse kwaliteitsborging en tips voor dit rapport. Ook de collega's Joost van Summeren, Bas Wols en Bernard Raterman, de contactpersonen bij de TKI-projecten en uiteraard alle leden van de themagroep Assetmanagement worden bedankt voor hun praktische inbreng en advies.

BTO

Datamining voor assetmanagement – inventarisatie en voorbeelden uit de watersector

BTO 2016.007 | Januari 2016

Opdrachtnummer
400554-070

Projectmanager
drs. P.G.G. (Nellie) Slaats

Opdrachtgever
BTO- Thematisch onderzoek - Assetmanagement

Kwaliteitsborgers
dr. ir. E.J.M. (Mirjam) Blokker

Auteurs
ir. E. (Erwin) Vonk, dr. ir. D. (Dirk) Vries

Verzonden aan
Dit rapport is verspreid onder BTO-participanten en is openbaar.

Jaar van publicatie
2016

Meer informatie

T +31 30 6069 547
E erwin.vonk@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

T +31 (0)30 60 69 511
F +31 (0)30 60 61 165
E info@kwrwater.nl
I www.kwrwater.nl

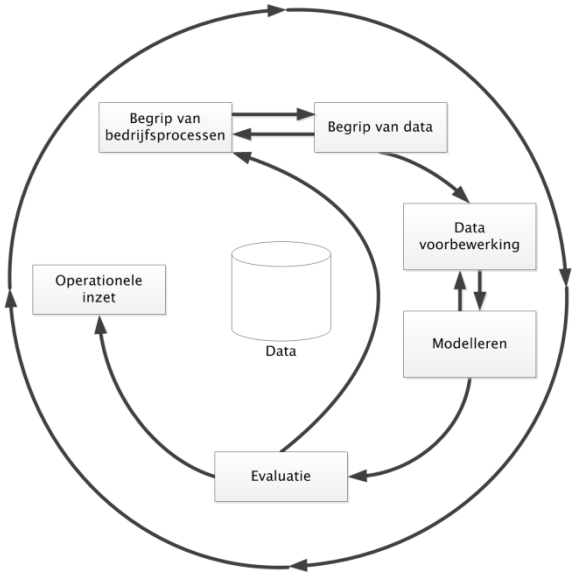
KWR Watercycle Research Institute

BTO *Managementsamenvatting*

Datamining biedt volop kansen om het assetmanagement van drinkwaterbedrijven te verbeteren

Auteur(s) Erwin Vonk MSc en dr.ir.Dirk Vries

Met een brede inventarisatie van assetmanagement-kennisvragen, een literatuurstudie naar datamining en de eerste praktijkervaringen uit twee TKI-projecten is voor drinkwaterbedrijven een eerste, voorzichtige stap gezet richting een datagedreven bedrijfsvoering. Drinkwaterbedrijven kunnen de verzamelde kennis inzetten bij beslissingen over het wel of niet inzetten van datamining en datagedreven analysetechnieken om hun operationeel assetmanagement te verbeteren. Met de huidige beperkingen aan de (kwaliteit en kwantiteit van) beschikbare data is voorzichtigheid geboden ten aanzien van datamining-ambities. Data-opschoonacties en data-kwaliteitscontroles binnen de bedrijven kunnen de beperkingen verkleinen.



Figuur 1. Het CRISP-DM model voor knowledge discovery (IBM, 2011). Dit procesmodel wordt doorgaans gehanteerd bij de uitvoering van datamining-projecten. Daarbij staat de beschikbare data centraal en is samenwerking nodig met de bronhouder van de data om tot een juist begrip van de bedrijfsprocessen en de data zelf te komen. Data voorbewerking en het eigenlijke modelleren zijn stappen die iteratief worden uitgevoerd. Na evaluatie met de opdrachtgever kan besloten worden om een resulterend datagedreven model operationeel in te zetten. De buitenste pijlen in deze figuur benadrukken het iteratieve karakter van datamining.

Belang: sectorbreed inventariseren waar de kansen voor datamining liggen

Datamining (het zoeken naar statistische verbanden in databases) is een stap in “Knowledge Discovery in Databases (KDD)”. Beide worden reeds omarmd in de marketing, medische zorg, ICT en financiële sector, maar de implementatie in de drinkwatersector is vooralsnog beperkt. Dit, terwijl het potentieel nieuwe mogelijkheden biedt waarmee de drinkwatersector zijn assetmanagement kan verbeteren. Waterbedrijven willen daarom meer inzicht in de meerwaarde die datagedreven analysemethodieken voor hen kunnen opleveren en zoeken een kennisbasis die hen houvast kan geven bij beslissingen over het wel of niet inzetten van KDD in assetmanagement.

Aanpak: vraag, aanbod en eerste praktijkervaringen verzamelen

Er is in eerste instantie een literatuurstudie uitgevoerd naar de voornaamste kenmerken, mogelijkheden en beperkingen van datamining. Daarnaast is de ‘kennisbehoefte’ in overleg met assetmanagers bij drinkwaterbedrijven in kaart gebracht. Tijdens een workshop bij KWR zijn de kennisvragen verder geprioriteerd op basis van urgentie en belang. Zo zijn zowel vraag als aanbod rondom datamining

geïnventariseerd. Die inventarisatie is gebruikt om inzicht te krijgen in de meest kansrijke toepassingsgebieden voor datamining ten behoeve van waterinfrastructuur assetmanagement. Daarnaast hebben resultaten vanuit twee TKI-projecten in dit werk model gestaan voor de inzet van datamining/KDD bij assetmanagementvraagstukken.

Resultaten: de kennisfundering voor toekomstige datamining-projecten

Datamining staat niet op zichzelf, maar is een onderdeel van de bredere 'knowledge discovery' procedure, waarbij men tracht bruikbare kennis uit dataverzamelingen te onttrekken. Datamining kan niet los worden gezien van expertsystemen die ervoor zorgen dat de onttrokken informatie uit data wordt aangeleverd. Ondanks de vele kansen die datamining biedt, is de methodiek ook gevoelig voor onjuist gebruik. Resultaten dienen door een opdrachtgever altijd kritisch bekeken te worden met het oog op veelvoorkomende valkuilen, zoals gebruik van niet-representatieve datasets, overfitting, 'data dredging' (correlaties zoeken zonder deze te valideren) en niet-causaliteit.

Bij het assetmanagement van drinkwaterbedrijven blijkt behoefte te zijn aan kennis van zowel de operationele prestatie van een asset als de actuele conditie en de storingsrisico's. Alle geïnventariseerde kennisvragen vallen binnen een van deze drie categorieën. Daarbij is nog een tweede dimensie in de kennisvragen te onderscheiden, namelijk die van urgentie in handelen. Deze as kan worden ingedeeld met aan de ene zijde de behoefte aan 'real-time' inzicht bij gebruik binnen het primaire bedrijfsproces, en aan de andere zijde een 'periodieke' raadpleging van kennisbronnen (expertsystemen) bij tragere bedrijfsprocessen, zoals bij het opstellen van onderhouds- en investeringsplannen. In de organisatorische context van een drinkwaterbedrijf kan knowledge discovery projectmatig georganiseerd worden, met als doel permanent expertsystemen te voeden.

Aangezien datamining geheel datagedreven is, zal een verkregen model hooguit zo goed zijn als de data waarmee het gemaakt is. Veelvoorkomende problemen omtrent databeschikbaarheid zijn terug te voeren tot vier categorieën: ontbrekende labels, te weinig 'events', een te korte meetperiode of een te lage meetfrequentie. Datakwaliteit is een ander aspect, dat betrekking heeft op mogelijke onjuistheden in de datasets, variërend van falende sensoren tot menselijke typfouten.

Voornaamste lessen uit reeds uitgevoerde datamining projecten zijn (1) het belang van feature engineering (een bestaande dataset verrijken met afgeleide parameters uit bijvoorbeeld modelsimulaties of berekeningen), (2) samenwerking met vakspecialisten op het gebied van waterinfrastructuur en operationele processen en ten slotte (3) de beperkte hoeveelheid beschikbare data. Van 'big data' is nog geen sprake bij de waterbedrijven. Toch lijkt er door een gestage toename aan sensoren in het leidingnet, slimme meters, groter wordende databases met meetgegevens en storingsregistraties meer en meer mogelijkheden te komen om middels datamining relevante vraagstukken voor de drinkwatersector te beantwoorden.

Implementatie: datamining vraagt gestroomlijnd datamanagement

Met de brede inventarisatie van assetmanagement-kennisvragen, de literatuurstudie naar datamining en met de eerste praktijkervaringen uit twee TKI-projecten is voor drinkwaterbedrijven een eerste, voorzichtige stap gezet richting een datagedreven bedrijfsvoering. Huidige beperkingen aan de data rechtvaardigen voorzichtigheid ten aanzien van datamining-ambities en nopen tot ondersteuning van data-opschoonacties en data-kwaliteitscontroles binnen de bedrijven. Maar met deze aspecten in het achterhoofd, kan worden toegewerkt naar meer datamining-successen.

Rapport

Dit onderzoek is beschreven in rapport *Datamining voor assetmanagement – inventarisatie en voorbeelden uit de watersector* (BTO 2016.007).

Meer informatie
Erwin Vonk, MSc
T (030) 60 69 547
E erwin.vonk@kwrwater.nl

KWR
PO Box 1072
3430 BB Nieuwegein
The Netherlands

KWR Watercycle Research Institute

Begrippen en definities

Attribuut
Eigenschap van een instantie. In de praktijk is dit meestal een kolom uit een database, waarin voor elke record een bepaalde meetwaarde is vastgelegd.

Conditie
De gezondheidstoestand van een bepaalde asset, uitgedrukt ten opzichte van een vooraf gedefinieerde norm ten aanzien van minimale assetkwaliteit.

Database
Gegevensverzameling.

Datamining
Het gericht zoeken naar statistische verbanden in databases.

Expertsysteem
Een computersysteem dat op basis van ingevoerde kennisregels binnen een bepaald gebied oplossingen biedt. Het systeem functioneert daardoor als een digitaal equivalent van een menselijke 'expert' binnen een organisatie.

Feature engineering
Procedure waarbij een bestaande dataset verrijkt wordt met afgeleide parameters die verkregen worden middels bijvoorbeeld modelsimulaties of berekeningen op bestaande attributen. Het gaat dus in essentie om het opwerken van ruwe data naar bruikbare invoerparameters door het gebruik van domeinkennis.

Instantie
Een enkele eenheid data uit een database, in de praktijk meestal een rij uit een tabel (ook wel 'record' of 'tupel' genoemd).

Knowledge discovery
Voluit vaak 'Knowledge Discovery in Databases' (KDD) genoemd. Dit is de algemene benaming voor het proces waarbij men tracht bruikbare kennis uit dataverzamelingen te onttrekken. Datamining is een van de stappen in dit proces.

Machine learning (machinaal leren)
Onderzoeksveld (grenzend aan kunstmatige intelligentie, statistiek en optimalisatie) dat zich bezig houdt met de ontwikkeling van algoritmes en technieken die computers in staat stellen zelf dingen te leren.

Operationele prestatie
De mate van 'goed' functioneren van een bepaalde asset ten opzichte van de gewenste mate van functioneren. Om dit te meten worden doorgaans bedrijfsspecifieke indicatoren geformuleerd. Hierbij kan gedacht worden aan bijvoorbeeld de energie-efficiency van een pomp, de verwijderingsefficiency van een zuiveringsstap of de CO_2-uitstoot van een voertuig.

Overfitting

Situatie waarbij een algoritme niet meer goed kan omgaan met nieuwe data, omdat de structuur en modelparameters tijdens het trainen volledig geoptimaliseerd zijn om de trainingsdata (inclusief de daarin aanwezig ruis en irrelevante gegevens) te reproduceren.

Storing

Het acute falen van een asset, waarbij de operationele prestaties een vooraf gedefinieerd prestatieniveau onderschrijden.

Training

De eerste stap in het leerproces van een machine learning algoritme, waarbij de modelparameters worden gekalibreerd. Het doel hiervan is om de door het model gegenereerde uitvoer zo goed mogelijk overeen te laten komen met de werkelijkheid (gewenste uitvoer).

Validatie

Tweede stap in het leerproces van een machine learning algoritme. Hierin wordt gecontroleerd of de modelinstellingen, zoals geleerd tijdens de training, ook goed presteren op een onafhankelijke dataset (de zogenaamde validatieset).

Inhoud

1 Inleiding

1.1 Aanleiding

Bij de bedrijfsvoering van drinkwaterbedrijven wordt doorgaans veel data gegenereerd, zoals procesdata (drukmetingen, niveaumetingen, waterkwaliteitsmetingen, et cetera), registraties (tijdsbesteding, storingen, waarnemingen, onderhoud, klant- en factuurgegevens) en asset-gegevens (ligging, conditie, vervangingskosten, et cetera). Hoewel deze gegevens in eerste instantie veelal een concreet en operationeel doel dienen, belandt een groot deel uiteindelijk in databases. In de praktijk worden veel gegevens niet of nauwelijks opgevraagd. Echter, door het slim combineren van data uit dergelijke, vaak verschillende, databases kan mogelijk waardevolle informatie gedestilleerd worden, informatie die niet bemachtigd had kunnen worden uit de bestaande databronnen los van elkaar. Deze informatie kan op zijn beurt leiden tot nieuwe inzichten, het ontwikkelen van strategische voordelen of stroomlijnen van interne procedures rondom asset management. Hierbij staat het begrip 'datamining' centraal, al dan niet in de bredere context van het zogenaamde 'Knowledge Discovery in Databases' (KDD).

Datamining wordt gedefinieerd als het zoeken naar patronen of correlaties in grote datasets met als doel de hieruit verkregen informatie op toegankelijke wijze aan de gebruiker te presenteren. Het wordt vaak spreekwoordelijk 'schatgraven in grote databases' genoemd, daar datamining inmiddels in verschillende bedrijfstakken op commerciële basis wordt ingezet en meerwaarde oplevert voor de efficiency van diverse bedrijfsprocessen. Zo brengen grote winkelketens het aankoopgedrag van consumenten ermee in kaart, worden in de sterrenkunde nieuwe ontdekkingen gedaan door de juiste informatie te destilleren uit radiometrische metingen, identificeren banken er verdachte financiële transacties mee en worden fouten in de productie ermee opgespoord door de microchip-industrie (Sagiroglu & Sinanc, 2013).

In een soortgelijke context als datamining wordt ook vaak de term 'big data' genoemd. Het verschil tussen deze begrippen is dat datamining refereert naar de procedure ('het zoeken naar patronen in databases'), terwijl big data betrekking heeft op de datasets zelf (Sagiroglu & Sinanc, 2013). In beginsel is big data niet zozeer een kans geweest voor veel ondernemingen, maar meer een praktisch probleem. Met name internetondernemingen hebben recentelijk te maken gekregen met dusdanig extreem grote datastromen dat traditionele opslag- en verwerkingstechnologieën niet meer volstaan. Dergelijke datastromen worden in de big data definitie gekenmerkt door een drietal aspecten: een groot volume (enkele gigabytes tot terabytes per uur die over en weer verzonden worden tussen computers), hoge snelheid (vrijwel real-time veranderingen in, of toevoegingen aan de datasets) en grote interne variëteit (de data bestaat uit een ongestructureerde combinatie van bijvoorbeeld tekstberichten, tabellen, getallenstromen en afbeeldingen). Momenteel is de heersende opvatting dat big data geen probleem hoeft te zijn, maar dat het in combinatie met datamining (het zogenaamde 'datastream mining') juist kansen biedt om waardevolle informatie te winnen en nieuwe inzichten te verkrijgen. Meer informatie van wat big data kan betekenen voor een waterbedrijf staat beschreven in de trends- en effectenstudie Big Data (KWR, 2013).

Ook voor de drinkwatersector lijkt datamining kansrijk te zijn om bedrijfsprocessen te verbeteren. Onder assetmanagers bij drinkwaterbedrijven bestaat een sterke behoefte aan het verkrijgen van meer inzicht in de conditie en faalkansen van diverse typen assets. Voor een aantal faalmechanismen worden reeds methoden voor herkenning gebruikt op basis van bestaande data (metingen en waarnemingen). Voorbeelden zijn de detectie van grote lekken middels het meten van volumestroom en druk, filterverstopping door meting van hydraulische weerstand of waterstanden. Bij vraagstukken waarbij echter nog geen duidelijk inzicht bestaat in de fysische processen die hierop van invloed zijn, zou datamining ingezet kunnen worden. Mogelijke toepassingen van datamining zijn bijvoorbeeld het herkennen en detecteren van faalcondities, verbanden met levenscycluskosten ontdekken en het inschatten van de slijtage van pompen.

1.2 Onderzoeksvragen

In de huidige situatie hebben assetmanagers bij drinkwaterbedrijven behoefte aan meer en vooral ook beter inzicht in de conditie van assets in de infrastructuur. Daarnaast is het aannemelijk dat diverse databases binnen de organisatie reeds data bevatten die door middel van een data-analyse (deels) antwoord kunnen geven op de kennisvragen. Drinkwaterbedrijven beheren veel data die, indien grondig geanalyseerd, van meerwaarde kunnen zijn voor kennis-gedreven assetmanagement van waterinfrastructuur.

Het doel van dit onderzoek is om *te inventariseren welke mogelijkheden datamining - binnen de bredere context van Knowledge Discovery in Databases (KDD), machine learning en expertsystemen - biedt om datasets van drinkwaterbedrijven om te vormen tot waardevolle kennis waarmee assetmanagement ondersteund kan worden*. Daarbij staan de volgende vragen centraal:

1. Wat zijn de voornaamste kenmerken van datamining?
2. Hoe kan datamining het assetmanagement van drinkwaterbedrijven ondersteunen?
3. Wat zijn de vereisten aan databeschikbaarheid en datakwaliteit ten aanzien van datamining?
4. Wat zijn de ervaringen tot nu toe met datamining binnen de drinkwatersector?

1.3 Leeswijzer

Dit rapport is gestructureerd op basis van de onderzoeksvragen. In elk hoofdstuk staat een onderzoeksvraag centraal, met in hoofdstuk 2 een overzicht van het datamining-proces. Vervolgens wordt in hoofdstuk 3 ingegaan op de specifieke kennisbehoefte van drinkwaterbedrijven ten aanzien van verbetering van het assetmanagement. Hoofdstuk 4 geeft een overzicht van vereisten ten aanzien van databeschikbaarheid en –kwaliteit, waarna tot slot in hoofdstuk 5 de conclusies en aanbevelingen van dit onderzoek volgen.

2 Datamining - schatgraven in databases

In dit hoofdstuk staat de eerste onderzoeksvraag centraal, waarbij uitleg gegeven wordt bij het begrip datamining. Als eerste komt in dit hoofdstuk de terminologie rondom data en datasets aan bod (paragraaf 2.1), waarna het proces rondom datamining ('knowledge discovery') wordt toegelicht (paragraaf 2.2). In paragraaf 2.3 komen de doelstellingen en algoritmes bij datamining aan bod. Tot slot wordt aan het einde van dit hoofdstuk dieper ingegaan op de valkuilen van datamining, waar men in de praktijk beducht op dient te zijn (paragraaf 2.4).

2.1 Datasets

Datasets bestaan doorgaans uit een zekere hoeveelheid metingen of registraties, welke als aparte records worden weggeschreven in een database. Bij datamining worden dit de *instanties* van een dataset genoemd. Elke instantie kan een, twee of meerdere zogenaamde *attributen* hebben (resulterend in univariate, bivariate of multivariate datasets). Daarnaast kan een kan onderscheid gemaakt worden tussen cross-sectionele datasets en tijdreeksen. Cross-sectionele data is niet tijdsgebonden, terwijl tijdreeksen dit wel zijn.

> Voorbeeld - Een dataset bevat een meetreeks met 200 metingen (instanties). Het instrument waarmee wordt gemeten is bivariaat, namelijk volumestroom en temperatuur. In de dataset worden volumestroom en temperatuur gezien als attributen van elke instantie.

Doorgaans wordt data opgeslagen in conventionele databases. Met de hedendaagse ontwikkelingen ontstaan er echter steeds vaker situaties waarbij datastromen dusdanig snel en groot zijn dat het niet meer mogelijk is om alles op te slaan (de zogenaamde 'big data'). Om uit dergelijke datastromen toch langetermijn trends te kunnen ontdekken is het noodzakelijk om deze op een zogenaamde 'streaming' manier te verwerken. Bij datastream mining wordt binnenkomende data (bijvoorbeeld uit sensornetwerken) gelijk bij ontvangst geanalyseerd, enige tijd in het werkgeheugen van een computer vastgehouden en enkele tijdstappen later uit het geheugen gewist.

2.2 Knowledge discovery - van data naar kennis

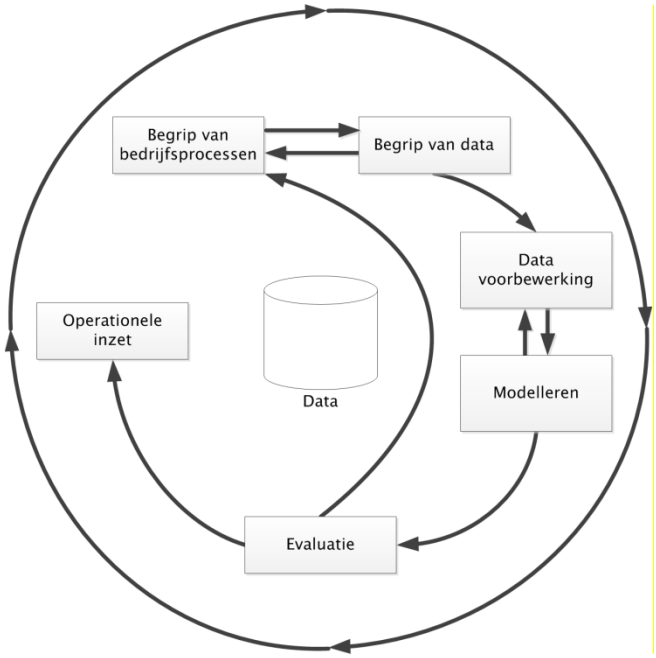
Datamining is niet een op zichzelf staande procedure en wordt daarom vaak beschouwd als onderdeel van het bredere proces Knowledge Discovery in Databases (KDD). In dit proces wordt ruwe data stapsgewijs getransformeerd naar bruikbare informatie. Het gehele proces bestaat uit de volgende stappen:

1. **Begrip van bedrijfsprocessen**
   Het vaststellen van de bedrijfsdoelen, relevant voor de data-analyse, is doorgaans de eerste stap in het datamining proces. Deze doelen worden daarna vertaald naar concrete succescriteria voor latere evaluatie van het datamining proces.

2. **Begrip van datasets**
   Een succesvol project vereist diepgaand begrip van de beschikbare databronnen en de fysische betekenis daarvan. In deze stap is overleg met de bronhouder van de data van belang om ervoor te zorgen dat er geen misverstanden bestaan over de eenheden, meetmethodieken en betekenis van de grootheden. Visualisatie is tevens in deze stap een krachtige techniek om op het oog correlaties en statistische verdeling van de data te beoordelen. Het gebruik van visualisaties vereenvoudigt de communicatie tussen opdrachtnemer en opdrachtgever en versterkt het selecteren van attributen en formuleren van afgeleide attributen (het zogenaamde '*feature engineering*', zie stap 3c).
3. **Voorbewerking van data**
   a. **Selectie van datasets**
      Op basis van vooraf gestelde hypotheses en onderliggende kennis over een bepaald onderwerp of proces worden alle datasets geselecteerd waarvan het vermoeden bestaat dat die relevante data bevatten.
   b. **Opschonen**
      De geselecteerde datasets dienen vervolgens ontdaan te worden van alle corrupte gegevens (incomplete of inconsistente metingen). Tevens worden doorgaans in deze stap ook de extreme waarden (uitbijters) uit de dataset verwijderd. Hier zijn diverse wetenschappelijke methodieken voor, die in dit rapport verder niet toegelicht zullen worden.
   c. **Transformatie en feature engineering**
      Het kan wenselijk zijn om fijnmazige data te aggregeren tot een hoger schaal- of abstractieniveau. Er kan gedacht worden aan tijdreeksen, waarvan uurlijkse metingen worden samengevoegd tot dagtotalen, of ruimtelijke data, waarvan de resolutie wordt verlaagd van vierkante meters naar vierkante kilometers. Verder wordt in deze stap de data ook genormaliseerd, waarbij alle waarden getransformeerd worden naar een vooraf gedefinieerd interval. Op basis van bestaande attributen kunnen eventueel ook extra of afgeleide attributen toegevoegd worden aan de dataset; iets dat '*feature engineering*' genoemd wordt. Het gaat dan om het verrijken van een dataset met afgeleide parameters die door bijvoorbeeld modelsimulaties bepaald worden. Tot slot is het in bepaalde gevallen ook wenselijk om 'smoothing' toe te passen, waarbij ruis zoveel mogelijk wordt verwijderd.
   d. **Volumereductie**
      Instanties in datasets kunnen gecorreleerde attributen bevatten die feitelijk dezelfde informatie verschaffen (denk aan bijvoorbeeld leeftijd en geboortedatum als attributen van een persoon). Dergelijke dubbele voorspellers worden zoveel mogelijk verwijderd, of vervangen door een attribuut dat informatie uit beide voorspellers combineert in één enkele parameter. Daarnaast kan de hoeveelheid data ook verkleind worden door bijvoorbeeld discretisatie van continue waarden.
4. **Modelleren**
   Het extraheren van de informatie (het vinden van correlaties en patronen) uit de dataset door deze vast te leggen in enkele modelparameters en een modelstructuur met behulp van machine learning algoritmes. Daarbij wordt een algoritme in eerste instantie getraind en (kruis)gevalideerd op een deel van de dataset en vervolgens getest (extern gevalideerd) op het resterende deel.

5. **Evaluatie**
   Hiertoe worden de resultaten doorgaans vertaald naar tabellen of diagrammen. In een evaluatiestap wordt beoordeeld of de gevonden correlaties overtuigend genoeg zijn voor gebruik in de praktijk, aan de hand van in stap 1 gedefinieerde (informatie)criteria.
6. **Eventueel: operationele inzet**
   Bij een positieve evaluatie kan besloten worden om de machine learning algoritmes, die getraind en gevalideerd zijn in de datamining fase, operationeel in te zetten. Daarbij dienen dan ook de datatransformatie- en reductietechnieken geautomatiseerd te worden. In de praktijk wordt de ontwikkelde data-analysemethodiek vaak opgenomen als onderdeel van een zogenaamd beslissings-ondersteunend systeem (ook wel expertsysteem genoemd).

Bovenstaande stappen zijn onderdelen in het zogenaamde Cross Industry Standard Process for Data Mining model (CRISP-DM). Uitgangspunt van dit model is dat knowledge discovery wordt weergegeven als een cyclisch proces (Figuur 1).



Figuur 1: Het CRISP-DM model voor knowledge discovery (IBM, 2011). *De buitenste pijlen in deze figuur benadrukken dat het om een iteratief proces gaat.*

2.3 Datamining

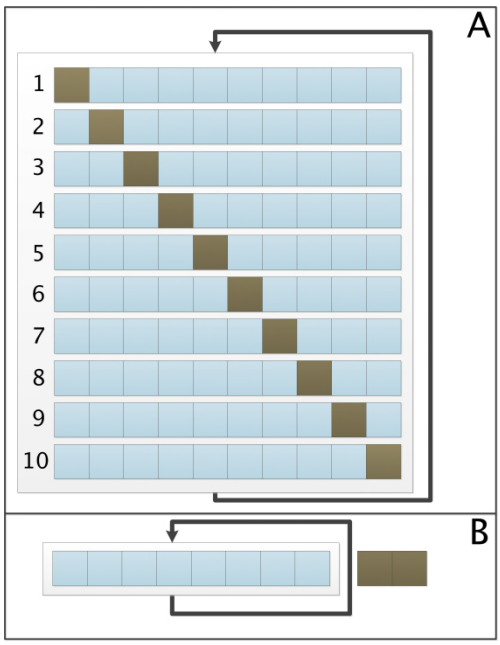
2.3.1 Machine learning

Voor het daadwerkelijk zoeken naar verbanden tussen verschillende attributen in datasets worden zogenaamde machine learning algoritmes gebruikt. Dit zijn statistische zoektechnieken waarmee op een geautomatiseerde manier een modelstructuur ontwikkeld en gekalibreerd wordt. Een machine learning algoritme moet in eerste instantie getraind worden op een zekere hoeveelheid data. Het algoritme ‘leert’ dan op basis van de aangeleverde trainingsdata de verbanden uit deze dataset te halen. Er zijn diverse manieren om een machine learning algoritme te trainen, maar de meest voorkomende methoden zijn het zogenaamde ‘supervised learning’ en ‘unsupervised learning’. Bij ‘supervised’ machinaal leren zijn instanties in de dataset vooraf gedefinieerd door middel van meta-informatie

(gelabeld). Dit betekent dat het algoritme elke instantie in een dataset dient te projecteren tot een gelabelde uitkomst. Bij 'unsupervised' machinaal leren wordt het algoritme 'vrij' gelaten op data die niet zijn voorzien van meta-informatie, of waarvan de meta-informatie wordt genegeerd (niet-gelabeld).

Het ontwikkelen van een geschikt machine learning algoritme verloopt in twee stappen: modelselectie en modelbeoordeling. Het doel van de modelselectie is om te komen tot een juiste algoritmekeuze. Kortom: het vinden van het optimale representatiemodel, evaluatiemaat en optimalisatietechniek; én het juist afstellen van de tuning-parameters voor deze combinatie van algoritme-componenten. Om verschillende algoritmes onderling te vergelijken worden deze getraind op een deel van de totale dataset en vervolgens *gevalideerd* op een ander deel. Nadeel hiervan is uiteraard dat een deel van de data 'verloren' gaat aan validatie, terwijl de data juist zo kostbaar kan zijn. Daarom wordt vaak k-voudige kruisvalidatie ('k-fold cross validation') toegepast in deze stap. Bij kruisvalidatie wordt de totale dataset in een aantal subsets onderverdeeld, waarbij het algoritme telkens op één van de subsets gevalideerd wordt en alle andere subsets al roulerend gebruikt worden om op te trainen. De gemiddelde validatiescore voor elke individuele algoritme-configuratie is vervolgens leidend in het vaststellen van het optimale model.

Uiteindelijk is het na de selectiefase belangrijk om het gekozen algoritme te beoordelen op voorspellingskwaliteit. Dit, om te testen of het algoritme de verbanden daadwerkelijk op een juiste manier geleerd heeft en ook in nieuwe gevallen in staat is om het juiste verband te voorspellen. Daartoe wordt het geselecteerde algoritme doorgaans getraind op de volledige kruisvalidatieset uit de selectiefase, en vervolgens getest op een nieuw stuk van de dataset die het algoritme nog niet eerder gezien heeft. Dit wordt de 'split sample test' genoemd.



Figuur 2: Schematische weergave van 10-voudige kruisvalidatie (A) en de split sample validatietest (B).De totale dataset wordt in subsets verdeelt, waarbij telkens een aantal subsets gebruikt worden voor training (blauwe blokken), en de resterende subsets voor validatie (bruine blokken). Doorgaans wordt kruisvalidatie gebruikt voor modelselectie (training van hyper-parameters) en de split-sample test voor modelevaluatie. Kortom, de blauwe blokken bij B zijn in fase A reeds gebruikt voor kruisvalidatie.

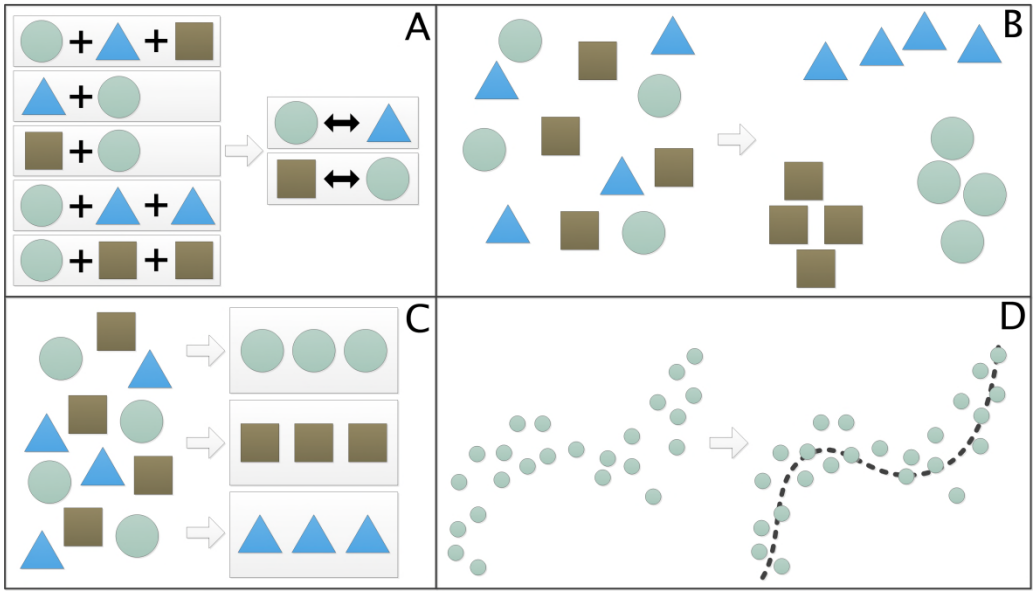
Voorbeeld - Een dataset bestaat uit 200 instanties. Men probeert met een selectie attributen van elke instantie een zekere doelvariabele te voorspellen. Daartoe splitst men de dataset in tweeën: 75% van de instanties wordt gebruikt voor modelselectie (kruisvalidatie) en de resterende 25% als testset voor modelbeoordeling. Het resultaat van de modelselectie is bemoedigend: het algoritme blijkt in staat om met zeer grote nauwkeurigheid de doelvariabele te voorspellen. Vervolgens neemt men de testset om te bepalen of het algoritme ook deze data goed weet te voorspellen. Dit blijkt tegen te vallen, op de resterende metingen faalt het algoritme in het geven van goede uitkomsten. Daardoor is het in de praktijk geen waardevol model: het is overgefit op de kruisvalidatie-set.

2.3.2 Datamining in statische datasets

Met datamining kan men meerdere doelen beogen. De mogelijke doelen worden mede bepaald door het type dataset (statisch of tijdreeks). Voor statische data worden in de literatuur doorgaans de volgende mogelijke doelstellingen onderscheiden (Bramer, 2007):

1. **Associatieregels vinden**
   Het afleiden van regels die het statistische verband tussen verschillende datapunten beschrijven (unsupervised).
2. **Clusteren**
   Het indelen van data in groepen, zonder vooraf klassen te definiëren. Hierbij is sprake van zogenaamd unsupervised machinaal leren.
3. **Classificeren**
   Een vorm van supervised machinaal leren, waarbij gegevens ingedeeld worden in vooraf gelabelde klassen.
4. **Voorspellen (regressie)**
   Het leggen van lineaire of niet-lineaire verbanden tussen variabelen. In het algemeen wordt onderscheid gemaakt tussen onafhankelijke (de predictors) en afhankelijke variabelen. Ook dit is een vorm van ‘supervised’ machinaal leren.

In Figuur 3 zijn de hierboven beschreven doelstellingen schematisch weergegeven.



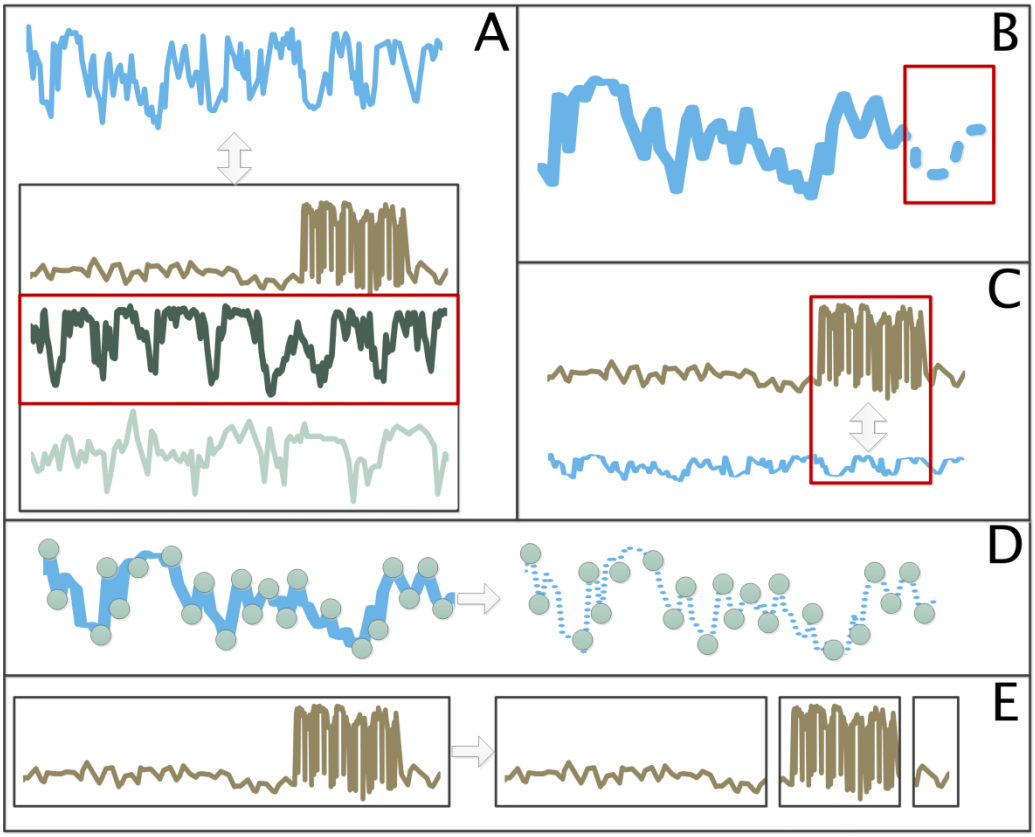
Figuur 3: Datamining-taken voor statische datasets: (a) associatieregels afleiden, (b) clusteren, (c) classificeren en (d) regressie. De symbolen representeren de verschillende instanties uit een dataset, waarbij de vorm van elk symbool de mate van 'gelijkheid' tussen deze instanties weergeeft.

2.3.3 Datamining in tijdreeksen

Analyse van tijdreeksen kan verschillende doelstellingen hebben (Esling & Agon, 2012; Ratanamahatana et al., 2010):

1. **Indexeren**
   Het doorzoeken van een database met tijdreeksen. Hierbij wordt een gegeven tijdreeks X als basis genomen, aan de hand waarvan gezocht wordt naar een andere tijdreeks Y, die het meest overeenkomt met de gegeven reeks.
2. **Voorspellen**
   Het voorspellen van een waarde op tijdstip n + 1 (of verder), gegeven een tijdreeks met n datapunten.
3. **Anomalieën detecteren**
   Het zoeken naar alle intervallen van een bepaalde tijdreeks X die als 'afwijkend' geclassificeerd kunnen worden op basis van een soortgelijke tijdreeks Y, waarin een andere variabele gemeten is voor dezelfde tijdstappen.
4. **Samenvatten**
   Het reduceren van het aantal datapunten van een tijdreeks, zonder daarbij de essentiële dynamiek verloren te laten gaan. Het doel is dus om de oorspronkelijke reeks zo goed mogelijk te benaderen, waarbij een zo hoog mogelijke reductie van datapunten nagestreefd wordt.
5. **Segmenteren**
   Het opdelen van een tijdreeks in een eindig aantal deelsegmenten, waarbij de begin- en eindpunten van elk segment bepaald worden door veranderingen in het regime (omslagpunten in de tijdreeks). Vanwege dit laatste aspect wordt segmenteren ook vaak 'change detection' genoemd.
6. **Clusteren**
   Het groeperen van individuele tijdreeksen in een database op basis van de mate waarin deze onderling overeenkomen.
7. **Classificeren**
   Het indelen van ongelabelde tijdreeksen in een beperkt aantal vooraf gedefinieerde klassen.

In Figuur 4 is zijn de verschillende doelstellingen (ook wel 'taken' genoemd) schematisch weergegeven.



Figuur 4: Datamining op tijdreeksen: (a) indexeren, (b) voorspellen, (c) anomalieën detecteren, (d) samenvatten, (e) segmenteren. De overige taken (clusteren en classificeren), zijn feitelijk niet anders dan bij statische datasets en zijn daarom niet in dit overzicht opgenomen (zie Figuur 3b en 3c).

2.3.4 Machine learning algoritmes in vogelvlucht

Centraal bij het doorzoeken van datasets staat de keuze voor een geschikt machine learning algoritme. Elk machine learning probleem bevat drie basiselementen:

- een representatiemodel: een model verkregen met een techniek die de data op bepaalde wijze representeert (dit kan een regressie, classificatie of clusteringsprobleem zijn);
- een evaluatiecriterium: een maat waarop de prestaties van het representatiemodel worden geëvalueerd;
- en een optimalisatie-algoritme: een algoritme dat de beste waarden onder het evaluatiecriterium zoekt.

Afhankelijk van de specifieke datamining-taak dient een selectie aan combinaties van mogelijke modellen, evaluatiematen en optimalisatietechnieken getest te worden. Een selectie aan veelgebruikte basiselementen is weergegeven in Tabel 1.

Tabel 1: Voorbeelden van veelgebruikte algoritme-elementen

Representatiemodel	**Evaluatiecriterium**	**Optimalisatie**
k-Nearest Neighbour (k-NN)	Correlatiecoefficient (R^2)	Greedy search
Support Vector Machines (SVM)	MSE (Mean Squared Error)	Beam search
Naive Bayes	Likelihood	Branch-and-bound
Logistische regressie	Information gain	Gradient descent
Neurale netwerken	Cost/utility	Conjugate gradient
Lineaire regressie	Margin	Quasi-Newton methoden
Beslisbomen		Linear programming
		Quadratic programming
		Evolutionary optimization

In de loop der jaren zijn vele representatiemodellen ontwikkeld. Deze zijn in te delen in een drietal typen: logische, geometrische en probabilistische algoritmes (Flach, 2012). In deze paragraaf wordt een korte omschrijving en opsomming gegeven van deze modellen. Een meer uitgebreide toelichting is opgenomen in Bijlage 1.

- *Logische algoritmes.* De zogenaamde 'logische algoritmes' construeren typisch modellen die intern georganiseerd zijn als netwerk- of boomstructuren. De data gaat aan de ene kant van het model deze structuur in en komt er via een stapsgewijze transformatie aan de andere kant weer uit. De meest bekende algoritmes en modellen van dit type zijn beslisbomen en Kunstmatige Neurale Netwerken (ANN).
- *Geometrische modellen* interpreteren data als punten in een meerdimensionale ruimte, waarbij het aantal attributen van de dataset tevens het aantal dimensies van deze ruimte is. In de meest eenvoudige vorm kan gedacht worden aan een tweedimensionale grafiek, waarin van elke instantie bijvoorbeeld het tijdstip uitgezet wordt tegen de waterstand. Door een geometrische benadering is het mogelijk om concepten als bijvoorbeeld 'afstand' en 'locatie' te gebruiken voor het classificeren van datapunten. Geometrische modellen hebben het voordeel dat de resultaten eenvoudig(er) te visualiseren zijn, onder de voorwaarde dat eventueel gevonden patronen in twee- of driedimensionale projecties worden gereduceerd of bepaald. Enkele veelgebruikte algoritmes en modellen zijn lineaire regressie, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), K-means clustering en eigenwaardendecompositie of Principal Component Analysis (PCA).
- *Probabilistische modellen* gaan uit van een bepaalde waarschijnlijkheid dat attributen in data aan elkaar zijn gerelateerd. Bijvoorbeeld: een voetbal wordt gekenmerkt door een ronde vorm, een zekere diameter en een bepaald materiaal waarvan het gemaakt is. Een Naive Bayes classifier beschouwt elk van deze eigenschappen als onafhankelijk van elkaar en berekent de kans dat iets een voetbal is, gegeven de waarschijnlijkheid dat het object deze eigenschappen bezit.

Tabel 2: Overzicht van veelgebruikte machine learning representatiemodellen en hun voornaamste kenmerken.

Respresentatiemodel	Klasse	Doel	Inzichtelijkheid	Type model
Beslisbomen	Supervised	Classificatie	Goed	Logisch
Regressiebomen	Supervised	Regressie	Goed	Logisch
Feedforward KNN	Supervised	Regressie/classificatie	Slecht	Logisch/geometrisch
Lineaire regressie	Supervised	Regressie	Goed	Geometrisch
k-NN	Supervised	Classificatie	Slecht	Geometrisch
k-means	Unsupervised	Clustering	Matig	Geometrisch
SVM	Supervised	Classificatie	Matig	Geometrisch
SVR	Supervised	Regressie	Matig	Geometrisch
Random forests	Supervised	Regressie/classificatie	Redelijk	Ensemble
Gradient Boosting Regression	Supervised	Regressie	Redelijk	Ensemble
Naive Bayes	Supervised	Classificatie	Slecht	Probabilistisch

2.3.5 Selectie van een geschikt algoritme

Het is vrijwel altijd sub-optimaal om vooraf al een keuze te maken ten aanzien van de te gebruiken machine learning algoritmes voor een analyse. De gebruikelijke aanpak bij data-analyses is om een (groot) aantal algoritmes te testen en te onderzoeken welke de beste resultaten geeft. Doorgaans wordt bij machine learning gestart met eenvoudige modellen, voordat men overgaat tot de geavanceerde methodieken (bijvoorbeeld: begin met Naive Bayes, daarna logistische regressie, vervolgens k-Nearest Neighbor en tot slot Support Vector Machines). Domingos (2012) constateert dat de beste resultaten op dit moment doorgaans behaald worden met model ensembles.

2.4 De valkuilen van datamining

Hoewel datamining een methode voor ‘knowledge discovery’ kan zijn, is er bij onjuist gebruik het risico van een foutieve interpretatie. In deze paragraaf worden de meest voorkomende valkuilen uitgelicht.

2.4.1 Gebruik van niet-representatieve datasets (sampling bias)

Datasets worden geacht een representatief beeld te geven van een situatie, door onafhankelijke observaties en consistentie in monsternames. Door diverse omstandigheden kan de data niet meer representatief zijn:

- bij *datamanipulatie* zijn (bewust) datapunten uit een dataset verwijderd, is informatie verkregen door interviews met suggestieve vragen, of is data gemanipuleerd door het (bewust of onbewust) selectief kiezen van steekproeven uit een grote populatie;
- een proefopzet die onvoldoende is: bijvoorbeeld door fouten in de meetopstelling, of een meetperiode die niet lang genoeg is om de relevante dynamiek van fysische processen te vangen, of door verschillende procedures van bemonsteringen waarvoor niet gecorrigeerd is.

2.4.2 Overfitting

Een grote valkuil bij datamining is overfitting. De oorzaak van overfitting is dat machine learning algoritmes in staat zijn om een eigen modelstructuur te creëren die tijdens het

trainen zo goed mogelijk wordt afgestemd op de trainingsset. Als er veel vrijheidsgraden toegekend worden aan het algoritme kan doorgaans de trainingsdata zeer goed voorspeld worden, zonder dat er rekening wordt gehouden met ruis of irrelevante gegevens in de dataset. Het gevolg is dat er met andere (nieuwe) datasets inaccurate voorspellingen worden gegenereerd. Kortom: het algoritme is weinig betrouwbaar en kan niet meer goed omgaan met nieuwe data, omdat de structuur en modelparameters volledig geoptimaliseerd zijn om de trainingsdata (inclusief de daarin aanwezig ruis en irrelevante gegevens) te reproduceren.

Om overfitting te voorkomen dient een data-analist (a) beperkingen op te leggen aan het maximaal aantal vrijheidsgraden van de modelstructuur en (b) gebruik te maken van een dataset met daarin genoeg variatie in gegevens ('informatierijkdom') om een representatief model te trainen. Uiteindelijk dient gezocht te worden naar een zo eenvoudig mogelijke modelstructuur, waarmee nog steeds met acceptabele nauwkeurigheid voorspellingen gedaan kunnen worden. Albert Einstein heeft dit principe ooit mooi verwoord met de stelling 'everything should be made as simple as possible, but not simpler'. In de praktijk zijn er een aantal methoden en vuistregels waarmee een data-analist de kans op overfitting kan beperken. Voorbeelden hiervan zijn het automatisch afbreken van de modeltraining zodra de betrouwbaarheid van de validatieset structureel afneemt, het uitvoeren van kruisvalidatie op de dataset of het opschonen van de getrainde modelstructuur middels het verwijderen van elementen die weinig toevoegen aan de voorspelling (ook wel 'pruning' genoemd).

2.4.3 Data dredging

Data dredging is het doorzoeken van grote datasets om daarin een correlatie tussen variabelen aan te tonen, zonder deze correlatie achteraf te valideren op een tweede dataset. Belangrijk is dat het zoeken naar correlaties tussen variabelen in een dataset op zichzelf een valide methode is, maar dat er nooit een hypothese zowel afgeleid als getoetst mag worden op dezelfde dataset. Daarom dient een onderzoek altijd een dataset voor training en validatie (kalibratie) te hebben en daarnaast een tweede dataset voor het testen.

2.4.4 Onjuist veronderstelde oorzakelijkheid (niet-causaliteit)

Een tweede aandachtspunt bij datamining is het probleem van onjuist veronderstelde oorzakelijkheid (niet-causaliteit): een sterke correlatie tussen twee variabelen betekent niet automatisch een causaal verband; de variabelen kunnen ook toevalligerwijs ten opzichte van elkaar variëren. Met name bij relatief kleine datasets kunnen snel correlaties gevonden worden zonder oorzakelijkheid. Daarnaast is het ook mogelijk dat twee variabelen met elkaar correleren doordat ze beide causaal afhankelijk zijn van een derde factor die niet gemeten is (indirecte oorzakelijkheid). Het is daarom altijd van belang om tijdens het dataminen te controleren of een gevonden verband ook een fysische verklaring heeft en zo de betekenisvolle en betekenisloze verbanden van elkaar te onderscheiden.

Voorbeeld - In een dataset wordt tijdens een datamining-analyse een sterke correlatie gevonden tussen het aantal ooievaars in een gemeente over een zekere periode en de hoeveelheid kindergeboortes in dezelfde gemeente in dezelfde periode. Ondanks de correlatie hebben beide zaken geen causaal verband met elkaar. Er mag dus niet gesteld worden dat het aantal kinderen een gevolg is van een toename van het aantal ooievaars. Kortom: een correlatie is pas betekenisvol indien er ook een fysische verklaring (oorzakelijk verband) voor is aan te wijzen.

3 Hoe kan datamining assetmanagement ondersteunen?

In dit hoofdstuk staat de onderzoeksvraag 'hoe kan datamining het assetmanagement van drinkwaterbedrijven ondersteunen?' centraal. Eerst wordt het bredere kader rondom deze vraag geschetst (paragraaf 3.1), waarna vervolgens wordt ingegaan op de procesmatige en organisatorische integratie van datamining, KDD en expertsystemen binnen drinkwaterbedrijven (paragraaf 3.2). Binnen het algehele kader van informatievoorziening ten behoeve van assetmanagement is door KWR een actuele inventarisatie gemaakt van de voornaamste vragen die momenteel spelen bij drinkwaterbedrijven, die in paragraaf 3.3 zijn toegelicht.

3.1 Kennisbehoefte bij assetmanagement

Assetmanagement omvat een integrale werkwijze bij drinkwaterbedrijven, waarbij de gehele keten aan infrastructuur (van bron tot tap) onderwerp van beschouwing is. Cruciaal voor een effectief beheer is daarbij de beschikbare informatie over de fysieke assets. De drinkwatersector kenmerkt zich door een grote mate van sociale verantwoordelijkheid (leveringsgarantie), kapitaalintensieve assets en een typisch lange levensduur van assets. Specifiek aandachtpunt is daarbij het leidingnet, waarvoor het kostbaar is om de conditie te bepalen als het eenmaal is aangelegd. Voor het nemen van degelijke investeringsbeslissingen en het inplannen van onderhoud is echter betrouwbare informatie wenselijk.

Recentelijk zijn ISO normen ontwikkeld om assetmanagement te structureren. In deze ISO 55000:2014 - 55002:2014 normen wordt een opsomming gegeven van de voornaamste aspecten waarover informatie nodig is om een organisatie effectief te ondersteunen bij het nemen van investeringsbeslissingen, het inplannen van onderhoud en het maximaliseren van de operationele prestaties van assets. Specifiek voor drinkwater distributie-infrastructuur zijn deze normen vertaald naar de ISO/DIS 24516-1 (Guidelines for Management of Assets of water supply and wastewater systems - Part 1: Drinking water distribution networks). Een belangrijke vereiste vanuit deze normen is dat infrastructuur wordt gemonitord op een aantal cruciale aspecten aan de hand van vooraf gedefinieerde Prestatie-Indicatoren (PI's). Doorgaans worden bedrijfsmatige doelstellingen periodiek geëvalueerd aan de hand van deze PI's.

Gerelateerd aan de data die vrijkomt uit continue monitoring en die nodig is voor kwantitatieve beoordeling van assets vallen er uit de ISO normen een aantal aspecten te destilleren die in de context van dit onderzoek van belang zijn. De drie samenhangende aspecten die een volledig beeld geven van de kwaliteit van een asset zijn conditie, operationele prestatie en storingsrisico. Voor elk van deze aspecten worden doorgaans PI's gespecificeerd, waarvan enkele voorbeelden zijn weergegeven in Tabel 3.

Tabel 3: De drie kernaspecten van infrastructureel assetmanagement. Voor een volledig overzicht van prestatie-indicatoren, zie Beuken (2015).

Aspect	**Voorbeeldvragen**	**Voorbeeld PI's**
Conditie	• Hoe staat het met de kwaliteit van de asset op dit moment? • Hoeveel jaar gaat deze asset naar verwachting nog mee? • Wat is de actuele waarde?	• Restlevensduur. • Conditiescore. • Mate van slijtage in relatie tot de norm.
Operationele prestatie	• Hoe goed vervult de asset zijn taak?	• Energieprestatie. • Geleverde waterkwaliteit. •
Storingsrisico	• Wat is de relatie tussen de asset-conditie en de storingsfrequentie? • Wat is de actuele storingskans van de asset? • Wat is de impact van een storing op het gehele systeem? • Onder welke omstandigheden is er een verhoogde of acute kans op een storing?	• Betrouwbaarheid (gemiddelde tijdsduur tussen faalgebeurtenissen). • Beschikbaarheid (gemiddelde uptime). • Ondermaatse leveringsminuten.

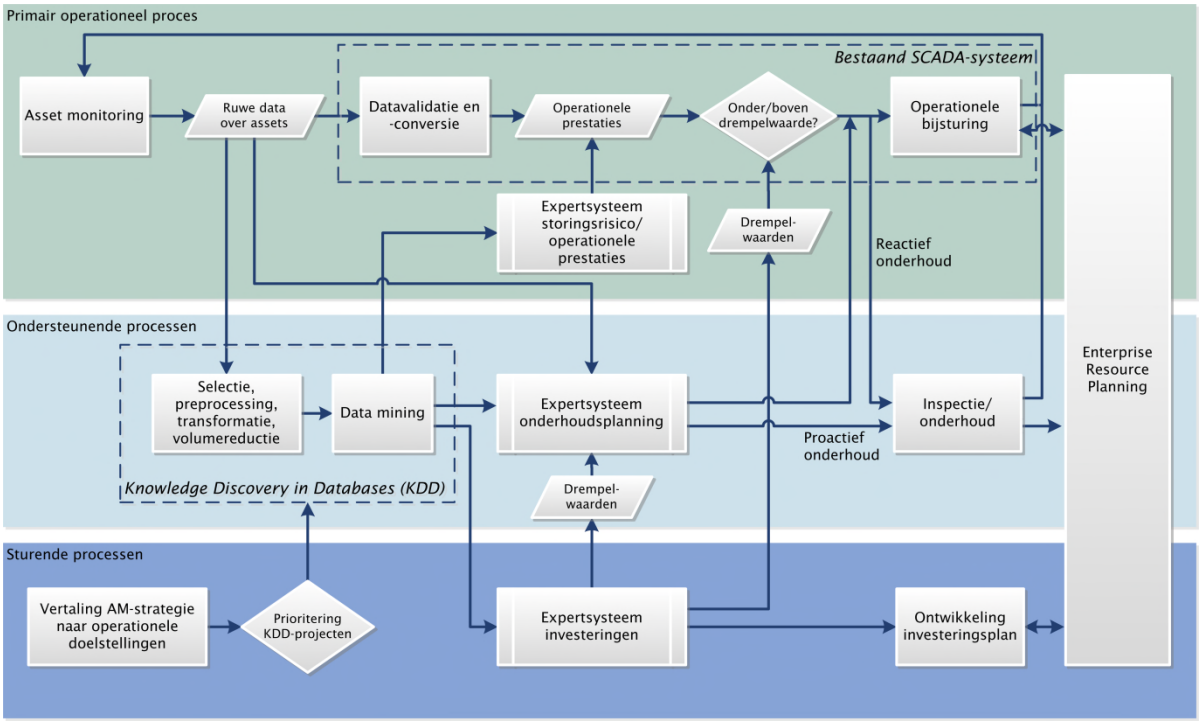
3.2 De rol van datamining in de bedrijfsvoering

Assetmanagement kan gezien worden als een integraal onderdeel van alle bedrijfsprocessen bij drinkwaterbedrijven. Het KDD-proces is binnen die context een specialistisch ondersteunend proces dat projectmatig georganiseerd is en ten doel heeft een specifiek bedrijfsproces te verbeteren (Figuur 5). Datamining resulteert in concrete producten, namelijk getrainde en gevalideerde machine learning algoritmes, die operationeel ingezet kunnen worden (dit is in feite de laatste stap in het CRISP-DM schema in paragraaf 2.2). Als een machine learning algoritme operationeel wordt ingezet is dat doorgaans in de vorm van een expertsysteem. Dat is een computersysteem dat op basis van ingevoerde kennisregels binnen een bepaald gebied oplossingen biedt. Het systeem functioneert daardoor als een digitaal equivalent van een menselijke 'expert' binnen een organisatie.

De manier waarop expertsystemen operationeel gebruikt worden is sterk afhankelijk van de typische omloopsnelheid waarmee informatie binnen de drie typen bedrijfsprocessen (operationeel, ondersteunend en sturend) stroomt en de snelheid waarmee nieuwe data gegenereerd wordt binnen deze bedrijfsprocessen. Indien er data met hoge meetfrequentie beschikbaar is en er tevens behoefte is aan real-time inzicht in processen, dan is het mogelijk om zogenaamde real-time data-analyse uit te voeren. Dit biedt drie soorten mogelijkheden (Bifet, 2013):

1. **Early warning**
   Snel kunnen reageren zodra er afwijkingen aan het system worden gedetecteerd.
2. **Real-time inzicht**
   Real-time inzicht in cruciale bedrijfsprocessen op detailniveau, waardoor ook het strategisch management op de hoogte is van actuele ontwikkelingen op het operationele procesniveau.
3. **Real-time feedback**
   Gelijk na bijsturing of implementatie van een nieuwe strategie zien of de beoogde effecten ook daadwerkelijk optreden.

In het geval van (real-time) besturingen, is het primaire operationele proces gebaat bij real-time data-analyse. Managementprocessen kennen veelal een tragere informatiestroom, met een meer periodieke raadpleging van informatiebronnen. Datamining-projecten dienen op deze verschillen toegespitst te zijn en moeten bovendien aansluiten op de specifieke informatiebehoefte binnen elk bedrijfsproces. In een assetmanagement-context, specifiek voor drinkwaterbedrijven, laten typische kennisvragen zich daarom op elk bedrijfsniveau doorvertalen naar een ander type expertsysteem. De systemen worden gevoed door datamining projecten en worden gebruikt om tot meer inzicht te komen en betere beslissingen te nemen (Figuur 5).



Figuur 5: De positie van KDD, datamining en expertsystemen in een assetmanagement context. Data heeft binnen de diverse bedrijfsprocessen van een drinkwaterbedrijf typisch verschillende omloopsnelheden. Zo worden volumestroom en druk (belangrijke stuurparameters in het operationele proces) continu gemonitord, terwijl in de ondersteunende en sturende bedrijfsprocessen typisch meer 'trage' statische data wordt gebruikt voor analyses (o.a. conditiemetingen). Daarnaast is de informatiebehoefte in elk bedrijfsproces uniek. Dit vertaalt zich naar drie verschillende expertsystemen, die aansluiten bij de behoeften van de verschillende bedrijfsprocessen: een real-time systeem voor het voorspellen van storingen en foutdiagnose, een systeem voor conditiebepaling waarmee het proactief onderhoud gestuurd kan worden en een beslissingsondersteunend systeem voor investeringsplanningen. KDD is een proces dat typisch projectmatig uitgevoerd wordt en waarvan de uitkomsten gebruikt kunnen worden voor het ontwikkelen van de diverse expertsystemen.

3.3 Actuele vragen asset managers

Binnen de context zoals geschetst in paragraaf 3.1 en 3.2 is door KWR een inventarisatie gemaakt van de voornaamste assetmanagement-vraagstukken bij de Nederlandse drinkwaterbedrijven. Via een online enquête en workshop is aan assetmanagers en data-analisten van de drinkwaterbedrijven gevraagd welke vragen rondom informatievoorziening ten behoeve van assetmanagement momenteel als meest urgent en belangrijk ervaren worden. Uit deze inventarisatie is, na groepering van overlappende vragen en herschikking in categorieën, een lijst met 12 vragen naar voren gekomen. De gecursiveerde kennisvragen zijn aan bod gekomen in de TKI-projecten DiAMANT-Water (D) en BWD2SWG (B).

Er is onderscheid gemaakt tussen de volgende categorieën, waarbij datamining een rol in het vergroten van inzicht in de gerelateerde kwesties kan spelen:

- **Afschatten van storingsrisico's**
  Hoe kunnen we:

  1. lekkages in het leidingnet real-time voorspellen?
  2. storingen aan regel- en keerkleppen real-time voorspellen?
  3. storingen aan pompen real-time voorspellen?
  4. *relaties leggen tussen de conditie van een asset en de storingsfrequentie? (D)*
  5. de invloed van bedrijfsvoering op de gemiddelde storingsfrequentie van een winning bepalen?
  6. *de invloed van bedrijfsvoering op de gemiddelde storingsfrequentie van een leiding bepalen? (D)*

- **Conditiebepaling**
  Hoe kunnen we:

  7. de huidige conditie van een leiding of leidingcohort bepalen uit ruwe data van inwendige of uitwendige inspectietools?
  8. mogelijke relaties vinden tussen de waterkwaliteit en actuele leidingconditie?
  9. een inschatting maken van de conditie-afname van winningen (bijvoorbeeld door putverstopping, ouderdom of ijzerversmering onderwaterpomp)?

- **Inzicht in operationele prestaties**
  Hoe kunnen we:

  10. *een verklaring vinden voor pieken in energie- of waterverbruik? (B)*
  11. een verklaring vinden voor wisselingen in de drinkwatersamenstelling?
  12. *de oorzaken van ongewenste veranderingen in de bedrijfsvoering (zoals 'hysterie' in pompaansturing[1]), die tot een incident kunnen leiden, achterhalen? (B)*

[1] Verschijnsel waarbij pompen zeer frequent automatisch aan- en uitgeschakeld worden.

4 Beschikbaarheid en kwaliteit van databronnen

Datamining is een krachtige methodiek om verbanden te ontdekken waarmee kennis verder ontwikkeld of getoetst kan worden. Het succes van datamining projecten staat of valt echter met de kwaliteit én beschikbaarheid van data. In dit hoofdstuk wordt, conform de derde onderzoeksvraag van dit rapport, verder ingegaan op die beschikbaarheidsaspecten (paragraaf 4.1) en kwaliteitsaspecten (paragraaf 4.2).

4.1 Beschikbaarheid databronnen

Voor het beantwoorden van de kennisvragen rondom assetmanagement van waterinfrastructuur zijn een aantal databronnen beschikbaar bij drinkwaterbedrijven. Hierbij kan gedacht worden aan centrale (zoals USTORE) en decentrale storingsregistratie, leidingnetmodellen, stuursignalen uit besturingssystemen als SCADA (Supervisory Control And Data Aquisition) en plant historians (bijvoorbeeld PI), energieverbruik-registratie, waterkwaliteitsbemonstering en continue waterkwaliteitsmonitoring rondom winningen, in zuiveringen en in het leidingnet.

Naast de data die beschikbaar is binnen drinkwaterbedrijven zijn er ook externe databronnen die geraadpleegd kunnen worden om analyses te complementeren of versterken. Gedacht kan worden aan bijvoorbeeld meteorologische data (zoals neerslag- en verdampingsreeksen) van het KNMI, ondergrondgegevens (waaronder de REGIS-II en GeoTOP modellen van TNO), gegevens geëxtraheerd uit social media en storingsregistraties van soortgelijke assets als bij drinkwaterbedrijven in andere sectoren (bijvoorbeeld RIONED voor rioleringssystemen), of andere landen.

Databeschikbaarheid is echter vaak een probleem vanwege een of meerdere van onderstaande aspecten:

- **Ontbrekende labels**
  Meetnetten en databases ontsluiten vaak een grote hoeveelheid gegevens. Regressie- en classificatie is echter het meest gebaat bij labeling van de afhankelijke variabelen, respectievelijk te onderscheiden klassen. Met andere woorden, de te voorspellen variabele moet bekend en gelabeld zijn. In de praktijk komt het erop neer dat bijvoorbeeld voor het voorspellen van storingen aan infrastructuur niet alleen reeksen met temperatuur, druk, volumestroom en stroomverbruik bekend moeten zijn, maar dat ook een registratie van storingen of van afwijkingen in de bedrijfsvoering (van sensoren) nodig is.
- **Veel data, maar weinig 'events'**
  Naast het volledig ontbreken van labeling of registratie van 'events' (zie hierboven) kan het ook voorkomen dat getracht wordt om voorspellingen te doen over een verschijnsel dat relatief weinig voorkomt en daardoor in de praktijk ook nog maar weinig is gemeten. Als een bepaald verschijnsel in het leidingnet voorspeld moet worden, dan dienen in de aangeleverde datasets honderden gevallen aanwezig te zijn waarbij dit verschijnsel in het verleden is opgetreden. Alleen op die manier kan een algoritme verbanden leren tussen optreden van het verschijnsel en de omstandigheden waaronder dit optrad.

- **Te korte meetperiode**
  Om ook op lange termijn betrouwbare voorspellingen te kunnen doen is het belangrijk om de beschikking te hebben over langetermijnreeksen met meetgegevens om langetermijndynamiek van processen te kunnen vangen. Hierdoor kan beter inzicht verkregen worden in de natuurlijke dynamiek van het systeem en kunnen trends uit de reeksen 'gefilterd' worden. Indien er alleen data beschikbaar is over een korte tijdspanne, dan bestaat het risico dat een getraind machine learning algoritme na verloop van tijd uit de pas gaat lopen met de werkelijkheid. Het gevolg is dat het aantal foutieve voorspellingen door de tijd hierdoor langzaam toeneemt (een verschijnsel dat 'drift' genoemd wordt). Dit komt omdat het algoritme niet getraind is op het herkennen van langjarige patronen, simpelweg omdat de meetreeks hier te kort voor is.
- **Te lage meetfrequentie**
  Niet alleen de meetperiode, maar ook de frequentie waarmee bemonsterd is of waarmee de data is weggeschreven moet in verhouding staan tot de dynamiek van het systeem waaraan gemeten wordt en de tijdsschaal waarop relevante verschijnselen in het systeem optreden. Los daarvan zal de meetfrequentie moeten aansluiten bij de vorm van kennisbehoefte. Indien men real-time inzicht wil hebben in een systeem, dan dient er ook real-time data beschikbaar te zijn om de gewenste (temporele) resolutie te kunnen waarborgen.

4.2 Datakwaliteit

Naast databeschikbaarheid dient ook de kwaliteit van de data aan minimale eisen te voldoen. Een model, verkregen uit datamining, is hooguit zo goed als de data waaruit het geëxtraheerd is. Hoewel de benodigde kwaliteit afhankelijk is van de specifieke vraag die ermee beantwoord moet worden, zijn er een aantal veel voorkomende fouten en afwijkingen te benoemen in datasets. Fouten variëren van typfouten tot het eenvoudigweg ontbreken van meetwaarden. Daarnaast kunnen meetinstrumenten onjuist gekalibreerd, onjuist geplaatst of onjuist uitgelezen zijn. Tot slot kunnen er verkeerde eenheden aan meetreeksen gekoppeld zijn. Naast deze door mensen geïntroduceerde fouten worden er ook diverse andere onnauwkeurigheden in databronnen geïntroduceerd door databewerkingen in data-opslagsystemen en fouten veroorzaakt door sensoren of meetinstrumenten. Sensorfouten kunnen bijvoorbeeld ontstaan door vervuiling van het meetinstrument, waardoor deze na verloop van tijd uit de pas lopen. Ook kunnen bij langjarige meetreeksen na verloop van tijd wijzigingen worden doorgevoerd in de meetopstelling, worden sensoren af en toe vervangen of kunnen er aanpassingen zijn gedaan in de meetfrequentie.

Fouten en onnauwkeurigheden in het meten en in het samenstellen van databases zorgen ervoor dat de data die uiteindelijk ter beschikking staat aan een data-analist kan afwijken van de werkelijkheid. Tot op zekere hoogte zijn fouten op te sporen door een geautomatiseerde toetsing uit te voeren voorafgaand aan verdere data-analyse (Von Asmuth & Van Geer, 2013). In deze studie worden twee typen toetsing aangeraden voor het opsporen van fouten: consistentie- en plausibiliteitstoetsen. Een meetwaarde is inconsistent indien deze fysisch gezien onmogelijk is (een stijghoogte in een peilbuis kan bijvoorbeeld niet worden gemeten als deze lager is dan de onderkant van het filter en een stofconcentratie in een waterkwaliteitsmeting kan niet negatief zijn). Toetsing op plausibiliteit geeft aan of een meetwaarde waarschijnlijk is, gegeven de statistische eigenschappen van alle voorgaande meetwaarden. Het gaat hier typisch om het detecteren van uitbijters. Hoewel het onderzoek specifiek op grondwaterreeksen toegespitst is, zijn de achterliggende principes breed toepasbaar.

5 Datamining in de praktijk

Diverse drinkwaterbedrijven hebben het initiatief genomen om de mogelijkheden van datamining en KDD in de praktijk uit te testen. KWR is bij twee TKI[2] -projecten rondom datamining betrokken geweest. Deze projecten hebben geleid tot een drietal casussen, die datamining voor assetmanagement gerelateerde vraagstukken illustreren. Bij casus 1 en 2 heeft Brabant Water als eindgebruiker en dataleverancier een rol gespeeld (paragraaf 5.1 en 5.2). Vitens is bronhouder van de data uit de derde casus (paragraaf 5.3). De casussen zijn uitgebreid beschreven in de TKI-rapporten KWR 2015.108 (Vries, Vonk, Jong, Van Duist, & Marel, 2015) en KWR 2016.006 (Vonk et al., 2015).

5.1 Casus 1 – Het verklaren van bruinwaterklachten

Het optreden van bruin water bij gedistribueerd drinkwater is een wijd verspreid fenomeen dat wereldwijd optreedt in veel drinkwaterdistributiesystemen. Hoewel bruin water voornamelijk een esthetisch probleem is en ongevaarlijk voor de gezondheid van consumenten, is het een niet ongewone reden voor klanten om een melding te maken bij drinkwaterbedrijven. Hierom willen bedrijven de relevante processen begrijpen en de bruinwaterincidenten en –meldingen reduceren.

In deze casus gebruiken we een uitgebreide data-analyse om het voorkomen van bruinwatermeldingen nader te onderzoeken. Hiervoor combineren we een grote verscheidenheid aan data, waaronder 5 jaar aan klantcontactgegevens, demografische gegevens en temperatuurgegevens. De analyse is gericht op een beter begrip van mogelijke invloeden van leidingnetwerkkarakteristieken, temperatuur en demografische factoren, zonder sterke aannames te maken over de processen of afhankelijkheden.

Voor dit onderzoek waren diverse databronnen bij Brabant Water beschikbaar, welke gecombineerd zijn met externe datasets van het CBS en KNMI. Een overzicht met data-karakteristieken is gegeven in Tabel 4. Eerste verkenningen van de data laten zien dat het gaat om relatief kleine datasets; er is geen sprake van zogenaamde big data.

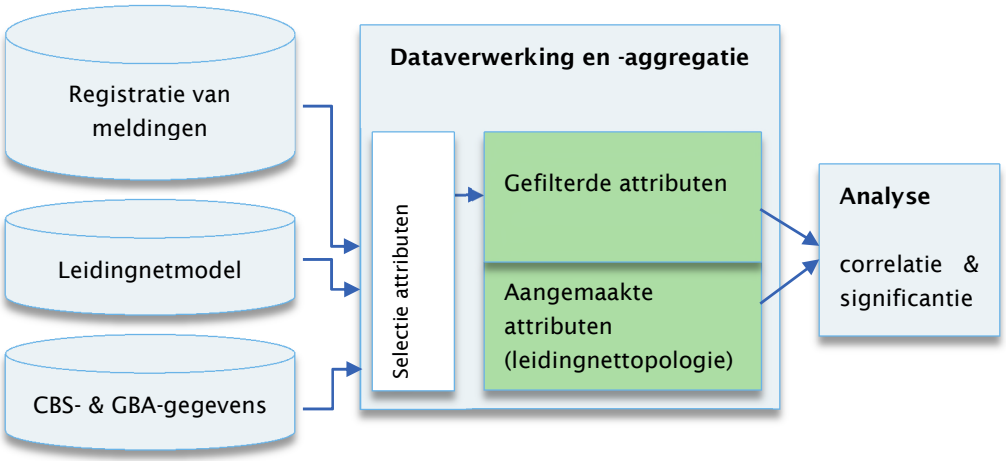
Tabel 4: Overzicht databronnen beschikbaar voor casus 1

Data	Bronhouder	Meetfrequentie	Databeheer/ software
Leidingnet	Brabant Water	eenmalig	InfoWorks
Klantmeldingen	Brabant Water	onregelmatig	Software klantcontacten
Buurtstatistieken	CBS	jaarlijks	GBA
Meetreeksen buitenluchttemperatuur	KNMI	dagelijks	KNMI

[2] TKI staat voor Topconsortia voor Kennis en Innovatie.

De databronnen zijn gekoppeld, gecombineerd en geanalyseerd (de 'data voorbewerking' stap van het CRISP-DM model uit paragraaf 2.2). De hoofdstappen in deze methodiek zijn de voorbewerking en selectie van data, het kwantificeren van de leidingnettopologie en de uitvoering van statistische analyses. Dit is schematisch weergegeven in Figuur 6. Daarbij zijn twee analysemethodieken toegepast:

1. Het uitvoeren van een correlatieanalyse van demografische gegevens, leidingnetgegevens en klantmeldingen met betrekking tot bruinwaterincidenten en algemene klantmeldingen;
2. Uitvoering van significantietesten tussen (enkelvoudige) relaties van temperatuur, leidingdiameter en leidingmateriaal met meldingen over vuilwaterlast.



Figuur 6: Methodiek op hoofdlijnen

Om het leidingnetwerkontwerp te karakteriseren zijn twee maten berekend op basis van het Infoworks-netwerkmodel (het zogenaamde 'feature engineering', zoals omschreven in paragraaf 2.2):

- De vermazingsgraad (G_{maas}): Een indicator voor de vertakte of vermaasde netwerken, berekend als de verhouding ban het aantal mazen (*M*) ten opzichte van het aantal aansluitingen (*A*).
- De netwerkdimensioneringsgraad (G_{dim}): Een indicator voor de totale verbruikslast op het distributienet, gedefinieerd als de verhouding van het totale volume van distributieleidingen (*V*) tot het gemiddelde uurlijks klantverbruik (*Q*) in een gebied. Een lage waarde van G_{dim} is indicatief voor een relatief klein distributievolume voor een gegeven watervraag, wat relatief hoge gemiddelde stroomsnelheden bevordert.

De resulterende statistieken zijn gecombineerd met alle overige datasets en gebruikt als input voor de statistische analyses. Alle data is gevisualiseerd middels een zogenaamde correlatiematrix, hetgeen een krachtige techniek blijkt om snel een dataset inzichtelijk te maken. Daarvoor is gebruik gemaakt van de Spearman's rank correlatie ρ, die afhankelijkheden beschrijft met een monotoon stijgende of dalende functie. Tot slot hebben we voor gevonden relaties bepaald of verschillen in de distributies statistisch significant zijn met een t-toets voor datasets van ongelijke monstergrootte en ongelijke variantie (de zogenaamde Welch's t-toets voor twee onafhankelijke steekproeven). Gezien de specifieke vraagstelling van deze casus zijn geen machine learning algoritmes gebruikt.

Het onderzoek heeft een aantal zaken aangetoond:

- klantmeldingen waren niet significant gecorreleerd aan demografische gegevens of afgeleide parameters van het leidingnet (G_{dim} en G_{maas});
- een statistisch significante, positieve correlatie tussen bruinwatermeldingen en de buitentemperatuur is aangetoond. Dit resultaat ondersteunt eerder onderzoek dat suggereert dat hoge temperaturen het optreden van bruinwater bevorderen.
- een significante negatieve correlatie tussen bruinwatermeldingen en leidingdiameter is gevonden.

5.2 Casus 2 – Welke factoren spelen een rol bij regionale verschillen in storingsfrequentie?

Brabant Water wil graag weten welke factoren invloed hebben op de conditiedegradatie en storingsfrequentie van leidingen in het distributienet. Daarbij is de verwachting dat dergelijke factoren kunnen dienen als voorspellers voor storingen. Met betrouwbare voorspellers voor de integriteit van het distributienet kan een effectiever saneringsbeleid worden bewerkstelligd, zodat prioriteit wordt gegeven aan leidingen die de grootste kans lopen om te storen.

Uit een recente, parallel uitgevoerde KWR-studie blijkt dat storingsfrequentie van leidingen verband houdt met het drukregime van pompstations (Wols, 2015). Bij deze casus is de relatie tussen data uit het procesinformatiesysteem (druk- en volumestroomgegevens) met gegevens van het leidingnet (storingen en leidingnetkarakteristieken) onderzocht. Specifiek is de volgende vraagstelling geformuleerd: is er een relatie tussen het aantal storingen per eenheid leidinglengte en het gemeten dynamische drukregime van het toeleverende pompstation?

Voor deze casus is gebruikt gemaakt van diverse databronnen. Een overzicht hiervan is gegeven in Tabel 5. Hoewel bepaalde datasets relatief hoogfrequente metingen bevatten, is er geen sprake van zogenaamde big data.

Tabel 5: Overzicht databronnen beschikbaar voor casus 2

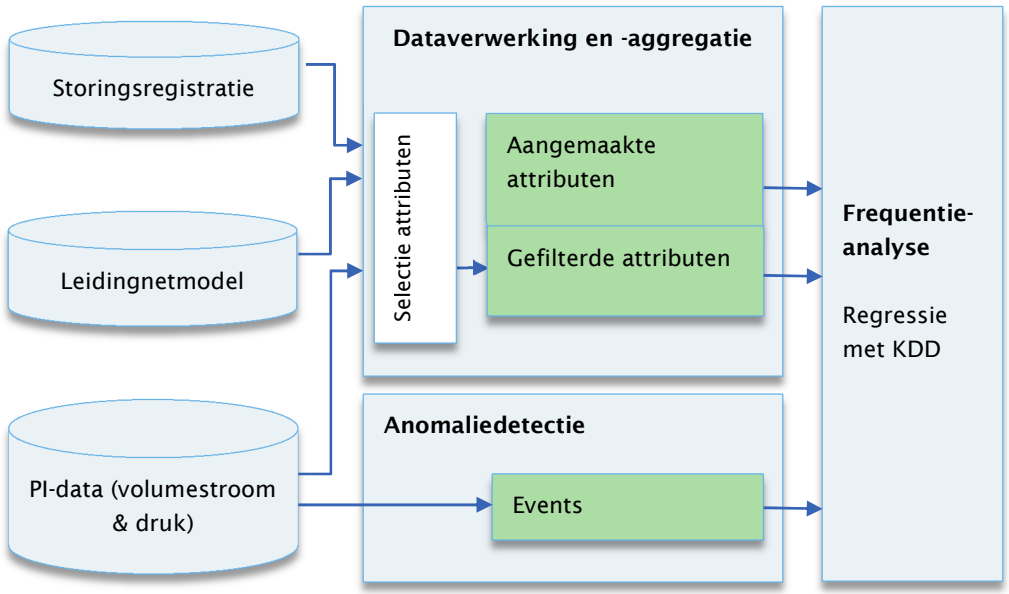
Data	Bronhouder	Meetfrequentie	Databeheer/ software
Leidingnet	Brabant Water	eenmalig	InfoWorks
Meetreeksen druk	Brabant Water	1/minuut	OSISoft PI
Meetreeksen volumestroom	Brabant Water	2/minuut	OSISoft PI
Storingen	Brabant Water	onregelmatig	USTORE
Begrenzing voorzieningsgebieden	Brabant Water	eenmalig	ArcGIS
Locaties WPB's, aanjagers en opjagers	Brabant Water	eenmalig	ArcGIS
Vertaaltabel sensorcoderingen-leidingtakken	Brabant Water	eenmalig	Excel

De databronnen in Tabel 5 zijn gekoppeld, gecombineerd en geanalyseerd. Om een relatie te leggen tussen drukschommelingen zoals gemeten op dit beperkte aantal puntlocaties en het optreden van storingen in een geografisch groot gebied is het noodzakelijk om zogenaamde

'drukzones' te definiëren: gebieden waarvoor aangenomen kan worden dat deze min of meer in gelijke mate de druk ondervinden van het bijbehorende pompstation. Middels deze drukzones kunnen dan de drukkarakteristieken zoals gemeten op een zekere meetlocatie gekoppeld worden aan storingen in het omringende gebied. Deze bewerking is in feite de data-voorbewerkingsstap uit de CRISP-DM methodiek (paragraaf 2.2).

Per drukzone zijn (a) meetreeksen voor volumestroom en druk opgevraagd uit de PI-database, (b) storingen opgehaald uit USTORE en (c) relevante leidingnetkarakteristieken opgehaald uit Infoworks exports. Drukreeksen zijn daarna geaggregeerd tot de juiste tijdschaal, om vervolgens direct gebruikt te worden als invoer voor de uiteindelijke analyse naar relaties tussen drukregime en storingen. In een parallel spoor zijn de drukreeksen en flowreeksen in hoge resolutie (5-minuten interval) gebruikt om middels een anomaliedetectiealgoritme geautomatiseerd zogenaamde 'events' te genereren. Een event is daarbij gedefinieerd als een moment waarop in de meetreeksen een afwijkend patroon te zien is. Een dergelijke afwijking kan veroorzaakt worden door bijvoorbeeld een leidingbreuk, verandering in de pompaansturing, openstaande brandkraan, bijzondere (feest-)dag, etcetera. De zogenaamde 'events' per drukzone zijn, samen met de geaggregeerde drukreeksen, gebruikt als invoer voor de analyse naar correlaties tussen drukregime en storingsfrequentie. Deze methodiek op hoofdlijnen is schematisch weergegeven in

Figuur 7.

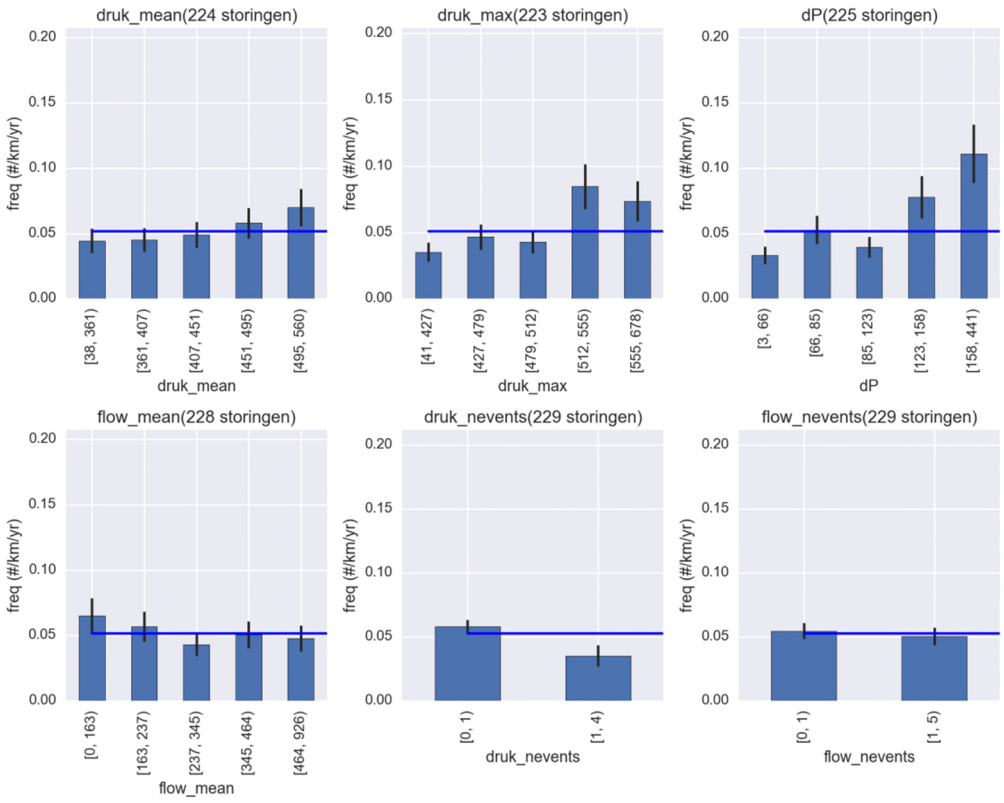


Figuur 7: van databron naar analyse van relaties tussen attributen in casus P

Er is een automatische anomaliedetectie uitgevoerd op alle druk- en volumestroommetingen. Hiervoor is gebruik gemaakt van een tijdreeks-regressiemodel, dat voor elk tijdstip een voorspelling doet voor de te verwachten meetwaarde op basis van type dag (weekdag, zaterdag, zondag) en op basis van het tijdstip van de dag. Vervolgens wordt de daadwerkelijke meetwaarde vergeleken met de modelvoorspelling, waarbij de mate van verschil gebruikt wordt om te bepalen of de meetreeks onverklaarbaar gedrag vertoont of juist in lijn ligt met wat verwacht mag worden. Voor deze methodiek is gebruik gemaakt van een zogenaamd Support Vector Regressiemodel (SVR), volgens de standaard Scikit-learn LibSVM implementatie in Python (Pedregosa et al., 2011). Anomaliedetectie wordt gedaan

binnen een 'moving window' van 20 tijdstappen, waarbij een patroon van meer dan 10 opeenvolgende afwijkingen van meer dan drie standaardafwijkingen ten opzichte van de voorspelling wordt aangemerkt als een anomalie. Deze methodiek is grotendeels gebaseerd op onderzoek van Mounce, Mounce, & Boxall (2011). De events die uit de anomaliedetectie voortkomen zijn als input voor de frequentie-analyse gebruikt. Voor die analyse is gebruik gemaakt van een tweede machine learning techniek, namelijk Gradient Boosting Regression, om de relatie tussen de genoemde attributen en het aantal storingen te bepalen (dit is de 'modelleer' stap uit het CRISP-DM schema, paragraaf 2.2).

Er is bij deze casus een relatie gevonden tussen de storingsfrequentie en het drukregime van pompstations. Bij AC-leidingen laten zowel de gemiddelde druk als de maximale verschildruk op een dag een relatie met de storingsfrequentie zien (Figuur 8). Voor niet-AC leidingen is er geen verband met gemiddelde druk. Er is geen direct verband gevonden tussen het optreden van anomalieën en storingen.

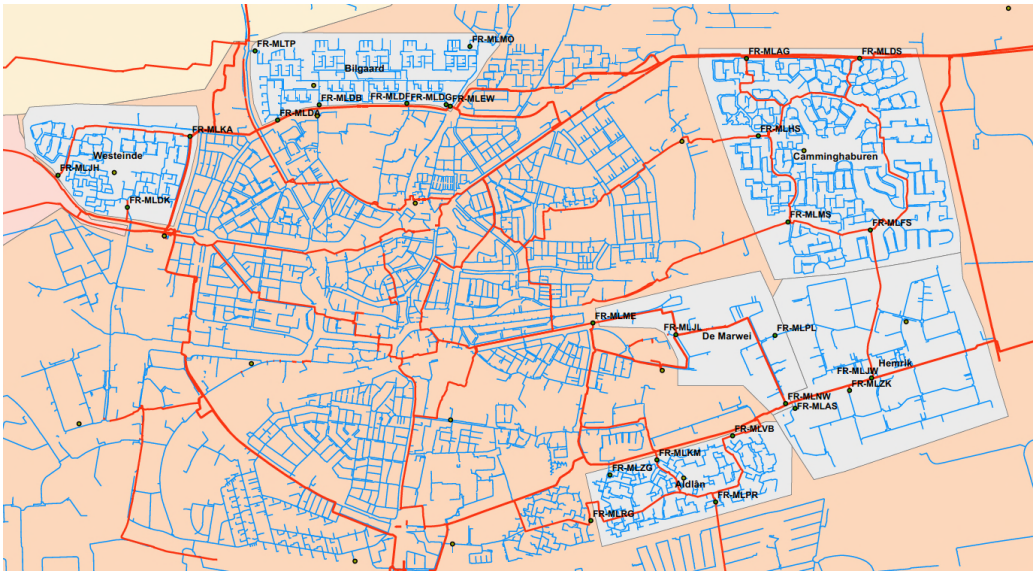


Figuur 8: Storingsfrequentie (alleen AC) per verklarende variabele; opgedeeld in klassen van gelijke grootte: (druk_mean) gemiddelde druk bij meetpunt, (druk_max) maximale druk op een dag, (dP) maximale verschildruk op een dag (flow_mean) gemiddelde volumestroom (m3/h), (druk_nevents) aantal anomalieën drukreeks (druk_nevents)) aantal anomalieën flowreeks (flow_nevents).

5.3 Casus 3 – Anomalieën herkennen uit sensormetingen in het leidingnet

5.3.1 Inleiding

Vitens heeft in 2012 een proeftuin voor het testen van nieuwe waterkwaliteitssensoren in het leidingnet aangelegd in de buurt van Leeuwarden (de zogenaamde Vitens Innovation Playground – VIP). Op elke sensorlocatie wordt zowel druk, volumestroom, geleidbaarheid als watertemperatuur gemeten. De gedachte daarbij is dat met deze grootschalige sensoring in het leidingnet in de toekomst anomalieën als gevolg van allerlei oorzaken gedetecteerd kunnen worden. Hierbij kan gedacht worden aan leidingbreuken, gebruik van brandkranen, bruinwater-situaties en (ernstige) verstoringen of veranderingen in de waterkwaliteit. De meerwaarde van het toepassen van een sensornetwerk in het leidingnet is dat Vitens ook de ambitie heeft om ruimtelijk een nauwkeurigere inschatting te kunnen geven van de locatie waar een anomalie zich bevindt. Hiermee kunnen responstijden voor onderhoudsploegen omlaag worden gebracht en is men in staat om sneller en gerichter buurtbewoners te informeren over dergelijke incidenten.



Figuur 9: Plattegrond van de proeftuin, met daarin de sensorlocaties.

De doelstelling van deze casus was om een methodiek te ontwikkelen waarmee voor elke locatie automatisch anomalieën herkend kunnen worden op basis van de (gecombineerde) meetsignalen voor die locatie. Daarvoor waren de databronnen uit Tabel 6 beschikbaar. Belangrijk kenmerk van deze analyse, ten opzichte van de analyses in casus 1 en 2, is dat het hier *ongelabelde* datasets betreft. Kortom, het doel is om anomalieën te vinden, maar voor een algoritme is vooraf niet bekend wat als anomalie opgemerkt dient te worden. Daarom is dit een unsupervised machine learning-probleem. Er zijn voor deze casus op basis van literatuuronderzoek de drie meest kansrijke methodieken uitgetest op de data: (1) dynamische bandbreedtemonitoring, (2) (residu) clustering en (3) SPIRIT.

Tabel 6: Overzicht databronnen beschikbaar voor casus 3

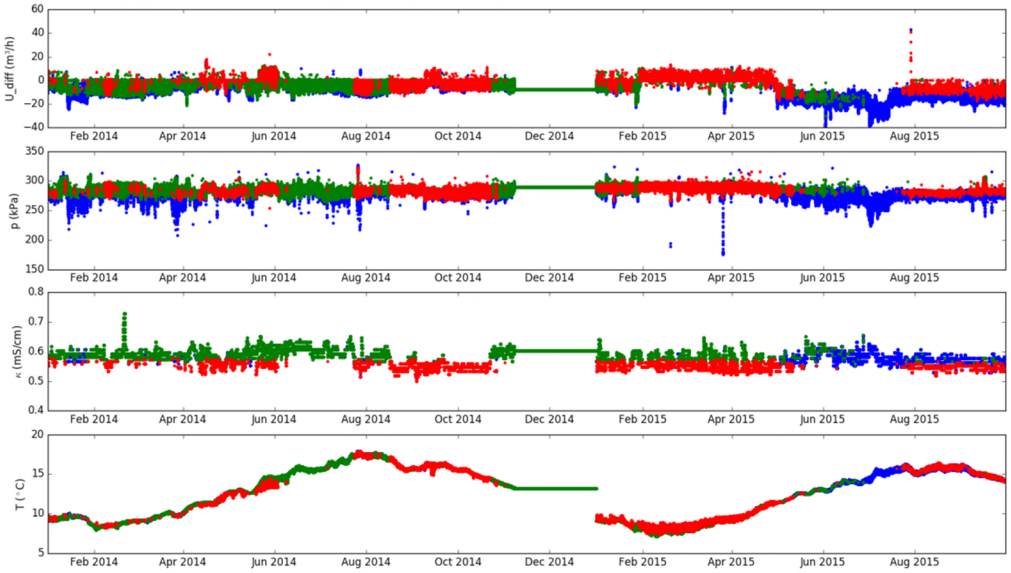
Data	Bronhouder	Meetinterval	Databeheer/ software
Sensorlocaties	Vitens	eenmalig	PDF
Meetreeksen druk	Vitens	onregelmatig (1-5 minuten)	OSISoft PI
Meetreeksen volumestroom	Vitens	onregelmatig (1-5 minuten)	OSISoft PI
Meetreeksen geleidbaarheid (EGV)	Vitens	onregelmatig (5-60 minuten)	OSISoft PI
Meetreeksen watertemperatuur	Vitens	onregelmatig (1-5 minuten)	OSISoft PI

5.3.2 Dynamische bandbreedtemonitoring

De methodiek voor dynamische bandbreedtemonitoring is grotendeels gebaseerd op eerder onderzoek van Mounce et al. (2011). Het systeem gebruikt een regressiemodel om een sensorsignaal te voorspellen op basis van externe variabelen, waarna de voorspelling vergeleken wordt met de daadwerkelijk gemeten waarde op een bepaald tijdstip. Een significant verschil tussen voorspelling en meting, dat structureel (gedurende meerdere tijdstappen) voortduurt kan aangemerkt worden als een anomalie. Dit is in essentie hetzelfde systeem als wat ook gebruikt is voor casus 2, met als verschil dat in deze casus meerdere inputsignalen zijn gebruikt voor het regressiemodel. Zo wordt naast tijdstip van de dag en type dag ook het temperatuursignaal meegenomen als verklarende variabele voor de volumestroom en druk. Twee regressie-modellen zijn uitgetest voor de voorspellings-component van deze methodiek, het Gradient Boosting Regression (GBR) model en Support Vector Regressie (SVR). De achterliggende techniek achter deze modellen is nader toegelicht in Bijlage I.

5.3.3 Clustering

Omdat a priori niet bekend is waar zich in de meetperiode daadwerkelijk anomalieën hebben voorgedaan, is analyse van de meetreeksen per definitie een unsupervised leerprobleem. Voor dergelijke probleemtypen worden vaak clusteringtechnieken gebruikt. Clustering-technieken groeperen alle instanties in een dataset in groepen met soortgelijke eigenschappen. Ook op de tijdreeksen uit de proeftuin zijn een aantal clusteringtechnieken uitgetest. De gedachte daarbij is dat het met clusteringalgoritmes wellicht mogelijk is om alle metingen op te delen in een groep van 'sterk afwijkende metingen', 'matig afwijkende metingen' en een groep 'normale metingen'. De clustering wordt daarbij uitgevoerd op alle signalen van één sensorlocatie tegelijkertijd. Kortom: het algoritme kijkt voor elk tijdstip of de combinatie van volumestroom, druk, temperatuur en geleidbaarheid overeenkomst vertoont met eerdere metingen die gedaan zijn, of dat het gecombineerde meetsignaal juist ver buiten deze groep van overeenkomstige metingen valt. Voor clustering waarbij elk data-attribuut (hier: sensorsignaal, of kenmerk van het signaal) even zwaar meegeteld en geclusterd wordt, is normalisatie noodzakelijk. Daartoe zijn alle signalen eerst (per dagtype) gecorrigeerd voor hun uurgemiddelde waarden, zodat natuurlijke periodiciteiten eruit gefilterd worden. De resterende signalen zijn in feite de 'residuen'; de restanten in het signaal die niet afwijken van het standaardpatroon. Deze residuen zijn naar hun variantie ten opzichte van het gemiddelde (z-score standaardisatie) geschaald. Op deze dataset met genormaliseerde residu-signalen zijn diverse clusteringalgoritmes uitgetest, waaronder k-Means en een aantal variaties op Gaussian Mixture Models (GMM), volgens de Scikit-learn implementatie van LibSVM in Python (Pedregosa et al., 2011).



Figuur 10: Resultaat residuclustering voor een sensorlocatie in de proeftuin. Blauwe datapunten zijn relatief sterk afwijkend ten opzichte van de rest van het signaal, terwijl rode en groene punten als 'normaal' aangemerkt kunnen worden.

In Figuur 10 lijkt er een onderscheid in verschillende clusters waar te nemen die relateren aan anomalieën (blauw) en reguliere data (groen en rood) met onderscheid in sensoruitval (groen). Er zijn echter ook resultaten waarbij de clustering niet zo duidelijk anomalieën kan onderscheiden, zie KWR-rapport 2015.108 (Vries, Vonk, et al., 2015).

De resultaten tonen dat de toepassing van clustering voor- en nadelen met zich meebrengt:

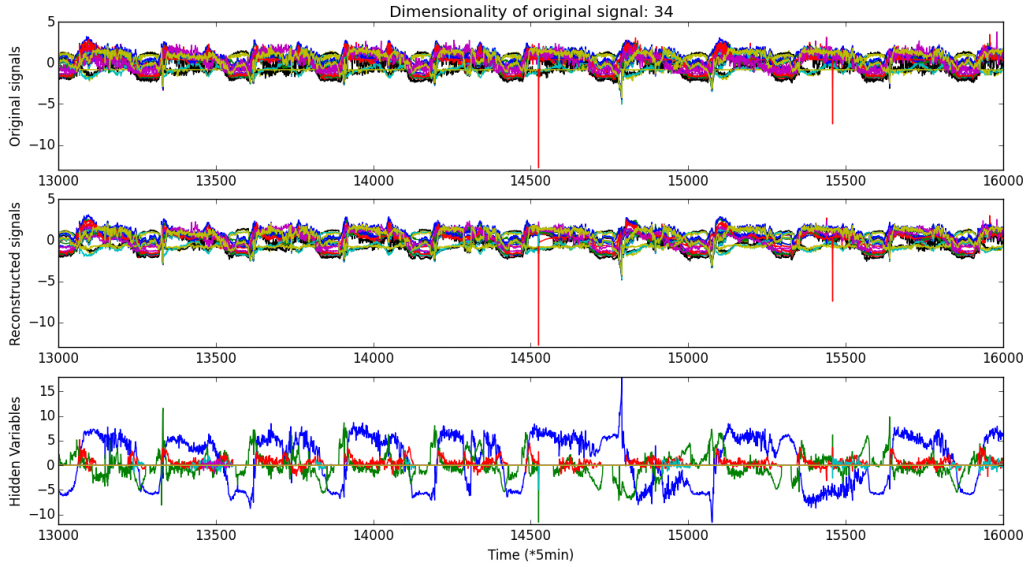
- Het (vermeende) voordeel van het gebruik van clustering met de huidige aanpak is dat kruiscorrelaties benut kunnen worden om te bepalen of ergens een anomalie optreedt. Per tijdstip wordt op basis van de 4 signalen (EGV, volumestroom, druk en temperatuur) gemonitord in welk cluster de meting valt. Sensoruitval (door bijvoorbeeld stroomuitval) behoort tot één cluster.
- Een nadeel is dat er bij de algoritmes *a priori* moet worden opgegeven hoeveel clusters uiteindelijk worden gevormd. Wanneer een meetreeks geen (extreme) anomalieën bevat, dan zal het clusteralgoritme toch proberen om het *a priori* opgegeven aantal clusters te vullen, met als gevolg dat er een vastgelegd aantal clusters zijn waarbij niet eenvoudig een interpretatie per cluster kan worden gegeven.
- Een ander nadeel is de moeilijkheid om causaliteit en dynamica naar attributen voor clustering te vertalen. Bij clustering wordt standaard gekeken naar hoe een bepaalde meting zich verhoudt tot de andere metingen in een meetreeks, bepaald door een afstandscriterium. Dynamische patronen in meetreeksen worden standaard niet meegenomen in de clusteringanalyse. De methodiek is het meest gevoelig voor een bepaalde, substantiële hoeveelheid data die met een bepaalde afstand van andere (ook substantiële hoeveelheid) data liggen. Daardoor kan een klein aantal ongebruikelijk hoge meetwaarden in een cluster met reguliere waarden worden ingedeeld.

5.3.4 SPIRIT

Een derde aanpak die in dit onderzoek getest is werkt, evenals de clustering, ook op basis van alle sensorsignalen tegelijkertijd. De gezamenlijke meetsignalen (volumestroom, druk, EGV en temperatuur) per meetlocatie kunnen gezien worden als een vierdimensionaal signaal. Het 'Streaming Pattern dIscoveRy on multIple Time-series' (SPIRIT) algoritme neemt dit vierdimensionale signaal en tracht op elke tijdstap het aantal dimensies zover mogelijk te reduceren tot een variabel, kleiner aantal 'verborgen variabelen'. Deze verborgen variabelen vormen samen in feite de basiscomponenten van het uiteindelijke vierdimensionale signaal. Op het moment dat plotseling in de tijdreeks veel meer of juist minder verborgen variabelen nodig zijn om toch het eindsignaal binnen een bepaalde nauwkeurigheid (bijvoorbeeld 95% van de variantie in het eindsignaal) te kunnen beschrijven, dan is dit een indicatie dat er een anomalie optreedt. Er kunnen meerdere meetlocaties per (real-time) analyse worden meegenomen, zoals in het voorbeeld van Figuur 11 waarbij druk en volumestroom van 17 locaties zijn verwerkt. In dit voorbeeld verschijnt een extra verborgen variabele (rode doorgetrokken lijn) gedurende een (door het algoritme gedetecteerd) afwijkende volumestroom, totdat de volumestroom weer normale waarden bereikt.

Er kleven meerdere nadelen aan het gebruik van het SPIRIT-algoritme:

- het algoritme vereist veel kalibratie (o.a. instellen van drempelwaarde van energieniveau, maar ook tijdsduur waarin een verandering mag optreden) om een verandering in aantal verborgen variabelen te laten overeenstemmen met (aangenomen) anomalieën;
- het algoritme is niet in staat om abrupte sensorfouten (bijvoorbeeld extreme maximumwaarden of nulwaarden) als anomalie te detecteren, omdat het foutieve signaal met evenveel variabelen kan worden gevangen als gedurende een reguliere situatie (bijvoorbeeld een volumestroomsensor die nulwaarden registreert gedurende de nacht).



Figuur 11: Voorbeeld van toepassing van het SPIRIT-algoritme op genormaliseerde druk- en volumestroommetingen afkomstig van 17 meetlocaties simultaan.

5.3.5 Resultaat

Het onderzoek binnen deze casus heeft niet geleid tot een methodiek die voldoende vertrouwen geeft om uit sensordata consistent en automatisch anomalieën te detecteren. De methodieken die verkend zijn hebben elk hun eigen nadelen en beperkingen. Zo is de dynamische bandbreedtemonitoring niet altijd voldoende nauwkeurig, omdat er bij een aantal meetlocaties gemeten wordt aan een sensor in een relatief kleine leiding met daardoor een onregelmatig(er) waterverbruik. Door de onregelmatigheid in de onderzochte signalen wordt het vaak een zeer open vraagstelling wat als anomalie aan te merken valt en wat niet. Verder zegt de kwaliteit van de modelfit bij de dynamische bandbreedtemonitoring niets over of de anomaliedetectie goed werkt, omdat de correlatiecoëfficiënt alleen iets zegt over de fit met het signaal. Bij de onderzochte methoden spelen ook bandbreedte en/of een cluster met een substantiële hoeveelheid geregistreerde anomalieën. Een goed voorbeeld zijn de EGV-signalen die bij de GBR-bandbreedtemethodiek een goede correlatie van de het GBR-model laten zien, maar tegelijkertijd ook een erg grote bandbreedte door de variatie in het patroon; waardoor nauwelijks anomalieën optreden.

Clustering heeft als voordeel dat alle signalen op een locatie tegelijk met elkaar vergeleken worden, maar de clusters zijn niet altijd eenvoudig te interpreteren en vereist (complexe) attributen om dynamiek van het signaal mee te nemen. Tot slot blijkt het gebruik van het SPIRIT-algoritme sterk mee te deinen met de variatie en dynamiek van de sensorsignalen. Dit betekent dat twee verborgen variabelen vaak 95 tot 98% van (de energie in het) signaal kunnen vatten, terwijl onder sommige omstandigheden slechts één variabele voldoende is (bijvoorbeeld bij sensoruitval, of gedurende een nachtpatroon van het volumestroomsignaal). Uitschieters door bijvoorbeeld een lek werden vaak met hetzelfde aantal verborgen variabelen meegenomen. Eenduidige resultaten in verandering van het aantal verborgen variabelen (of de eigenwaarden ervan), gekoppeld met (verwachte) anomalieën werden niet bereikt.

Een belangrijke les uit deze casus is dat het waardevol is om óf gelabelde anomalieën te gebruiken voor training van algoritmes, zodat prestaties van de algoritmes hiermee onderling vergeleken kunnen worden óf gebruik te maken van gevalideerde leidingnetmodellen die een verwacht patroon in waterkwaliteit en –levering accuraat kan voorspellen (via data-assimilatie, zie bijvoorbeeld Vries, Akker, & Van Summeren (2015)) zodat anomalieën eenduidig kunnen worden gedetecteerd.

5.4 De belangrijkste lessen uit de praktijk

Uit de praktijk van de drie casussen is duidelijk geworden dat het zogenaamde '*feature engineering*' een van de belangrijkste stappen in het datamining proces is. Feature engineering is het vertalen en opwerken van 'ruwe' invoer naar parameters die geschikt zijn om in het machine learning algoritme te stoppen. In casus 1 zijn de vermazingsgraad en dimensioneringsmaat voorbeelden van feature engineering. In casus 2 worden storingsregistraties berekend en worden parameters geformuleerd die drukregimes definiëren. In casus 3 is bij clusteren gebruik gemaakt van attributen die het tijdstip en dag van de week vastleggen. Voor feature engineering is vakspecifieke kennis onontbeerlijk. Dat feature engineering de sleutel is tot succesvol dataminen is ook een van conclusies in het werk van Domingos (2012).

Naast de genoemde feature engineering maken verschillende zaken samen datamining tot een succes:

- *Technisch:* software en/of een programmeeromgeving is nodig die het experimenteren met meerdere verschillende machine learning modellen, databronnen en problemen kan ondersteunen. Directe (zo nodig afgeschermde) toegang tot de databronnen voor data-analisten versnellen datamining en de data-analyse.
- *Organisatorisch:* een nauwe samenwerking tussen data-analisten en vakinhoudelijke deskundigen biedt meer garantie dat databronnen juist geïnterpreteerd en gekoppeld worden. Ook hier geldt dat een open en constructieve houding van alle betrokken partijen, inclusief organisatorische en projectmatige ondersteuning van de opdrachtgever, de kans op synergetische voordelen sterk kan vergroten.

Een laatste conclusie die getrokken kan worden uit de eerste praktijkervaringen is dat verhoudingsgewijs de huidige data nog verre van ‘big data’ te noemen is. Hiervan is pas sprake indien bijvoorbeeld grootschalige sensornetwerken hoogfrequente metingen zouden doorsturen naar een centraal punt, alwaar het opgeslagen en/of verwerkt zou worden. Het gaat dan om ordegrootte gigabytes aan data per seconde. Ter vergelijking: de datasets die voor genoemde casussen zijn gebruikt hebben een omvang van een ordegrootte van ten hoogste 2 GB. Dat is niet de big data waarvan wordt verwacht dat de verwerking ervan (ordegrootte Terabytes, i.e. duizenden Gigabytes) schaalbare databases en cloud-oplossingen vergen.

6 Conclusies en aanbevelingen

Het doel van dit onderzoek was geformuleerd als 'het inventariseren welke mogelijkheden datamining biedt om datasets van drinkwaterbedrijven om te vormen tot waardevolle kennis waarmee het assetmanagement verbeterd kan worden'. De inventarisatie is op basis van een literatuurscan, een enquête en een beschrijving van casussen uitgevoerd.

6.1 Conclusies

- Datamining wordt gezien als onderdeel van de bredere 'knowledge discovery' procedure, sterker, zorgvuldige selectie en het maken van combinaties met verschillende data is een belangrijke sleutel hierin. De methodiek is daarom gevoelig voor misinterpretatie. Resultaten dienen door een data-analist, vakdeskundigen en opdrachtgever altijd kritisch bekeken te worden met het oog op veelvoorkomende valkuilen: gebruik van niet-representatieve datasets, overfitting, data dredging en niet-causaliteit.
- Datamining stelt eisen aan databeschikbaarheid en -kwaliteit. Aangezien datamining geheel datagedreven is, zal een verkregen model hooguit zo goed zijn als de data waarmee het gemaakt is. Veelvoorkomende problemen omtrent databeschikbaarheid zijn terug te voeren tot een viertal categorieën: ontbrekende labels, te weinig 'events', een te korte meetperiode of een te lage meetfrequentie. Datakwaliteit is een ander aspect, dat betrekking heeft op mogelijke onjuistheden in de datasets, variërend van falende sensoren tot menselijke fouten.
- Uit een vragenronde bij waterbedrijven blijkt dat assetmanagement van drinkwaterbedrijven gebaat is bij kennis van zowel de operationele prestatie van een asset, de actuele conditie als de storingsrisico's. Daarnaast is er een tweede as: de behoefte aan 'real-time' inzicht bij gebruik binnen het primaire bedrijfsproces, versus een meer 'periodieke' raadpleging van kennisbronnen (expertsystemen) bij de tragere bedrijfsprocessen, zoals bij het opstellen van onderhouds- en investeringsplannen. 'Knowledge discovery' kan projectmatig georganiseerd en in de organisatie ingebed worden, dusdanig dat permanente expertsystemen worden gevoed en de gevraagde informatie leveren.
- Uit drie uitgelichte casussen volgt dat datamining geschikt is voor een diversiteit aan vraagstukken: variërend van kennisvragen rondom storingsfrequenties van assets tot kennisvragen rondom operationele prestaties. Voornaamste lessen uit de drie praktijk-casussen zijn het belang van ruim voldoende datakwaliteit en -hoeveelheid, feature engineering, samenwerking met vakspecialisten op het gebied van waterinfrastructuur en operationele processen en tenslotte de beperkte hoeveelheid beschikbare data.
- Van 'big data' is nog geen sprake bij de waterbedrijven. Toch lijkt er door een gestage toename aan sensoren in het leidingnet, slimme meters, groter wordende databases met meetgegevens en storingsregistraties meer en meer mogelijkheden te komen om middels datamining relevante vraagstukken voor de drinkwatersector te beantwoorden.

6.2 Aanbevelingen

Op basis van de bevindingen in deze studie bevelen wij aan om:

- bij de uitvoering van datamining projecten veel tijd en aandacht te besteden aan de feature engineering stap;
- dataspecialisten te laten samenwerken met vakspecialisten op het gebied van waterinfrastructuur, teneinde een correcte interpretatie van databronnen te waarborgen en resultaten te interpreteren;
- om bij het formuleren van een datamining project altijd vooraf al na te denken over de beschikbaarheid van voldoende databronnen. Deze vormen namelijk altijd de voornaamste beperking bij analyses en zijn doorslaggevend bij de vraag of datamining überhaupt haalbaar is.

7 Referenties

Auria, L., & Moro, R. A. (2008). Support Vector Machines (SVM) as a Technique for Solvency Analysis. *DIW Discussion Papers*. German Institute for Economic Research.

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support Vector Regression. *Neural Information Processing - Letters and Reviews*, *11*, 203–224.

Beuken, R. (2015). *Prestatie-indicatoren en stuurparameters voor het distributienet*. Nieuwegein: KWR Watercycle Research Institute.

Bifet, A. (2013). Mining Big Data in Real Time. *Informatica*, *37*, 15–20.

Bramer, M. (2007). *Principles of Data Mining*. London: Springer-Verlag.

Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Communications of the ACM*, *55*, 78–86.

Esling, P., & Agon, C. (2012). Time-Series Data Mining. *ACM Computing Surveys*, *45*, 12–25.

Flach, P. (2012). *Machine Learning - The Art and Science of Algorithms that Make Sense of Data*. Cambridge: Cambridge University Press.

GÃ¼ler, C., Thyne, G. D., McCray, J. E., & Turner, K. A. (2002). Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. *Hydrogeology Journal*, *10*, 455–474.

George, P. J. P. Chen Z. Shaw. (2009). Fault detection of drinking water treatment process using PCA and Hoetellingâ€™s T2 chart. *World Academy of Science, Engineering and Technology*, *50*, 970–975.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. New York: Springer.

Hill, D. J., & Minsker, B. S. (2010). Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling & Software*, *25*, 1014–1022.

Hotelling, H. (1933). Analysis of a complex of statistical variables in a factor analysis. *Psychometrika*, *30*, 179–185.

IBM. (2011). IBM SPSS Modeler CRISP-DM Guide. IBM Corporation. Retrieved from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf

Iyer, S., Sinha, S. K., Tittmann, B. R., & Pedrick, M. K. (2012). Ultrasonic signal processing methods for detection of defects in concrete pipes. *Automation in Construction*, *22*, 135–148.

Jolliffe, L. T. (2002). *Principle Component Analysis*. New York, NY, U.S.A.: Springer-Verlag.

KWR. (2013). *BTO Trends - Effectenstudie Big Data*. Nieuwegein: KWR Watercycle Research Institute.

Mashford, J., De Silva, D., Marney, D., & Burn, S. (2009). An Approach to Leak Detection in Pipe Networks Using Analysis of Monitored Pressure Values by Support Vector Machine. *Network and System Security, 2009. NSS '09. Third International Conference on* (pp. 534–539). doi:10.1109/NSS.2009.38

Mounce, S. R., Mounce, R. B., & Boxall, J. B. (2011). Novelty detection for time series data analysis in water distribution systems using support vector machines. *Journal of Hydroinformatics*, *13.4*, 672–686.

Muralidharan, V., & Sugumaran, V. (2012). A comparative study of NaÃ¯ve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis. *Applied Soft Computing*, *12*, 2023–2029.

Pal, M., & Mather, P. M. (2005). Support vector machines for classification in remote sensing. *International Journal of Remote Sensing*, *26*, 1007–1011.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Phil. Mag.*, *6*(2), 559–572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, A., Thirion, B. V., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Praus, P. (2005). Water quality assessment using SVD-based principle component analysis of hydroilogical data. *Water SA*, *31*(4).

Quinlan, J. R. (1979). Discovering rules by induction from large collections of examples. *Expert systems in the micro electronic age*. Edinburgh: Edinburgh University Press.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo: Morgan Kaufmann Publishers.

Ratanamahatana, C., Lin, J., Gunopulos, D., Keogh, E., Vlachos, M., & Das, G. (2010). *Data Mining and Knowledge Discovery Handbook* (pp. 1049–1077). Springer US.

Rish, I. (2001). An empirical study of the naive Bayes classifier. *Proceedings of IJCAI-01 workshop on Empirical Methods in AI*. Sicily.

Ruiz, J. M. Colomer J. Melendez. (2006). *Multiway principal component analysis and case base-reasoning approach to situation assessment in a wastewater treatment plant*. Universidad Politecnica de Cataluna.

Sagiroglu, S., & Sinanc, D. (2013). Big Data: A Review. *International Conference on Collaboration Technologies and Systems (CTS)* (pp. 42-47). IEEE.

Von Asmuth, J. R., & Van Geer, F. C. (2013). *Kwaliteitsborging grondwaterstands- en stijghoogtegegevens: op weg naar een landelijke standaard*. Nieuwegein/Utrecht: KWR Watercycle Research Institute / TNO.

Vonk, E., Vries, D., Van Summeren, J. R. G., Verbree, J. M., Wols, B. A., & Raterman, B. W. (2015). *Kennis uit waterdata in en rondom het leidingnet*. KWR - BTO 2016.006.

Vries, D., Akker, B. Van den, & Van Summeren, J. (2015). *Prototype softsensor voor het kalkafzettend vermogen van leidingwater - BTO 2015.030*. Nieuwegein: KWR Watercycle Research Institute.

Vries, D., Vonk, E., Jong, W. de, Van Duist, H., & Marel, J. Van der P. Van der Wielen. (2015). *Herkennen van anomalieën in waterdata: demo in de Vitens proeftuin* (No. KWR 2015.108). KWR Watercycle Research Institute.

Widodo, A., & Yang, B. S. (2007). Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing*, *21*, 2560-2574.

Wols, B. B.A Van Summeren J. Mesman G. Raterman. (2015). *Fysieke kwetsbaarheid leidingen voor klimaatverandering*. KWR - BTO.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*, 1-37.

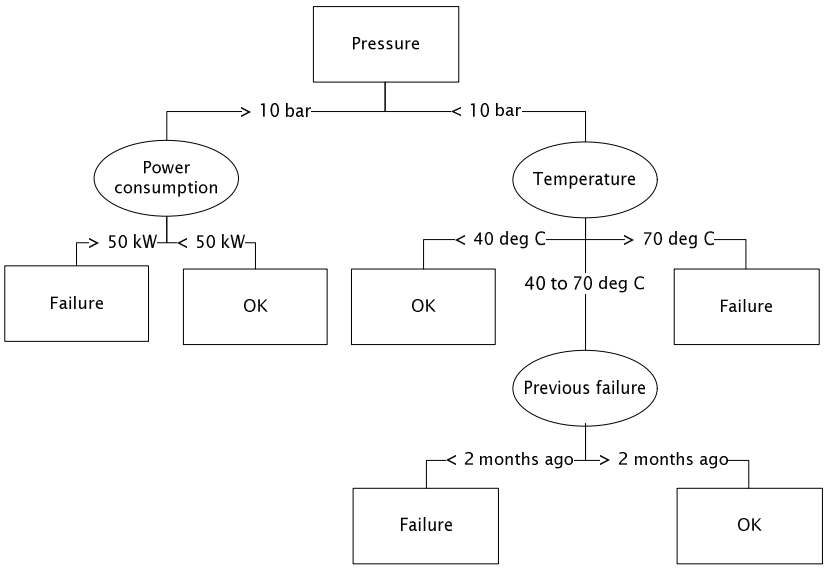
Zhang, H. (2004). The Optimality of Naïve Bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*. Miami Beach: AAAI Press.

Bijlage I Modellen voor machine learning

Logische modellen

Beslisbomen (decision tree learning)

De afgelopen jaren zijn er talrijke algoritmes ontwikkeld om machinaal beslisbomen af te leiden. Met dergelijke beslisbomen kan op een natuurlijke manier informatie geclassificeerd worden. Ook is het mogelijk om beslisbomen te gebruiken als regressiemodel (waarmee een stuksgewijs lineaire functie gecreëerd wordt). Het grote voordeel van beslisbomen is de inzichtelijkheid voor een gebruiker; het resultaat is een zogenaamd 'white box model'. Daarnaast kunnen beslisbomen door softwareontwikkelaars ook eenvoudig in programmeercode omgezet worden. In Figuur 12 is een grafisch voorbeeld weergegeven van een beslisboom die middels machine learning is afgeleid.



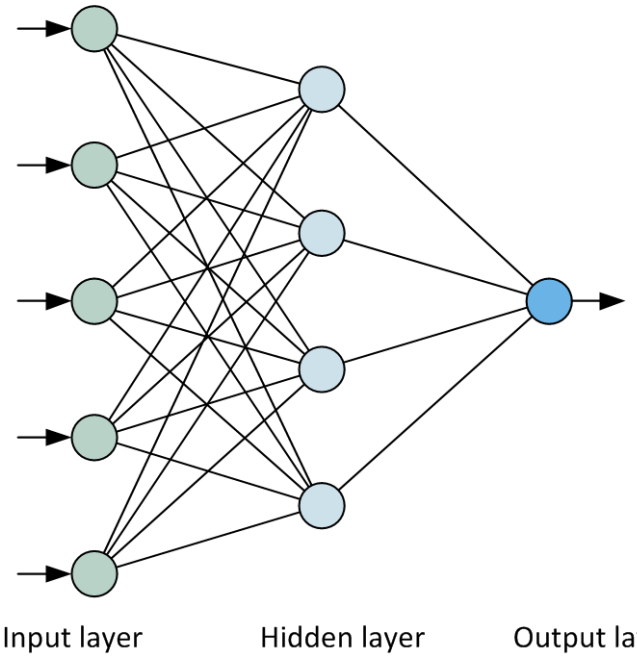
Figuur 12: Voorbeeld van een beslisboom. Elke knoop in de boomstructuur is een relevante attribuut van de dataset waarop het machine learning model getraind is.

Decision tree learning is typisch geschikt voor problemen waarbij in datasets veel categorische variabelen aanwezig zijn (zoals een variabele 'temperatuur', die bijvoorbeeld de waarde 'warm', 'lauw' of 'koud' kan hebben). Ook zijn beslisbomen relatief robuust in situaties waarbij vooraf bekend is dat er veel foutieve, ontbrekende of onnauwkeurige metingen in een dataset zitten. Doorgaans worden beslisbomen getraind met een zogenaamde 'divide-and-conquer' algoritme. Een dergelijk algoritme begint bovenaan de boom en bepaalt voor iedere nieuwe vertakking welke variabele op dat moment de grootste bijdrage levert aan een correcte classificatie (de zogenaamde 'information gain' is hierbij een maat voor de kwaliteit van de gemaakte vertakking). Er bestaat een grote diversiteit aan algoritmes voor het machinaal afleiden van beslisbomen. Enkele bekende zijn CART (Breiman et al., 1984), ID3 (Quinlan, 1979) en diens opvolger C4.5 (Quinlan, 1993).

Een belangrijk aandachtspunt bij het werken met de genoemde algoritmes is de kans op overfitting. Dit is met name een risico bij relatief kleine datasets en wanneer er veel ruis in de data aanwezig is. Bij dergelijke gevallen ontstaan beslisbomen waar lange vertakkingen aan groeien. Een techniek die in dergelijke situaties vaak wordt toegepast is 'pruning' (analoog aan het 'snoeien' van een boom). Bij het zogenaamde 'reduced-error pruning' wordt, nadat een boom machinaal geconstrueerd is, uitgezocht hoeveel elke vertakking bijdraagt aan de totale kwaliteit van de modelvoorspellingen. Takken die relatief veel complexiteit toevoegen aan het model, maar daarentegen slechts een minimale invloed hebben op het validatieresultaat, worden uit de beslisboom verwijderd. Een alternatief is 'rule post-pruning', waarbij de beslisboom na training geconverteerd wordt tot een set beslisregels. Elke regel is daarbij de route vanaf de bovenkant van de beslisboom tot een uitkomst onderaan. Rule post-pruning rangschikt de toegevoegde waarde van alle beslisregels door voor elke regel te evalueren welk effect het heeft als bepaalde termen hieruit verwijderd worden (Bramer, 2007).

Neurale netwerken

Kunstmatige Neurale Netwerken (KNN) zijn statistische simulatiemodellen die gebaseerd zijn op de manier waarop biologische zenuwstelsels, zoals het menselijk brein, inwendig georganiseerd zijn. Net als in een biologisch brein bestaat een KNN uit een uitgebreide verzameling neuronen die onderling met elkaar verbonden zijn. De onderlinge verbindingen tussen neuronen kunnen worden versterkt of verzwakt. Hierdoor is een neuraal netwerk te trainen op het uitvoeren van bepaalde taken, zoals bijvoorbeeld het geven van een juiste uitkomst op basis van een bepaalde set invoergegevens. Eenmaal getraind is het KNN als model inzetbaar op ongelabelde data voor een specifiek doel: het herkennen van continue verbanden uit data (regressie). Eventueel is het ook mogelijk om overeenkomstige gebieden uit een dataset groeperen (clusteren) of om data indelen in klassen (classificatie), maar dit zijn minder gangbare toepassingsgebieden.



Figuur 13: Grafische weergave van een eenvoudig feedforward neuraal netwerk (multilayer perceptron).

Een ‘feed-forward’ kunstmatig neuraal netwerk (‘perceptron’) bestaat uit een laagsgewijze rangschikking van afzonderlijke neuronen die middels onderlinge verbindingen signalen naar elkaar verzenden. De verbindingen op zichzelf worden gekenmerkt doordat ze een ieder een unieke weegfactor hebben die de erdoor verstuurde signalen kan versterken of verzwakken. Elk neuron ontvangt via de inkomende verbindingen één of meerdere invoersignalen, waaronder (optioneel) een zogenaamd bias-signaal. De inkomende signalen worden in het neuron getransformeerd tot een zeker uitvoersignaal. Daartoe worden de inkomende invoersignalen vermenigvuldigd met de weegfactoren van iedere verbinding, waarna hieruit de gewogen som berekend wordt. Middels een lineaire, sigmoïde of stapsgewijze transformatiefunctie genereert het neuron vervolgens één enkel uitvoersignaal, dat door alle uitgaande verbindingen wordt verzonden naar neuronen elders in het neurale netwerk.

Het aantal neuronen in een netwerk, de onderlinge verbindingen tussen de neuronen, de activatiefuncties van de neuronen en het aantal netwerklagen bepalen samen de zogenaamde netwerkarchitectuur. Deze is afhankelijk van de specifieke taak waarvoor het netwerk is ontworpen. Een netwerk bestaat in ieder geval uit een invoerlaag en een uitvoerlaag, met daartussen nog een variabel aantal verborgen lagen. De kwaliteit van het KNN wordt, naast de kwaliteit van de data en de gekozen architectuur, bepaald door het zogenaamde trainingsalgoritme. Een dergelijk algoritme optimaliseert de gewichten in het netwerk zodanig dat de gewenste uitkomst zo goed mogelijk wordt benaderd. De kwaliteit en instellingen van het trainingsalgoritme zijn van grote invloed op de uiteindelijke kwaliteit van het netwerk. Voor eenvoudige netwerken is het veelgebruikte backpropagation algoritme erg geschikt. Bij complexere KNN’s moet vaak gezocht worden naar geavanceerdere optimalisatietechnieken, zoals Levenberg-Marquardt of zogenaamde metaheuristieken. Een nadeel van KNN is dat het model als ‘black box’ moet worden beschouwd: de gewichten geven weinig inzicht in de onderliggende wetmatigheden.

Toepassingen - Er is een breed scala aan toepassingen voor KNN’s. De techniek is een essentiële component in software voor handschriftherkenning en spraakherkenning. Daarnaast wordt het ook ingezet voor het automatisch analyseren van camerabeelden en het herkennen van objecten in dergelijke beelden. Binnen de watersector blijken KNN’s nuttig voor regressie-doeleinden, bijvoorbeeld als operationeel voorspellingsmodel voor de watervraag en rivierafvoer. Ook kan het ingezet worden als virtuele sensor, om empirisch te voorspellen. Voorbeelden van mogelijk te voorspellen parameters zijn: SI en TACC in het distributienet (op basis van calciumconcentraties, temperatuur, zuurgraad), microbiologische activiteit in het distributienet (op basis van leidingtemperatuur, pH en verblijfstijden), leidingfalen (op basis van drukvariaties, sluitfouten in volumestromen, leidingparameters (ouderdom, materiaal, etc.) en omgevingsfactoren (grondsoort, weer, graafwerkzaamheden). Als zodanig zijn er ook recente voorstellen om KNN’s te gebruiken voor datagedreven anomaliedetectie bij sensornetwerken (Hill & Minsker, 2010).

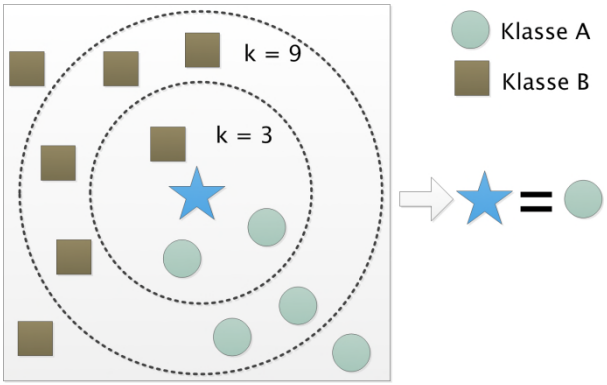
Geometrische modellen

Lineaire regressie

Het lineaire regressiemodel is een veelgebruikte en krachtige methode uit de statistiek voor het bepalen van lineaire verbanden tussen variabelen. Er wordt met deze methode een lineaire functie 'gepast' op een set gegevens. Het bepalen van de 'best passende lijn' wordt doorgaans gedaan met de kleinste-kwadratenmethode, waarbij van elke punt uit de dataset het verschil wordt berekend met de lineaire functie, waarna vervolgens dit verschil gekwadrateerd en gesommeerd wordt. Deze kwadratensom wordt daarna geminimaliseerd om de best passende lijn te vinden. Er zijn uitbreidingen van deze methoden naar datasets met meerdere attributen. Dit wordt 'meervoudige lineaire regressie' genoemd.

K-Nearest Neighbors (kNN)

Het kNN algoritme is geschikt voor het classificeren van datasets. Daartoe wordt telkens een nieuwe ongelabelde instantie uit de dataset aangeboden, waarna het algoritme deze instantie vergelijkt met een voorgedefinieerd aantal K instanties uit de dataset die de grootste gelijkenis vertonen met de geselecteerde instantie (de zogenaamde 'nearest neighbors'). Het algoritme wijst vervolgens de geselecteerde instantie toe aan een bepaalde klasse op basis van de klasse waartoe de meest overeenkomstige instanties toe behoren. Bij gebruik van een kNN algoritme moet een zogenaamde 'afstandsmaat' gekozen worden. Deze afstandsmaat, feitelijk een functie die de mate van verschil tussen instanties bepaalt (Wu et al., 2008), is tevens de achilleshiel van het kNN algoritme. Als deze maat niet zo gekozen wordt dat belangrijke factoren zwaar wegen, dan zijn de resultaten suboptimaal.



Figuur 14: De werking van het kNN-algoritme grafisch weergegeven.

Toepassingen – Binnen de watersector zijn nog niet veel toepassingen van het kNN algoritme bekend. Dit komt omdat veruit de meeste problemen in de watersector gaan over regressie, terwijl kNN een classificatiemethode is. Wel is het recentelijk in een Amerikaanse studie onderzocht als onderdeel van een methodiek om defecten in betonnen leidingen op te sporen (Iyer, Sinha, Tittmann, & Pedrick, 2012). Het kNN algoritme maakte daarbij onderdeel uit van een grotere aaneenschakeling van verschillende machine learning algoritmes die elk een specifieke taak uitvoeren.

K-means clustering

Clustering is een vorm van unsupervised learning, waarbij individuele instanties van een dataset gegroepeerd worden op basis van overeenkomstige attributen. Bij K-means clustering kiest een data-analist vooraf hoeveel clusters er tijdens de analyse gevormd moeten worden, door een K aantal zogenaamde 'zwaartepunten' voor de clustering te definiëren. Het zwaartepunt (ook wel 'centroïde' of 'prototype' genoemd) vormt het middelpunt van een cluster. Tijdens het eigenlijke clusterproces doorloopt het algoritme

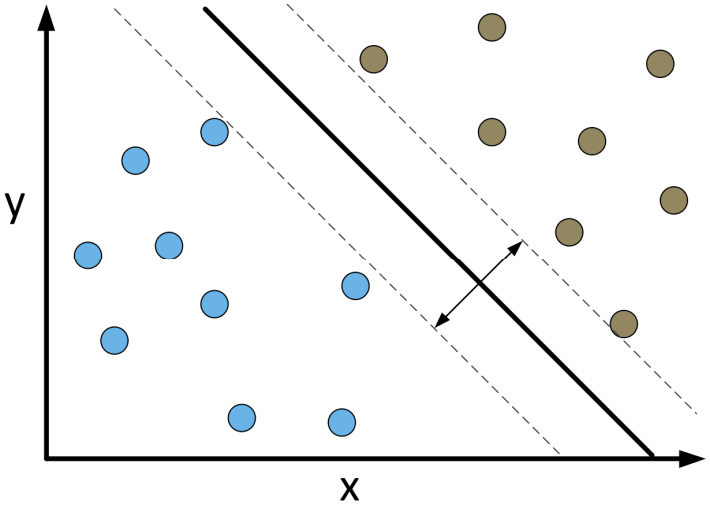
stapsgewijs alle instanties in een dataset, waarbij elke instantie vervolgens toegewezen wordt aan het cluster waarvan het zwaartepunt het dichtste in de buurt ligt. Doordat instanties stapsgewijs toegewezen worden aan alle clusters, veranderen ook de zwaartepunten daarvan. Daarom wordt tijdens elke iteratiestap een herberekening van de cluster-zwaartepunten gedaan, totdat deze uiteindelijk niet meer veranderen. Om de 'afstand' tussen een instantie en een cluster-zwaartepunt te bepalen wordt een zekere afstandsmaat gedefinieerd: de zogenaamde 'similarity measure' (Hastie, Tibshirani, & Friedman, 2009). Net als bij het kNN algoritme is ook hier de keuze voor een juiste afstandsmaat cruciaal voor een correcte clustering.

Figuur 15: Grafische weergave van het k-means clustering algoritme.

Toepassing – In een recent onderzoek van (GÃ¼ler, Thyne, McCray, & Turner, 2002) is k-means clustering gebruikt om verschillende typen chemicaliën in grondwater te groeperen tot clusters van soortgelijke verontreinigingen. Het doel hiervan is om in hydrologische studies eenvoudiger het gedrag van diverse soorten chemicaliën te kunnen bestuderen, zonder elk bestandsdeel afzonderlijk te beschouwen.

Support Vector Machines (SVM's)

Support Vector Machines zijn zogenaamde kernelgebaseerde technieken, een groep van algoritmes voor gecontroleerd machinaal leren met name gebruikt kunnen worden voor zowel classificatie van data. Recentlijk zijn er echter ook uitbreidingen op deze techniek gekomen, waardoor regressie ook mogelijk wordt. In de meest eenvoudige vorm bevat een dataset een zekere hoeveelheid instanties, waarbij elke instantie een x-coördinaat en een y-coördinaat heeft en daarnaast tot



Figuur 16: De grenslijn tussen twee dataklassen, met daaromheen de marge die door de SVM gemaximaliseerd wordt.

een zekere klasse behoort. Met een SVM kan dan in het x-y vlak de best passende grenslijn gezocht worden die de ene klasse van de andere onderscheidt.

De optimale grenslijn tussen twee klassen wordt gevonden middels het maximaliseren van de zogenaamde marge tussen de grenslijn en de set punten die het dichtste bij de grenslijn liggen. Deze maximalisatie is een wiskundig optimalisatieprobleem dat bekend staat als kwadratisch programmeren (Basak, Pal, & Patranabis, 2007). Hoewel in dit voorbeeld slechts twee attributen (x en y) in beschouwing genomen zijn, kunnen SVMs ook gebruikt worden voor het classificeren op basis van tientallen attributen (waarbij de grenslijn een meerdimensionaal vlak wordt, ook wel een hypervlak genoemd). Tevens hoeft een scheidslijn niet lineair te zijn; door de data te transformeren kunnen ook niet-lineaire grenzen worden bepaald. Naast lineaire kernelfuncties kunnen ook polynomen, Radiale Basis Functies (RBF) en sigmoïde functies gekozen worden. Over het algemeen zijn SVMs relatief ongevoelig voor overfitting, waardoor de modelresultaten goed generaliseerbaar zijn (Auria & Moro, 2008). SVMs hebben toepassing gevonden op allerlei gebieden waar natuurlijke patronen en metingen vertaald moeten worden naar vooraf gedefinieerde klassen. SVMs kunnen naast classificatie ook gebruikt worden voor regressie; de zogenaamde Support Vector Regression (SVR).

Toepassingen - Handschriftherkenning, spam-filters en objectherkenning op foto's. Lekdetectie in leidingen (Mashford, De Silva, Marney, & Burn, 2009), classificatie van landgebruiktypen op basis van luchtfoto's (Pal & Mather, 2005) en conditiemonitoring van machines (Widodo & Yang, 2007).

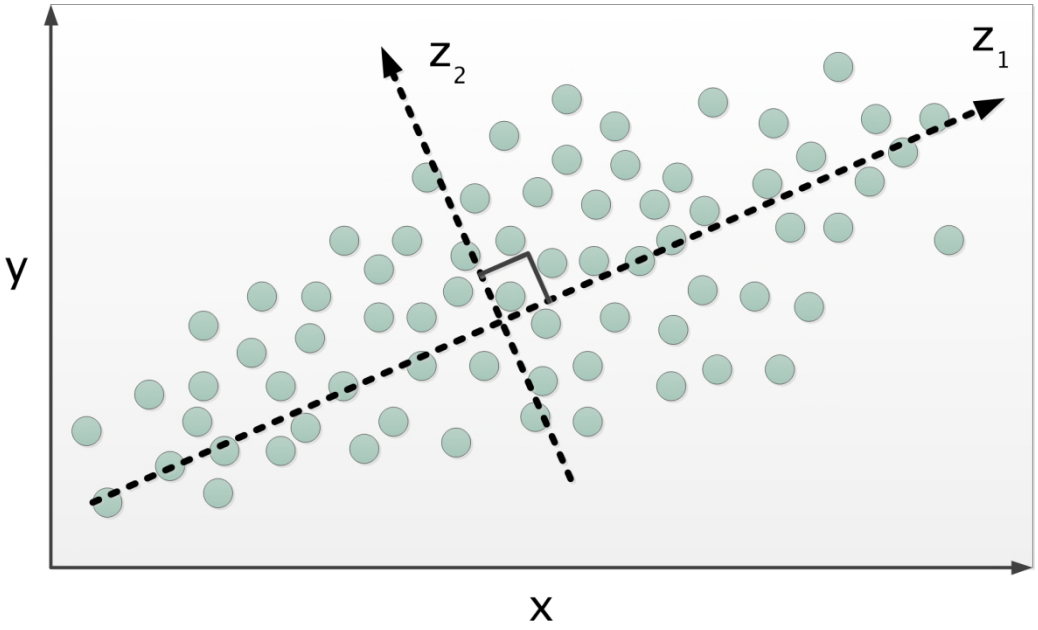
Principal component analysis

Principal Component Analysis (PCA) is een statistische analysemethode voor datasets opgebouwd uit meerdere variabelen (Hotelling, 1933; Pearson, 1901). De methode geeft inzicht in dataspreiding (variatie) van de verschillende variabelen, hun onderlinge afhankelijkheden en is toe te passen als datareductiemethode.

PCA werkt op basis van een wiskundige transformatie van een dataset naar een nieuwe set variabelen (coördinatenstelsel). In het nieuwe stelsel ligt de grootste dataspreiding langs de eerste coördinaat-as (de eerste hoofdcomponent, of 'first principal component'), de een-na-grootste spreiding langs de tweede as, enzovoort. De hoofdcomponenten zijn onderling onafhankelijk en opgebouwd uit een *combinatie* van de variabelen uit de originele (ongetransformeerde) variabelen. Om de hoofdcomponenten van de dataset te vinden wordt allereerst de onderlinge samenhang in de data beschreven met behulp van een covariantiematrix, waarna vervolgens de hoofdassen berekend worden. De spreiding van elke hoofdcomponent is hierbij een maat voor de informatierijkheid. Componenten worden tenslotte gerangschikt van grote naar lage dataspreiding (op basis van hun 'eigenwaarden').

Als een voldoende groot deel van de variatie in de data wordt verklaard door een beperkt aantal hoofdassen, kan gekozen worden voor datareductie door de overige assen achterwege te laten. Hoewel PCA uitgaat van lineaire verbanden tussen verschillende variabelen, zijn er ook niet-lineaire uitbreidingen ontwikkeld zoals 'additive principal component' en 'principal curves' analyse (Jolliffe, 2002). Mogelijke problemen met PCA kunnen zich voordoen met incomplete datasets: gaten in de data kunnen worden opgevuld met schattingen, maar bij

veel ontbrekende waarden is de analyse niet langer valide. Verder kan het bepalen van de covariantiematrix lang duren voor grote datasets.



Figuur 17: Een fictief voorbeeld van tentamencijfers van studenten als functie van college-aanwezigheid. Uit de samenhang blijkt dat als iemands aanwezigheid bekend is, er een ruwe schatting is te maken van het tentamencijfer (of vice versa). Met enig verlies aan nauwkeurigheid is de data daarom ook te beschrijven met een enkele variabele: een nieuwe maat die een combinatie is van aanwezigheid en cijfer. Deze wordt beschreven door hoofdas 1 ('Z1'). De tweede hoofdas ('Z2') beschrijft dan de overgebleven variatie, namelijk hoe groot de spreiding is in cijfer en aanwezigheid binnen een groep personen met dezelfde maat. Z1 en Z2 zijn de zogenaande hoofdcomponenten van deze dataset.

Voor datamining doeleinden is PCA geschikt vanwege de datareductie: het aantal dimensies van een dataset kan verkleind worden, waarbij die dimensies worden weggelaten die de minste onafhankelijke metingen bevatten. Belangrijk is daarbij wel dat er tevens nieuwe dimensies geïntroduceerd worden door PCA. Dit kan lastig zijn bij de interpretatie van gegevens. Er kan hierbij een grenswaarde worden aangegeven, zoals bijvoorbeeld de voorwaarde dat in ieder geval 95% van de variatie in de data nog beschreven wordt door de gereduceerde dataset.

Toepassingen - PCA is een veelgebruikte techniek, o.a. bij het comprimeren van digitale foto's of figuren, in de neurobiologie, maar is evengoed bruikbaar in de drinkwatersector. Voorbeelden zijn het bepalen van verbanden in waterkwaliteitsgegevens met veel ruis (Praus, 2005), detectie en karakterisatie van veranderingen in procesvariabelen (Ruiz, 2006), het reduceren van correlaties tussen sensor-signalen en het detecteren van afwijkingen in het zuiveringsproces (George, 2009).

Probabilistische modellen

Naive Bayes classifier

De naive Bayes classifier is, zoals de benaming ook suggereert, bedoeld voor classificatie. Het algoritme is gebaseerd op voorwaardelijke kansrekening (de zogenaamde Bayesiaanse statistiek). Als zodanig beschouwt een naive Bayes classifier elk van de attributen van een instantie als onafhankelijk van elkaar en berekent vervolgens op basis van het al dan niet hebben van bepaalde attributen de kans dat een instantie bij een bepaalde klasse behoort. Bijvoorbeeld: een voetbal wordt gekenmerkt door een ronde vorm, een kenmerkende grootte en een bepaald type zacht materiaal waarvan het gemaakt is. Men zou dan de kans kunnen berekenen dat iets een voetbal is, gegeven dat het een ronde vorm heeft, of gegeven dat het een ronde vorm heeft in combinatie met dat het van zacht materiaal gemaakt is. De naive Bayes classifier wijst de instantie dan op basis van deze kenmerken toe aan de categorie voetbal, of aan de categorie niet-voetbal, op basis van het meest waarschijnlijke. Het kwantificeren van de waarschijnlijkheid van elke classificatie is een belangrijk kenmerk van naive Bayes classifier.

Over het algemeen is een naive Bayes classifier in staat om met relatief weinig trainingsdata een nauwkeurige inschatting te maken van de modelparameters. Daarnaast werkt het algoritme snel, vooral ook bij grote datasets waarin instanties veel attributen hebben (Rish, 2001). De belangrijkste aanname achter de classifier is dat de attributen van elke instantie statistisch onafhankelijk van elkaar zijn. In de werkelijkheid is dit echter zelden waar. Toch blijkt verassend genoeg uit onderzoek dat deze onjuiste aanname geen invloed heeft op de kwaliteit van de modelvoorspellingen (Zhang, 2004).

Toepassingen - In de praktijk worden naive Bayes classifiers veel gebruikt voor het geautomatiseerd archiveren van documenten in voorgedefinieerde categorieën en voor het herkennen van spam bij e-mails. Binnen de watersector zijn zeer weinig toepassingen van dit algoritme bekend. Wel blijkt de naive Bayes classifier geschikt te zijn voor foutdiagnose van centrifugaalpompen. Hierbij wordt op basis van gemeten pompvibratie een nauwkeurige inschatting gemaakt van de conditie van pomponderdelen. Zo blijken bepaalde afwijkende vibratiepatronen veroorzaakt te worden door cavitatie, terwijl beschadigde lagers een ander karakteristiek vibratiepatroon kennen (Muralidharan & Sugumaran, 2012). Ook beschadigingen aan het schoepenwiel of combinaties van conditie-achteruitgang kunnen uit vibratiepatronen herkend worden.

Ensembles

Ensemble algoritmes zijn methodieken die meerdere machine learning methoden combineren om tezamen tot een voorspelling te komen. De meestgebruikte ensemble-methodiek is het zogenaamde boosting. Dit is een techniek waarbij niet een enkel algoritme wordt getraind, maar in plaats daarvan een serie algoritmes, die gezamenlijk voorspellingen doen. Bij boosting worden weegfactoren toegekend aan de trainingsinstanties. De weegfactoren worden gevarieerd zodat elk nieuw algoritme in het ensemble zich richt op de instanties die door zijn voorgangers verkeerd werden voorspeld (Domingos, 2012). Boosting wordt meestal toegepast bij beslisbomen. Elke boom uit de serie wordt getraind op de voorspellingsresiduen van de vorige boom. In plaats van een enkele (zeer complexe) beslisboom, wordt zo een hele set aan eenvoudige beslisbomen gevormd, die gezamenlijk

opereren. Veelgebruikte algoritmes uit deze categorie zijn Random Forests en Gradient Boosting Regression.