Modellering Experimenten V2

Digital Twin 3.0

Handpicked Agencies

Breda

Door: Koen Pijnenburg

Introductie

Naar aanleiding van de eerste iteratie van de modellering experimenten zullen er aanpassingen gemaakt worden aan de modellen. Onderstaand is de conclusie van de eerste iteratie te vinden.

Het doel was dat de machine learning modellen tenminste een R2 score behalen van 0,9. Momenteel behaalt geen enkel model dit doel, de ARIMA modellen kwamen hier het dichtst bij in de buurt.

Deze modellen zijn toegepast om de Twindle applicatie uit te breiden. De voorspellingen van deze modellen lopen altijd 1 uur achter. Dit is een eigenschap van de modellen waardoor ze niet toepasbaar zijn voor dit probleem.

Daarnaast kan er maximaal een uur in de toekomst gekeken worden. De impact die dit heeft op de applicatie is te klein.

Om dit doel te behalen en de implementatie te verbeteren kunnen de aanpassingen gemaakt worden:

1. Onderzoeken wat invloed heeft op veranderingen in de meetwaarden en deze data toevoegen aan de modellen.
2. Verder in de toekomst kijken, 3 uur of meer.
3. Betere controleren voor welk tijdstip de voorspelling gemaakt worden.
4. Bestandsgrootte modellen verminderen om schaalbaarheid te vergroten.

In dit document zullen voor deze uitdagingen oplossingen worden gezocht. Daarnaast is in overleg met de opdrachtgevers en belangrijkste stakeholder, Techtenna, besloten om eerste een model voor de CO2 waarde te ontwikkelen en deployen. Wanneer dit succesvol is bevonden kan dit worden uitgebreid naar de overige meetwaarden.

Inhoudsopgave

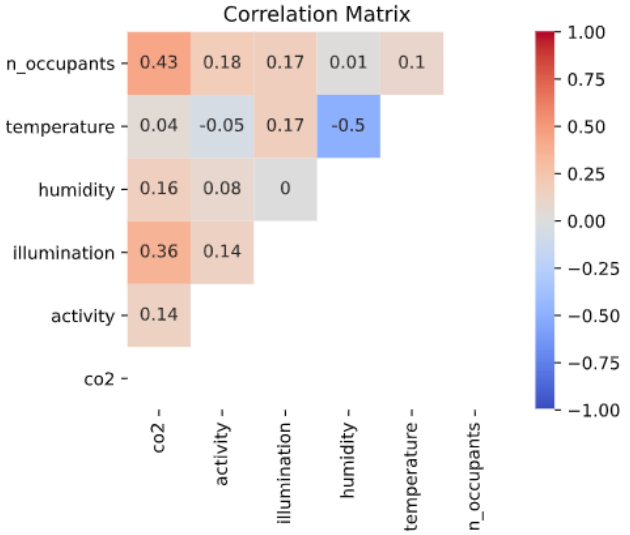
Gegevensoverzicht

Tijdens de tweede oplevering van het project werd door Marco van Techtenna opgemerkt dat het aantal personen wat zich in een ruimte bevindt invloed heeft op de CO2 niveaus. Via de Google Calendar API is opgevraagd hoeveel personen zich op een bepaalde tijd in een ruimte bevinden. Deze gegevens zijn toegevoegd aan de dataset zoals beschreven in de vorige iteratie.

Uit een vergelijkbaar onderzoek (Kallio et al., 2021) is gebleken dat temperatuur, luchtvochtigheid en activiteitsniveau een mogelijke invloed kunnen hebben op de CO2 waarde.

Correlaties

In de onderstaande heatmap valt een overzicht te zien van de correlaties die zich bevinden in de gegevens. De eerste kolom geeft aan dat de meeste meetwaarde een positieve invloed hebben op de CO2 waarden.



Afbeelding 1: Correlation Heatmap

Bevindingen

De CO2 waarden worden inderdaad beïnvloed door het aantal personen in de ruimte, activiteitsniveau (activity & illumination), luchtvochtigheid en temperatuur. Het zijn echter gemiddeld (< 0.5) tot zwakke (< 0.25) correlaties wat betekent dat deze gegevens op zich waarschijnlijk niet genoeg zijn om accurate voorspellingen te kunnen maken.

Eisen

In dit hoofdstuk zal worden toegelicht aan welke eisen de modellen moeten voldoen. Dit zal gedaan worden door de evaluatie methoden te beschrijven en doelen te stellen waar de modellen aan moeten voldoen.

Doelen

Vanuit de product owner is het doel gesteld dat de modellen tenminste 90% accuraat moeten zijn. De meetwaarden die gemodelleerd dienen te worden zijn continu en niet uit te drukken in procent accuracy. R2 Score kan gebruikt worden voor dit soort modellen.

Het geeft een nummer tussen 0.0 en 1.0 aan wat geïnterpreteerd kan worden als een percentage. Wanneer een R2 Score van 0.9 wordt behaald zal het model aan de eisen van de product owner voldoen.

Evaluatiemethoden

Om te kunnen bepalen hoe goed een model presteert zullen meerdere evaluatiemethoden worden toegepast. Deze kunnen onderverdeeld worden in datavoorbereiding en scoring methodieken.

Scoring
De modellen zullen beoordeeld worden op twee scores; R2 Score, Root Mean Square Error (RMSE).

R2 Score
R2 is een score voor lineaire regressiemodellen. Deze waarde geeft het percentage van de variantie in de afhankelijke variabele aan dat de onafhankelijke variabelen gezamenlijk verklaren. R2 meet de sterkte van de relatie tussen uw model en de afhankelijke variabele op een handige schaal van 0,0 - 1,0.

RMSE
Dit is een cijfer dat in dezelfde eenheid is als het target. Geeft de gemiddelde foutmarge weer. Wordt gebruikt om duidelijk te maken in hoeverre de gemiddelde voorspelling kan afwijken.

Model Beschrijvingen

In dit hoofdstuk wordt een overzicht van de modellen gegeven. Dit wordt gedaan door de voorbereiding van de data toe te lichten en de modelkeuze uit te leggen. Er zullen twee modellen uitgewerkt worden. Een voor lange termijn voorspellingen en een voor korte termijn voorspellingen.

Voorbereiding - Korte termijn

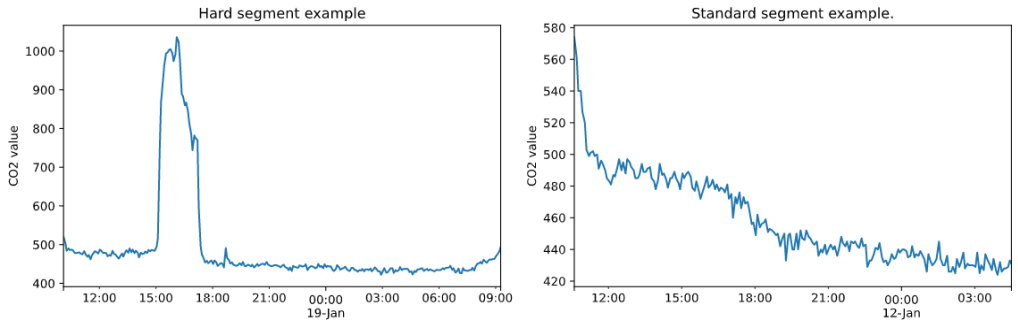
In het onderzoek van Kallio et al. wordt in hoofdstuk 3.2 het data voorbereidings process toegelicht. Deze voorbereiding zal toegepast worden voor de data die verzameld is voor dit project. Het bestaat uit de volgende stappen:

1. Data uit meerdere ruimtes samenvoegen.
2. Aggregatie en resampling; het opvangen van oneven gemeten en missende data.
3. Outlier verwijdering; Iedere meetwaarde heeft minimaal en maximaal acceptabele waarden, zie tabel 1. Alles hierbuiten wordt verwijderd.

Meetwaarde	**Minimaal**	**Maximaal**	**Eenheid**
CO2	350	5000	parts per million (PPM)
Temperature	0	40	Celsius (°C)
Humidity	0	80	%
Activity	0	12	PIR activity
Pressure	950	1100	Hectopascal (hPa)

Tabel 1: Minimale en maximale waarden

Na dit proces uit te voeren zijn er 23.866 datapunten beschikbaar. Deze worden opgedeelt in een train- en test set. Hiervoor wordt een methode gebruikt die ‘moeilijk’ te voorspellen data prioriteert als testdata. Data wordt geclassificeerd als ‘moeilijk’ wanneer er een grote pieken of dalen in de metingen zitten. Onderstaand is een afbeelding te zien wat het verschil tussen een moeilijk en standaard segment is.



Afbeelding 2: Moeilijk- en standaard segment voorbeelden

Voorbereiding - Lange termijn

In het onderzoek van Kallio et al. wordt verwezen naar een onderzoek waarin een model wordt ontwikkeld om op lange termijn de CO2 te voorspellen. In dit onderzoek van Putra et al. wordt de data opgesplitst in dag en uur. Dit process process is gekoppeld met de outlier verwijdering van vorig hoofdstuk om tot de volgende dataset te komen:

```
co2          float64 # Target
boardroom    float64
friday         uint8
monday         uint8
saturday       uint8
sunday         uint8
thursday       uint8
tuesday        uint8
wednesday      uint8
0              uint8 # Hours (0 - 24)
1              uint8
2              uint8
3              uint8
4              uint8
5              uint8
6              uint8
7              uint8
8              uint8
9              uint8
10             uint8
11             uint8
12             uint8
13             uint8
14             uint8
15             uint8
16             uint8
17             uint8
18             uint8
19             uint8
20             uint8
21             uint8
22             uint8
23             uint8
```

Korte termijn modellen

Tijdens het onderzoek van Kallio et al worden meerdere mogelijke modellen beschreven. De onderstaande modellen zullen worden toegepast. Deze zijn specifiek uitgekozen omdat ze accurate voorspellingen kunnen geven op auto gecorreleerde data.

- Ridge Regression
- Decision Tree
- Random Forest
- Multilayer Perceptron

Aangezien de CO2 waarden in een ruimte vaak niet drastisch veranderen in korte tijd is er gekozen om als baseline een ‘last-observation carried forward’ model te gebruiken. Dit geeft een de laatst gemeten CO2 waarde aan als de voorspelde waarde.

Lange termijn modellen

In het onderzoek van Putra et al. wordt alleen vermeld dat zij een neural network met de Levenberg-Marquardt optimizer gebruiken. Het aantal hidden layers en nodes per layer is wordt niet genoemd. Met de volgende modellen is geëxperimenteerd.

- Multilayer perceptron
- Keras neural network

Net zoals bij het korte termijn model zal er hier gebruik gemaakt worden van een baseline model. De gemiddelde waarden op een dag en tijdstip variëren niet veel. Deze kunnen dus gebruikt worden als baseline. Onderstaand is hier een voorbeeld van te zien.

```
day     hour
Friday  0       438.764583
        1       436.865625
        2       436.252381
        3       436.031994
        4       437.247917
...
Name: co2, dtype: float64
```

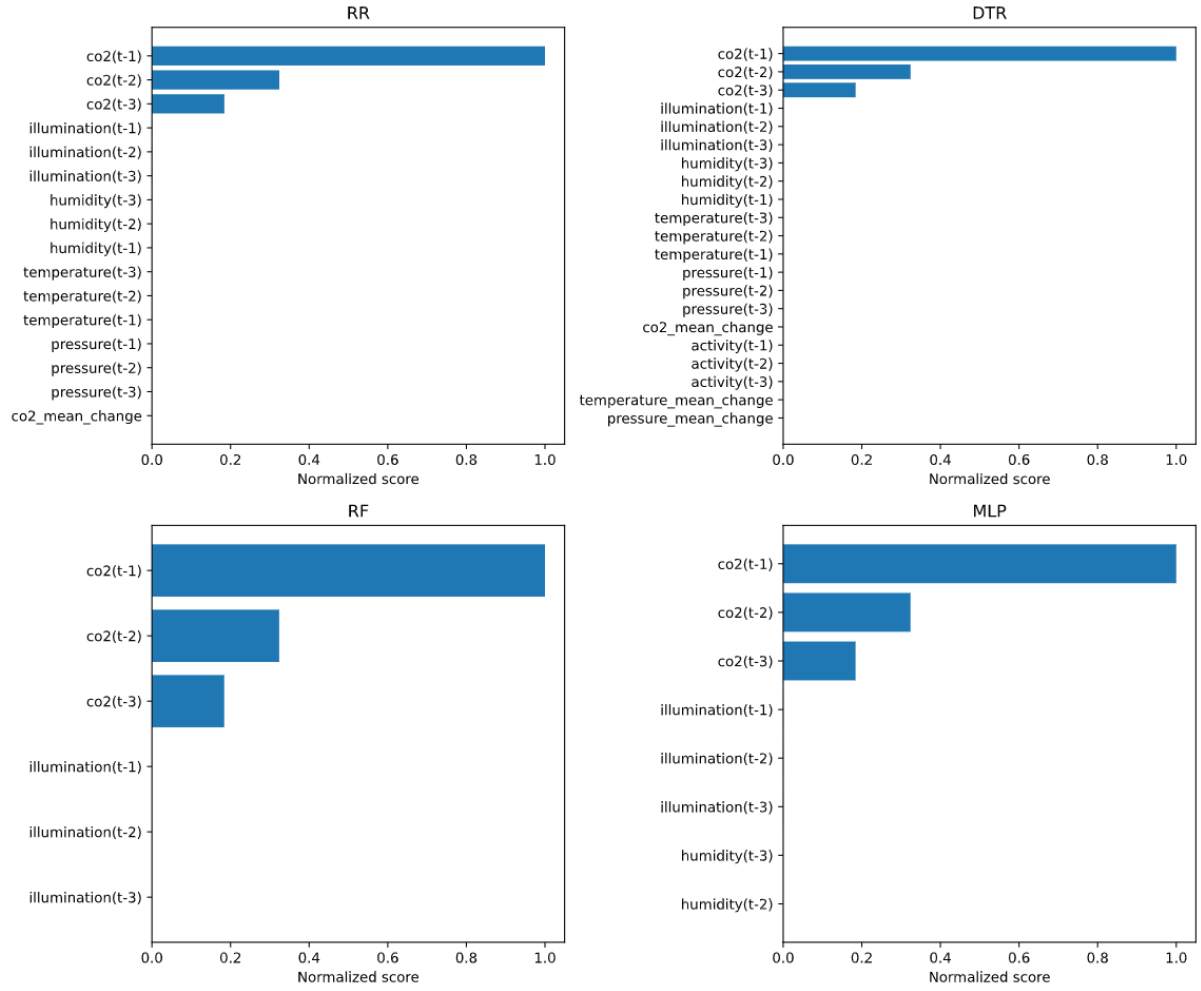
Model Beoordelingen

De in het voorgaande hoofdstuk genoemde modellen zijn ontwikkeld en geëvalueerd. In dit hoofdstuk zal worden toegelicht wat de resultaten waren en wat er eventueel verbeterd kan worden in komende iteraties.

Korte termijn

Targets & features

Per model kan het verschillen welke features tot het beste resultaat leiden. Om dit te optimaliseren is er gebruik gemaakt van de `SelectKBest` class. Deze klasse maakt gebruik van `f_regression` om te bepalen welke features het meest gecorreleerd zijn tot het target. Tussen de drie en dertig features zijn getest door middel van `GridSearch`. Onderstaand valt per model te zien welke features de hoogste score behalen.

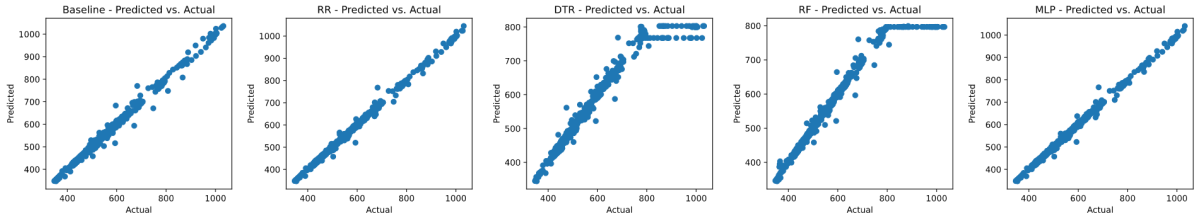


Afbeelding 3: Feature selectie

Bij ieder model zijn de drie voorgaande CO2 metingen de belangrijkste features. De features daarna zijn minder gecorreleerd maar dragen wel informatie bij aan het model.

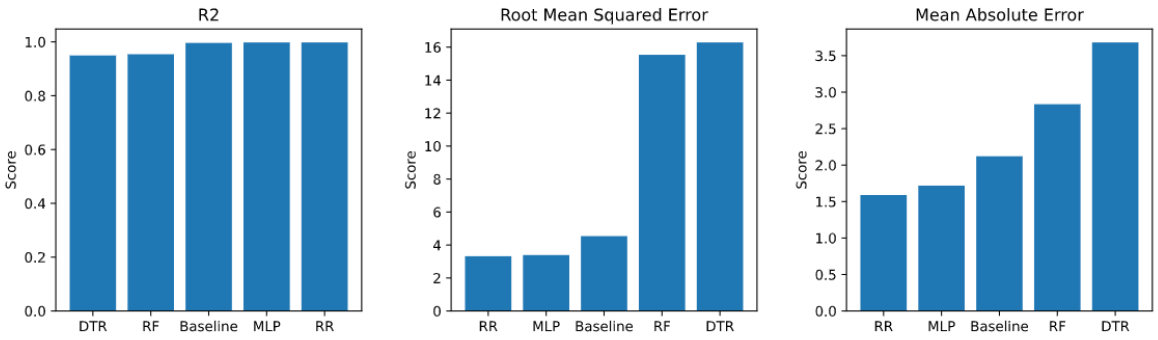
Resultaten

In de onderstaande afbeeldingen worden op verschillende manieren de modellen geëvalueerd. Per afbeelding zal een korte toelichting gegeven worden.



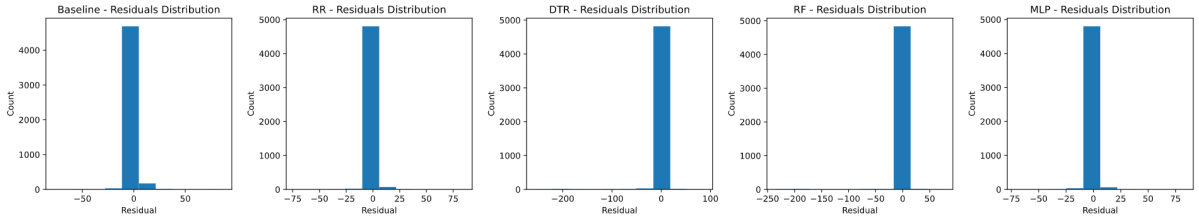
Afbeelding 4: Daadwerkelijk en voorspelt per model

Door de voorspelde waarde tegenover daadwerkelijke waarde te plotten kunnen patronen in de fouten van model worden opgespoord. Ideaal gezien vormen deze plotten een diagonale lijn, zoals bij het baseline, RR en MLP model. Dit betekend dat de voorspelde waarde dicht bij de daadwerkelijke waarde ligt. Bij de DTR en RF modellen worden de voorspellingen onnauwkeurig wanneer ze hoger worden.



Afbeelding 5: Evaluatie scores per model

De r-squared (R2) score van een model geeft aan hoe goed de fit is van een model. Hoe dichter dit bij 1.0 is hoe beter. Het MLP en RR model presteren beter dan de baseline. Dit valt ook te zien in de foutmarges van de modellen. Het RR model presteert het beste op RMSE en MAE.



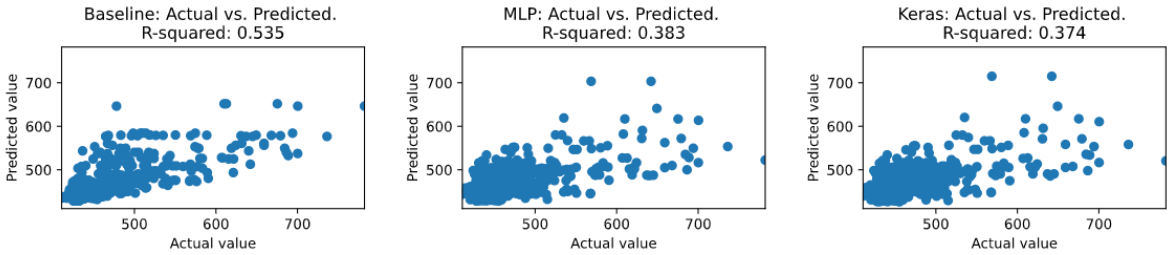
Afbeelding 6: Distributie van residuals per model

Het kan zijn dat er patronen zijn in de residuals, het verschil tussen de daadwerkelijke en voorspelde waarde. Dit is het geval bij het RF en DTR model. Beide hebben uitbijters in de residuals. Dit betekend dat er soms grote fouten gemaakt worden door het model.

Bevindingen

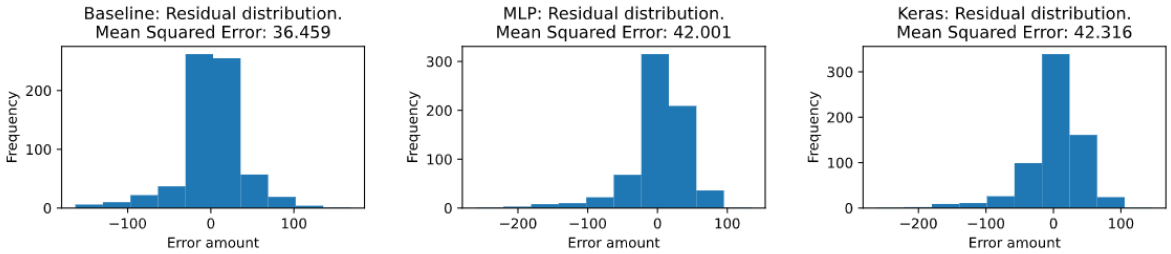
Net zoals bij het onderzoek van Kallio et al. presteert het ridge regression model het beste. Deze zal daarom toegepast worden in de applicatie.

Lange termijn

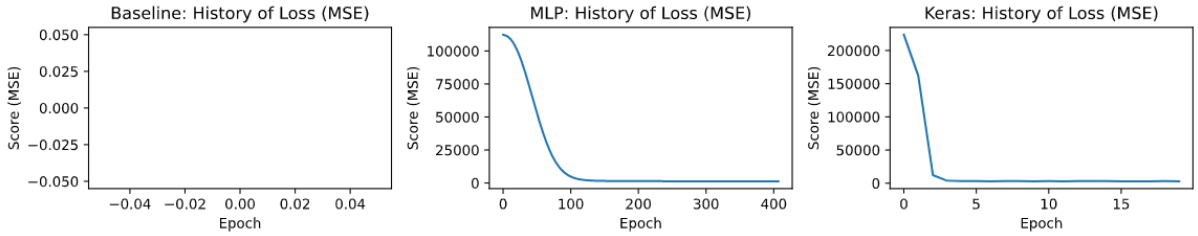


Afbeelding 6: Daadwerkelijk vs. voorspelt per ANN model

Ieder model, baseline inbegrepen, vertoont hetzelfde patroon. Er zitten meer lage waarden in de dataset dan hoge. Hierdoor zijn alle modellen gebiased richting lagere voorspellingen.



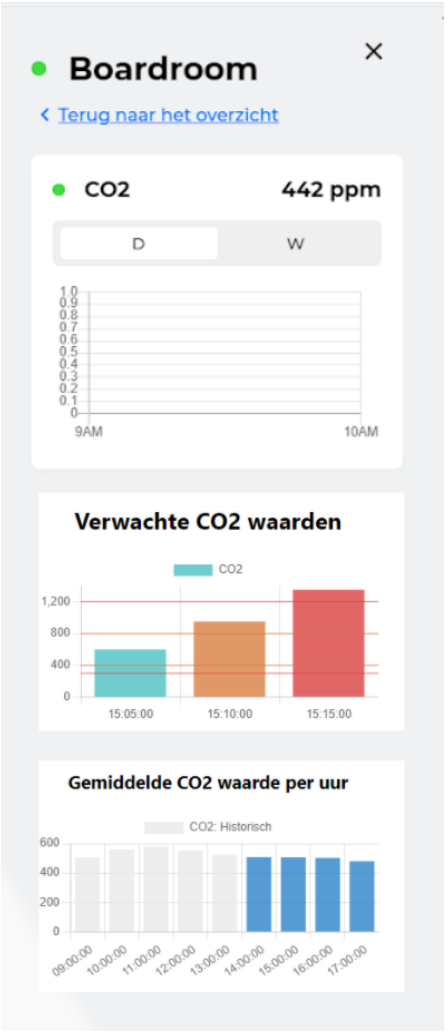
Afbeelding 7: Distributie van residuals per model



Afbeelding 8: Loss over epoch per model

Implementatie

In de afbeelding hiernaast is een mock up te zien van de implementatie van de modellen. Op deze manier kan er gezien worden wat er in de nabije- en verre toekomst verwacht wordt.



Conclusie

Vanuit de vorige iteratie waren er een viertal uitdagingen die opgelost moesten worden. Onderstaand worden deze, en de oplossingen, toegelicht.

1. **Onderzoeken wat invloed heeft op veranderingen in de meetwaarden en deze data toevoegen aan de modellen.**
   CO2 gehaltes worden vooral bepaald door het aantal personen in een ruimte. Door gebruik te maken van de google calendar API kan opgevraagd worden hoeveel personen zich op een bepaalde tijd in de ruimte bevonden.

2. **Verder in de toekomst kijken, 3 uur of meer.**
   In het onderzoek van Kallio et al. werd geconcludeerd dat voorspellingen voor de lange termijn vaak niet accuraat zijn. Hierdoor is er gekozen om een model te ontwikkelen wat op korte termijn accuraat is, deze zal ondersteunt worden door een visualisatie van gemiddelde voor een bepaalde dag en tijd.

3. **Betere controleren voor welk tijdstip de voorspelling gemaakt worden.**
   Door andere data voorbereidingstechnieken te gebruiken wordt er geen voorspelling gedaan voor een bepaald tijdstip. Hierdoor hoeft deze ook niet meer gecontroleerd te worden.

4. **Bestandsgrootte modellen verminderen om schaalbaarheid te vergroten.**
   De Ridge Regression modellen zijn minder dan 64 KB groot. Dit betekend dat het makkelijk schaalbaar is wanneer ze voor digital twins met veel ruimten moeten worden gebruikt.

Bronnen

Kallio, J., Tervonen, J., Räsänen, P., Mäkynen, R., Koivusaari, J., & Peltola, J. (2021). Forecasting office indoor CO2 concentration using machine learning with a one-year dataset. Building and Environment, 187, 107409. https://doi.org/10.1016/j.buildenv.2020.107409

Putra, J. C. P., Safrilah, & Ihsan, M. (2018). The prediction of indoor air quality in office room using artificial neural network. Published. https://doi.org/10.1063/1.5042896