

Improving Quality of Life



Digitalisering Noordzee: 6 februari 2023



“**De gamechanger:** Microsoft sloot onlangs een miljardenalliantie met OpenAI, het bedrijf achter de slimme chatbot ChatGPT, zonder twijfel dé technologiehype van de voorbije maanden.” 6 februari 2023





AI in Watersector

Use cases voor AI in Water



monitoren van de
waterkwaliteit met sensoren



voorspellen van de
beschikbaarheid van
waterbronnen



machine learning en
beeldherkenning om dreigende
overstromingen te voorspellen
en te voorkomen.



analyseren van
weersverwachtingen en het
voorspellen van droogtes



optimaliseren van
irrigatiesystemen

Use cases voor Datascience in Water

1. Het monitoren van de waterkwaliteit met sensoren en het analyseren van de data om vervuiling te detecteren en te voorkomen.
2. Het voorspellen van de beschikbaarheid van waterbronnen en het bepalen van optimale distributie om aan de vraag te voldoen.
3. Het gebruik van machine learning en beeldherkenning om dreigende overstromingen te voorspellen en te voorkomen.
4. Het analyseren van weersverwachtingen en het voorspellen van droogtes om zo de waterbronnen beter te beheren en te gebruiken.
5. Het optimaliseren van irrigatiesystemen door middel van het analyseren van gegevens van bodemvochtigheid, neerslag, temperatuur en andere factoren om het watergebruik efficiënter en effectiever te maken.

Toepassing van AI in de use cases (1,2)

1. **Het monitoren van de waterkwaliteit met sensoren en het analyseren van de data om vervuiling te detecteren en te voorkomen:**

Voor deze usecase zouden sensoren gebruikt kunnen worden om waterkwaliteit te meten, zoals de pH-waarde, zuurstofgehalte, zoutgehalte, temperatuur en troebelheid. De verzamelde data zou vervolgens geanalyseerd kunnen worden met machine learning algoritmes om vervuiling te detecteren en de bron ervan te achterhalen. Het algoritme zou waarschijnlijk gebruik maken van patroonherkenningstechnieken om te leren wat de normale waarden zijn voor de verschillende parameters en afwijkingen te detecteren die wijzen op vervuiling.

Om deze data in te winnen zouden sensoren in het water geplaatst kunnen worden op strategische punten, zoals in de buurt van fabrieken of dichtbij rioleringen. De data kan vervolgens naar een centrale locatie worden gestuurd waar de algoritmes draaien om de data te analyseren.

Voor deze usecase zou het nodig zijn om grote hoeveelheden data te verzamelen om het algoritme effectief te laten zijn. Dit zou waarschijnlijk betekenen dat sensoren continu data moeten verzamelen om trends op te merken en om effectief te leren wat de normale waarden zijn. Het kan nodig zijn om gebruik te maken van geavanceerde statistische technieken om onderscheid te maken tussen normale schommelingen en verontrustende afwijkingen.

2. **Het voorspellen van de beschikbaarheid van waterbronnen en het bepalen van optimale distributie om aan de vraag te voldoen:**

Voor deze usecase zouden historische gegevens over de waterkwaliteit, het weer, verdamping, grondwaterstanden en waterverbruik gebruikt kunnen worden. Deze data zou ingewonnen kunnen worden via sensoren en weersatellieten, evenals openbare data uit meteorologische instituten en overheidsinstanties. Met behulp van machine learning algoritmes zou er een model kunnen worden gebouwd om toekomstige waterbeschikbaarheid te voorspellen en om optimale distributiemethoden te bepalen.

Toepassing van AI in de use cases (3,4,5)

3. **Het gebruik van machine learning en beeldherkenning om dreigende overstromingen te voorspellen en te voorkomen:**

Voor deze usecase zouden beelden van satellieten, drones en sensoren op strategische locaties gebruikt kunnen worden om gegevens te verzamelen over waterstromen, waterstanden en mogelijke overstromingsgebieden. Vervolgens zou machine learning algoritmes gebruikt kunnen worden om patronen te vinden in deze data en dreigende overstromingen te voorspellen. Op basis daarvan zouden er dan vroege waarschuwingen kunnen worden uitgegeven om de nodige maatregelen te nemen, zoals bijvoorbeeld het verhogen van dijken.

4. **Het gebruik van machine learning en beeldherkenning om dreigende overstromingen te voorspellen en te voorkomen:**

Voor deze usecase zouden historische weergegevens, getijdengegevens, gegevens over terrein- en waterveranderingen, en mogelijke oorzaken van overstromingen, zoals stijgende zeespiegels, gebruikt kunnen worden als trainingsdata voor het machine learning algoritme. Deze gegevens zouden verzameld kunnen worden via verschillende bronnen, zoals satellieten, sensoren, weerstations en andere meteorologische instanties.

5. **Het identificeren van potentiële lekkages in waterleidingen en het voorspellen van toekomstige lekkages:**

Voor deze usecase zou er data nodig zijn van de waterleidingen, zoals de leeftijd, het materiaal, de diameter, de ligging en historische gegevens over lekkages. Deze data zou vervolgens geanalyseerd kunnen worden met behulp van machine learning algoritmen om zwakke plekken en potentiële lekkages te identificeren. Deze gegevens zouden verzameld kunnen worden via sensoren die op de leidingen zijn geïnstalleerd en door middel van historische data die wordt bijgehouden door waterbedrijven.



Welk van de vijf use cases is het meest waardevol vanuit een economisch oogpunt

De tweede usecase, die gaat over het voorspellen van de beschikbaarheid van waterbronnen en het bepalen van optimale distributie om aan de vraag te voldoen, waarschijnlijk de meest waardevolle zijn vanuit economisch oogpunt.

Dit omdat het kan helpen bij het optimaliseren van de watervoorziening en de kosten van distributie en opslag van water kan verminderen.

Plan van aanpak opstellen voor een datascience team t.b.v. use case 2

1. Probleemdefinitie en scope afbakening: Het doel van dit project zou zijn om een algoritme te maken dat de beschikbaarheid van waterbronnen kan voorspellen en bepalen wat de optimale distributie is om aan de vraag te voldoen. Dit zou kunnen helpen om de watertoevoer te optimaliseren, wat uiteindelijk zou kunnen leiden tot lagere kosten en betere controle over de beschikbaarheid van water.
2. Dataverzameling en -analyse: Om een model te maken dat in staat is om de beschikbaarheid van waterbronnen te voorspellen, is het belangrijk om gegevens te verzamelen over de waterkwaliteit, het weer, de verdamping, de grondwaterstanden en het waterverbruik. Deze gegevens zouden afkomstig kunnen zijn van satellieten, sensoren, weerstations, meteorologische instituten en overheidsinstanties. De gegevens moeten worden geanalyseerd en opgeschoond en vervolgens gebruikt worden om het model te trainen.
3. Modelontwikkeling en -validatie: Het ontwikkelen van het daadwerkelijke model vereist een combinatie van statistische en machine learning technieken. Het model zou verschillende variabelen moeten omvatten, zoals historische weergegevens, waterverbruik en waterstanden. Het model zou vervolgens gevalideerd moeten worden met behulp van beschikbare gegevens. Een belangrijk onderdeel van de validatie is het evalueren van de voorspellende nauwkeurigheid van het model.
4. Implementatie en monitoring: Nadat het model is ontwikkeld en gevalideerd, kan het worden geïmplementeerd in de praktijk en worden gebruikt om de beschikbaarheid van waterbronnen te voorspellen en de optimale distributie te bepalen. Het is belangrijk om het model regelmatig te monitoren en de resultaten te evalueren om waar nodig aanpassingen te maken.
5. Rapportage en communicatie: Het is belangrijk om rapporten te genereren en regelmatig te communiceren met alle belanghebbenden, zoals de lokale overheid, om hen op de hoogte te houden van de voortgang van het project en de resultaten ervan.

Model opzetten voor deze use case

Het model zou gebaseerd kunnen zijn op machine learning algoritmen, zoals Time Series Forecasting. Het doel van dit model zou zijn om te voorspellen hoeveel water er beschikbaar zal zijn en wat de optimale distributie is om aan de vraag te voldoen. De data die gebruikt zou kunnen worden om dit model te trainen, zou historische gegevens over waterkwaliteit, het weer, verdamping, grondwaterstanden en waterverbruik zijn. Deze gegevens zouden verzameld kunnen worden via sensoren en weersatellieten, evenals openbare data uit meteorologische instituten en overheidsinstanties.

Mogelijke stappen in de ontwikkeling en implementatie van dit model zouden kunnen zijn:

1. Data cleaning en pre-processing: Er zou een grondige analyse van de data moeten worden gedaan. Dit zou kunnen inhouden dat ontbrekende of onjuiste gegevens worden geïdentificeerd en gecorrigeerd.
2. Feature engineering: Dit zou het creëren van nieuwe features uit de bestaande data kunnen omvatten, zoals seizoensgebonden trends, om betere voorspellingen te ondersteunen.
3. Model training en evaluatie: Het model zou getraind kunnen worden met historische gegevens en vervolgens geëvalueerd kunnen worden door het testen van de voorspellingen van het model aan de hand van nieuwe gegevens. Hierdoor kan worden bepaald hoe nauwkeurig het model is.
4. Implementatie en monitoring: Nadat het model is gevalideerd, kan het worden geïmplementeerd in de waterdistributiesystemen. Het model zou moeten worden gemonitord om te zorgen dat het nauwkeurig blijft.

NB. Dit is een hoog niveau voorbeeld van hoe een model zou kunnen worden ontworpen voor deze usecase. Het precieze modelontwerp en implementatie zou afhangen van de specifieke omstandigheden van de locatie waar het zal worden gebruikt.

Model voor use case gericht op Zwolle, Nederland

Voor het voorspellen van de beschikbaarheid van waterbronnen en het bepalen van optimale distributie om aan de vraag te voldoen in het gebied van Zwolle, zouden we historische gegevens kunnen gebruiken van het lokale waterverbruik, meteorologische gegevens, grondwaterstanden, waterkwaliteit en gegevens over het waterdistributienetwerk.

De data zou vooraf in verschillende formaten en bronnen bestaan, en moet worden geuniformeerd en gestructureerd voordat deze kan worden gebruikt om het model te trainen.

Voor de doelstelling van ons model, zou het tijdreeksanalyse gebruiken om historische patronen te identificeren en toekomstige veranderingen te voorspellen. Het model zou rekening houden met seizoensgebonden patronen, trends en afwijkingen om te bepalen hoeveel water er beschikbaar zal zijn.

Om tot de optimale waterdistributie te komen, zou het model ook rekening houden met de afstand tussen de waterbronnen en de verspreiding van de vraag naar water over verschillende gebieden. Het zou zich richten op verschillende factoren verband houdend met de waterverdeling zoals transporttijd, kosten, en toevoer van de waterdistributienetwerk.

Naarmate het model wordt getraind, zal het geleidelijk aan accurater worden. Het model zou vervolgens kunnen worden geïntegreerd in een dashboard of systeem dat automatisch aanbevelingen doet voor optimale waterdistributie en voorraadbeheer in de regio Zwolle.

Detaillering 1

Voor historische gegevens van het lokale waterverbruik zou het team kunnen samenwerken met het waterbedrijf in het gebied om gegevens te verkrijgen over het totale waterverbruik per dag, week of maand in de voorgaande jaren.

Voor de meteorologische gegevens, zoals neerslag en verdamping, zou het team gegevens kunnen verzamelen van meteorologische instituten en overheidsinstanties die actief zijn in het gebied van Zwolle.

Voor grondwaterstanden zou het team sensoren kunnen plaatsen in verschillende gebieden in de stad om gegevens op te halen en deze in realtime te kunnen incorporeren in het model.

Voor de waterkwaliteit zou het team kunnen samenwerken met de waterbeheerders in het gebied om gegevens te verkrijgen over de waterkwaliteit van het oppervlaktewater en het grondwater.

Ten slotte zou het team gegevens over het waterdistributienetwerk kunnen verkrijgen van het waterbedrijf om te begrijpen hoe het water door de stad wordt gedistribueerd en welke gebieden meer water nodig hebben dan andere.

Al deze gegevens zouden vervolgens kunnen worden ingevoerd in het model om de beschikbaarheid van waterbronnen te voorspellen en de optimale distributie te bepalen om aan de vraag te voldoen.

Detailering 2

Voor de grondwaterstanden zou het team kunnen samenwerken met het lokale waterschap, dat metingen verricht bij verschillende peilbuizen in de buurt van Zwolle.

Voor de kwaliteit van het water zou het team gegevens kunnen verzamelen bij het waterbedrijf en waterschap over de chemische en bacteriologische samenstelling van het water, zoals de concentratie van zouten, zuurstof, nitraat- en fosfaatverbindingen, en pathogene micro-organismen.

Ten slotte zou het team data over het distributienetwerk kunnen verzamelen bij het waterbedrijf, zoals informatie over de capaciteit en prestaties van de waterpompen, leidingen en reservoirs. Deze gegevens zouden kunnen worden gebruikt om de optimale distributie van water te bepalen en om inzicht te krijgen in eventuele bottlenecks in het distributienetwerk.

Detaillering 3

Het datateam zou gegevens kunnen verzamelen over de fysieke eigenschappen van het waterdistributienetwerk, zoals de diameter en lengte van de buizen, de locatie van pompen en kleppen, en de hoogteverschillen tussen verschillende delen van het netwerk. Deze gegevens zouden kunnen worden gebruikt om de stroomsnelheid van het water in het netwerk te berekenen en om te bepalen welke delen van het netwerk kwetsbaar zijn bij piekbelastingen.

Met al deze gegevens zou het datateam kunnen beginnen met het in kaart brengen van de correlaties tussen verschillende variabelen en het ontwikkelen van een model dat de beschikbaarheid van waterbronnen en de optimale distributie kan voorspellen. Het zou ook belangrijk zijn om het model te testen met historische gegevens en te verfijnen op basis van de resultaten van deze tests.

Detaillering 4

Als we nog dieper ingaan op de details, zou het datateam ook informatie moeten verzamelen over de watervoorziening in het gebied van Zwolle, zoals de locatie van waterbronnen en waterzuiveringsinstallaties.

Om de beschikbaarheid van waterbronnen te voorspellen, zouden ze historische gegevens kunnen gebruiken over de hoeveelheid water die uit deze bronnen is gewonnen en de watervorraden in de omgeving. Deze gegevens zouden in kaart gebracht kunnen worden met behulp van geografische informatie systemen (GIS), die ook kunnen helpen om de afstand te bepalen tussen de waterbronnen en de vraaglocaties.

Tot slot zou het datateam ook rekening moeten houden met specifieke eigenschappen van het Zwolse gebied, zoals de geografie, demografie en economie. Deze factoren zouden kunnen worden gebruikt om de vraag naar water te voorspellen, bijvoorbeeld door te kijken naar de bevolkingsgroei, de ontwikkeling van nieuwe bedrijven en de invloed van seizoensgebonden evenementen.

Logisch datamodel

- De kernentiteit zou "Waternverbruik" zijn, die gerelateerd is aan andere kernentiteiten zoals "Weergegevens", "Grondwaterstanden", "Waterkwaliteit", "Waternvoorziening" en "Waternistributienetwerk".
- "Weergegevens" zou gerelateerd zijn aan sub-entiteiten zoals "Neerslag", "Temperatuur" en "Luchtvochtigheid".
- "Grondwaterstanden" zou gerelateerd zijn aan sub-entiteiten zoals "Stijghoogte", "Grondwaterdata" en "Vergelijkingsdata".
- "Waterkwaliteit" zou gerelateerd zijn aan sub-entiteiten zoals "Chemische samenstelling", "Bacteriologische samenstelling", "Meetgegevens", "Normwaarden" en "Kwaliteitscontrole".
- "Waternvoorziening" zou gerelateerd zijn aan sub-entiteiten zoals "Waternbronnen" en "Waternzuiveringsinstallaties".
- "Waternistributienetwerk" zou gerelateerd zijn aan sub-entiteiten zoals "Leidingen", "Kleppen", "Pompen", "Netwerkstructuur", "Fysieke eigenschappen" en "Simulatiemodellen".

Detailering:

- - "Neerslag" zou gerelateerd zijn aan sub-entiteiten zoals "Regendata" en "Gemiddelde neerslag". "Regendata" zou gerelateerd zijn aan sub-entiteiten zoals "Datum", "Tijd" en "Hoeveelheid regen" terwijl "Gemiddelde neerslag" gerelateerd zou zijn aan sub-entiteiten zoals "Periode", "Locatie" en "Gemiddelde hoeveelheid regen".
- - "Stijghoogte" zou gerelateerd zijn aan sub-entiteiten zoals "Locatie", "Afstand tot oppervlaktewater", "Datum en tijd" en "Stijghoogtedata".
- - "Waternistributienetwerk" zou gerelateerd zijn aan sub-entiteiten zoals "Leidingen", "Kleppen", "Pompen", "Locaties", "Diameters" en "Hoogteverschillen".

Dit datamodel zou de koppelvlakken voor alle benodigde datasets kunnen definiëren en zou kunnen helpen bij het structureren van het verzamelde gegevens om het model te trainen en de voorspellingen te doen.

Volledig datamodel

- Kernentiteit: Waterverbruik
- Gerelateerde entiteiten:
 - Weergegevens
 - Grondwaterstanden
 - Waterkwaliteit
 - Watervoorziening
 - Waterdistributienetwerk

- Entiteit: Weergegevens
- Gerelateerde entiteiten:
 - Neerslag
 - Temperatuur
 - Luchtvochtigheid

- Entiteit: Neerslag
- Gerelateerde entiteiten:
 - Regendata
 - Gemiddelde neerslag

- Entiteit: Regendata
- Gerelateerde entiteiten:
 - Datum
 - Tijd
 - Hoeveelheid regen

- Entiteit: Gemiddelde neerslag
- Gerelateerde entiteiten:
 - Periode
 - Locatie
 - Gemiddelde hoeveelheid regen

- Entiteit: Grondwaterstanden
- Gerelateerde entiteiten:
 - Stijghoogte
 - Grondwaterdata
 - Vergelijkingsdata

- Entiteit: Stijghoogte
- Gerelateerde entiteiten:
 - Locatie
 - Afstand tot oppervlaktewater
 - Datum en tijd
 - Stijghoogtedata

- Entiteit: Waterkwaliteit
- Gerelateerde entiteiten:
 - Chemische samenstelling van het water
 - Bacteriologische samenstelling van het water

- Entiteit: Chemische samenstelling van het water
- Gerelateerde entiteiten:
 - Soort chemische samenstelling
 - Tijdsperiode waarin de data is verzameld
 - Locatie

- Entiteit: Bacteriologische samenstelling van het water
- Gerelateerde entiteiten:
 - Soort bacteriologische samenstelling
 - Tijdsperiode waarin de data is verzameld
 - Locatie

- Entiteit: Watervoorziening
- Gerelateerde entiteiten:
 - Waterbronnen
 - Waterzuiveringsinstallaties

- Entiteit: Waterbronnen
- Gerelateerde entiteiten:
 - Hoeveelheid water die is gewonnen
 - Watervorraden in de omgeving

- Entiteit: Waterzuiveringsinstallaties
- Gerelateerde entiteiten:
 - Capaciteit van de zuiveringsinstallatie
 - Effluentkwaliteit

- Entiteit: Waterdistributienetwerk
- Gerelateerde entiteiten:
 - Fysieke eigenschappen van het netwerk (diameter, lengte, hoogteverschillen, pompen, kleppen)
 - Stroomsnelheid van water in het netwerk
 - Kwetsbare delen van het netwerk bij piekbelastingen.

Script schrijven voor Time Series Forecasting dat aansluit bij de 2e use case en gebaseerd is op het door ons opgestelde datamodel

Je hebt gelijk dat het datamodel meerdere gerelateerde entiteiten bevat. Als we machine learning-algoritmen zoals Time Series Forecasting willen toepassen om de beschikbaarheid van drinkwater te voorspellen, zouden we inderdaad meerdere gegevensbronnen moeten gebruiken, zoals weergegevens en grondwaterstanden.

Hier is een aangepast script dat rekening houdt met meerdere entiteiten:

```
...
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
import matplotlib.pyplot as plt
%matplotlib inline

# Load and preprocess data
water_data = pd.read_csv('water_data.csv', parse_dates=['date'])
weather_data = pd.read_csv('weather_data.csv', parse_dates=['date'])
groundwater_data = pd.read_csv('groundwater_data.csv', parse_dates=['date'])

# Merge data
data = pd.merge(water_data, weather_data, on='date')
data = pd.merge(data, groundwater_data, on=['location', 'date'])

# Select features and target
X = data[['temperature', 'precipitation', 'groundwater_level']]
y = data['water_demand']

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
# Fit model
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```

```
# Evaluate model
r2 = r2_score(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
```

```
print('R2 score:', r2)
print('Mean squared error:', mse)
print('Mean absolute error:', mae)
```

```
# Visualize results
plt.scatter(y_test, y_pred, alpha=0.5)
plt.plot(y_test, y_test, color='black')
plt.xlabel('Actual Water Demand')
plt.ylabel('Predicted Water Demand')
plt.title('Water Demand Prediction')
plt.show()
...
```

Dit script laadt gegevens over watervraag, weergegevens en grondwaterstanden in en voegt ze samen tot één dataset. Vervolgens worden de kolommen met de temperatuur, neerslag en grondwaterstanden geselecteerd als functies en de watervraag als doelvariabele. Het script past een Random Forest-regressiemodel toe op de features en traint het op een trainingset voordat het de prestaties op een testset evalueert en de resultaten visualiseert met een scatterplot.