Algoritme

Korte beschrijving van het algoritme

Naam *De naam die gebruikt wordt om het algoritme aan te duiden.*	Verbeteren van sensordata luchtkwaliteit door een model
Organisatie *De volledige naam van de organisatie waar het algoritme ingezet wordt.*	DataFryslân
Korte omschrijving *Een korte beschrijving van het algoritme.*	Het gebruik van sensoren om luchtmetingen te doen, heeft brede maatschappelijke gevolgen. Het heeft invloed op de gezondheid van mensen, de industrie en op natuurgebieden (biodiversiteit). Om deze complexe problematiek adequaat aan te pakken, is het van essentieel belang om gebruik te maken van modelmatige berekeningen die bouwen op metingen ter plekke. Daarom onderzoeken we in dit /LAB de ontwikkeling en toepassing van betaalbare, maar betrouwbare luchtmetingsensoren. We zoomen in sensoren die gebruikt zijn binnen het project ‘Zicht op Stikstof’. Deze sensoren zijn gebruikt voor onder andere stikstofdioxide (NO2) en fijnstof (PM10). In dit /LAB hebben we gekeken naar NO2 (luchtverontreiniging afkomstig van het verkeer en industrie). Hoewel deze sensoren een grotere foutmarge hebben dan de duurdere meetstations van het RIVM, hebben ze waardevolle data opgeleverd. Hoewel we over een aanzienlijke hoeveelheid gegevens beschikken dankzij deze sensoren, is er nog geen modelmatige berekening gedaan om achtergrond ruis en data-vervuiling eruit te halen ten behoeve van gekalibreerde NO2-metingen. Hier komt de toepassing van kunstmatige intelligentie (AI) en data science om de hoek kijken. Door deze technieken toe te passen op de verzamelde NO2-meetgegevens van de ontwikkelde sensoren, kunnen we schonere meetdata verkrijgen. Hiervoor zijn gegevens van het RIVM gebruikt voor kalibratiedoeleinden. Het ontwikkelde model is

	gebaseerd op de verzamelde gegevens van de NO2-sensoren in Utrecht, evenals de meetdata van diverse RIVM-meetstations. De resultaten van dit /LAB laten zien dat betaalbare sensoren kunnen worden ingezet om een gedetailleerder beeld te krijgen van de lokale luchtkwaliteit. Het ontwikkelde model, gebaseerd op de gegevens uit Utrecht, heeft aangetoond dat het combineren van AI met goedkope sensorwaarden meerwaarde heeft om inzicht te krijgen in lokale metingen.
Type algoritme *Is het algoritme zelflerend? In een niet-zelflerend algoritme specificeert de mens de regels die de computer moet volgen. Als het een zelflerend algoritme is, leert de machine over de patronen in de data.*	Supervised model
Methoden en modellen ***Standaard methoden of modellen die het algoritme gebruikt.***	Random Forest in Python
Beleidsterrein ***Trefwoorden over het beleidsterrein waarin het algoritme wordt ingezet.***	Ruimtelijk
Status ***De status van het algoritme: in ontwikkeling, in gebruik, of buiten gebruik.***	Alleen gebruikt tijdens de analyse van de data.
Doel ***Het doel waarvoor het algoritme ontwikkeld is en/of hoe de inzet ervan bijdraagt aan het behalen van die doelen.***	Afbakening: • Het te ontwikkelen model gaat wordt gebaseerd op de verzamelde data van de NO2 sensoren in Utrecht en RIVM meetstation data. Gewenst resultaat:

	• Een model voor het corrigeren van de data van de NO2 sensoren om schonere en gekalibreerde data te verkrijgen • Inzicht in de kwaliteit van NO2 sensoren en in hoeverre de sensoren de ground truth goed voorspellen Doel: • Schone en gekalibreerde data van low-quality sensoren die NO2 meten.
Impact ***De impact van het algoritme op burgers en bedrijven. Bijvoorbeeld: hoe werkt het algoritme en wat zijn de verwachte consequenties daarvan voor het individu of bedrijf?***	Het betreft een algoritme dat voorspelde waarden geeft dat niet van invloed is op personen. De impact van het algoritme is met name relevant voor het politieke domein, waar luchtkwaliteit een steeds belangrijke speler is en men vaker kritiek heeft om simulatie-uitkomsten (en niet lokaal meten). De consequenties gelden vooral voor de uitkomsten van het algoritme, niet zozeer voor het algoritme zelf. De keuze voor het gebruikte model om de sensor-data te analyseren komt voor uit de extensieve literatuur over dit soort data en hoe die het beste zijn te kalibreren.
Proportionaliteit ***Een afweging van de voor- en nadelen van de inzet van het algoritme en waarom dit redelijk gerechtvaardigd is.***	Dit soort big data zijn lastig met niet-machine learning technieken te kalibreren. Daarnaast is er post-hoc gekalibreerd en niet real-time. Het algoritme en de uitkomsten daarvan zijn niet ingezet voor beleid. Het algoritme is een middel om antwoord te geven op de vraag of er met AI schonere en gekalibreerde data opgeleverd kan worden. Het gebruik van het beste geïdentificeerde model lijkt derhalve gerechtvaardigd en ook common practice.
Menselijke tussenkomst ***Een omschrijving van hoe uitkomsten van het algoritme door een mens gecontroleerd en bijgesteld (kunnen) worden.***	De input voor het model is bepaald door de onderzoekers. De keuze voor de onafhankelijke variabelen in het model zijn wel gebaseerd op de literatuur. De mens geeft sturing aan de input en valideert de output aan de hand van de valideer dataset.
Monitoring ***Een overzicht van hoe de inzet van het algoritme wordt gemonitord.***	De geautoriseerde onderzoekers van DataFryslân hebben zicht op de uitkomsten, het script en wat ermee gedaan wordt. De resultaten zijn gepresenteerd en daarbij is de literatuur gevolgd in het gebruik van modellen om sensor-data te kalibreren.
Risico	Het risiconiveau van de inzet van dit algoritme zijn beoordeeld als 'laag' omdat er is geprobeerd om gevalideerde data van het RIVM te voorspellen met

Een overzicht van de voorziene risico's bij de inzet van het algoritme.	variabelen die vooral ruimtelijk van aard zijn. Het betreft geen data over personen.
Prestatienormen ***Een omschrijving van de verwachte prestaties van het algoritme en hoe die worden gemeten.***	Om het model te trainen en te valideren is de data opgesplitst: • 70% van de set wordt gebruikt om te trainen • 30% om het model te valideren • De splitsing is willekeurig gedaan, om bias tegen te gaan
Databronnen ***Een overzicht van de databronnen die op dit moment gebruikt worden door het algoritme en/of in het begin gebruikt zijn bij het maken van het algoritme.***	Het betreft data van het RIVM (open data) omtrent de 5 meetpalen in Utrecht van 2020 t/m 2023. Daarnaast is data ontvangen van N = 3.150.742 metingen met sensoren. De data is vergelijk gecleaned (verwijdering sensoren met te weinig observaties, errorwaarden en outliers. Vervolgens is de data genormaliseerd zodat sensoren onderling vergelijkbaar zijn. Deze data is vervolgens geaggregeerd naar uur-niveau om te kunnen vergelijken met de RIVM-data. In de data zitten gegevens over luchtvochtigheid (RH), datums (DT), stikstofdioxide (NO2) en Ozon (O3).

Heeft u vragen? U kunt altijd mailen naar info@datafryslan.nl over de gebruikte AI.