Onderzoeksverslag
Smartbuilding

DE HAAGSE HOGESCHOOL

COLOFON

ONDERWIJSINSTELLING

Instelling	: Haagse Hogeschool
	: Faculteit IT & Design
Adres	: Johanna Westerdijkplein 75
	: 2521 EN
	: Den Haag
Telefoon	: 070-445 8888
Begeleider/Project lid	: Dr. J.D. Schagen (Jan Dirk)
Telefoon	: 06-14926995
E-mailadres	: J.D.Schagen@hhs.nl

OPDRACHTGEVER

Instelling	: Haagse Hogeschool te Delft
	: Lectoraat Energie en de Gebouwde Omgeving (LEGO)
Adres	: Rotterdamseweg 137
	: 2628 AL
	: Delft
Telefoon	: 015-260 6200
Hoofdlector	: Dr. Laure Itard
Telefoon	: 015-2606387
E-mailadres	: I.c.m.itard@hhs.nl

VOORWOORD

Dit document geeft verslag over de door ons uitgevoerde opdracht binnen het ‘Lectoraat Energie en Gebouwde Omgeving’ (hierna genoemd LEGO). Data Science: het onttrekken van kennis uit een grote hoeveelheid data, waarbij de data gestructureerd of ongestructureerd is. Dat is de kern van het onderzoeksproject bijhorende dit verslag.

De tien projectleden, waarvan negen studenten en één docent hebben allen een verschillend specialisme maar de affiniteit voor Data Science heeft ons bij elkaar gebracht. De afgelopen 18 weken hebben wij onze theoretische kennis, die wij tijdens de minor ‘Applied Data Science’ hebben opgedaan, toegepast bij dit onderzoek. Dit hebben wij uitgevoerd door middel van het opstellen van een onderzoeksrapport. Met dit rapport en het traject hierbij behorende, hebben wij een onderzoek gedaan voor ‘LEGO’.

Allereerst gaat onze dank uit naar onze opdrachtgever, Laure Itard en haar lectoraat, die een hoop tijd in ons hebben geïnvesteerd. Mede dankzij haar en de overige betrokken binnen het lectoraat, hebben wij dit onderzoeksrapport naar genoegdoening kunnen opleveren.

Wij hebben ons allen erg positief ontwikkeld gedurende de onderzoeks- en analyseperiode. Voor ons was het een leuke, maar voornamelijk een leerzame periode.

Daarnaast bedanken wij Jan Dirk Schagen, Jeroen Vuurens en Toni Andrioli, onze begeleiders vanuit faculteit ITD . Deze docenten hebben ons ondersteund met de theorie die wij hebben kunnen toepassen tijdens de opdracht. Mede hierdoor is dit eindresultaat verkregen.

Tot slot willen wij onze mede studenten en overige betrokkenen bedanken voor alle hulp.

SAMENVATTING

Het schoolgebouw van de Haagse Hogeschool te Delft maakt gebruik van een klimaatregelsysteem. De inkomende sensorwaardes worden opgeslagen in een database. Dit systeem was voorheen ‘Octalix’; nu wordt het systeem van ‘Priva’ gebruikt.
Het lectoraat ‘Energie en Gebouwde Omgeving’ (‘LEGO’) wilt onderzoeken, hoe anomalieën real-time kunnen worden gedetecteerd en gepresenteerd. Het uitgangspunt is het verminderen van het aantal klachten over het binnenklimaat per jaar. Om dit te realiseren worden de methoden: ‘Rule Based System’, ‘Bayesian Belief Network’, ’Deep Learning’ en ‘Clusteranalyse’ onderzocht. Het doel is bepalen welke methoden, alleen of samengesteld, het best bijdragen aan het real-time detecteren van anomalieën.

Als eerste is onderzocht welke anomalieën zich op dit moment voordoen. ‘Rule Based System’ heeft 66 anomalieën gevonden over een periode van 01-01-2012 t/m 23-06-2016. Hiervan zijn 64 anomalieën in dezelfde categorie: “CO2 sensor is waarschijnlijk kapot”. ‘Bayesian Belief Network’ heeft alle resultaten gevonden welke voorkwamen in het aangeleverde BBN model; overbezetting, luchtklep, ventilatie, PIR -, CO2 - en Airflow sensor. De ‘Clusteranalyse’ heeft twee anomalieën gevonden. Dit zijn een defecte aanwezigheidssensor en een defecte luchtklep. Ten tweede, op welke manier sensor data gebruikt kan worden om anomalieën op te sporen. Alle vier methodes kunnen gebruikt worden voor het opsporen van anomalieën. Alleen de ‘Cluster analyse’ en de ‘Rule Based System’ hebben daadwerkelijk anomalieën gevonden. Als laatste, hoe een gevonden anomalie real-time gemeld kan worden. Dit is mogelijk via web-applicatie. Het werkt reeds voor de ‘Rule Based System’ methode. Het is echter mogelijk om meerdere systemen te koppelen. Hierdoor kunnen verscheidene methodes, anomalieën doorgeven; welke methode dit dan ook afkomstig van is.
Een medewerker, de resultaten van ‘Machine learning’ te laten visualiseren, is ongewenst. Een ideale situatie is een dashboard dat middels een back-end systeem de uitkomst biedt.
De onderzoeksgroep heeft niet de geplande doelen bereikt. De niet bereikte doelen zijn: het toetsen van de validiteit van de vier methodes. Ook wordt aanbevolen om een vervolgonderzoek pas te starten wanneer een dataset verzameld is, waarvan minimaal 10% van de dataset bestaat uit gelabelde anomalieën. Tot slot raadt de onderzoeksgroep, de toekomstige onderzoeksgroep, aan om ervoor te zorgen dat de database in orde is. Hiermee wordt bedoeld dat alle kolommen logische benamingen krijgen, die aansluiten op het desbetreffende lokaal, waardoor de data in de toekomst op een efficiënte manier verkregen kan worden.

Uit de resultaten en conclusies van de vier methoden blijkt dat in een ideale situatie zowel de ‘clusteranalyse’ als een ‘LSTM’ wordt gecombineerd in een ‘Backend analyse systeem’. Dit analyse systeem is gekoppeld aan een webapplicatie waarin de gebruiker real-time meldingen ontvangt over gevonden anomalieën.

INHOUDSOPGAVE

1 INLEIDING

1.1 Context

1.1.1 Onderzoekscontext

De opdrachtgever van het onderzoek is het 'lectoraat Energie en de Gebouwde omgeving'. 'LEGO' is opgezet in 2010, omdat binnenklimaateisen en bezetting van gebouwen een onverwacht karakter hebben (c). "Hierdoor zijn klimaatbeheersingssystemen vaak niet efficiënt afgesteld en verbruiken ze meer energie dan oorspronkelijk verwacht werd" (Itard, 2016). Het installeren van hernieuwbare energiebronnen betekent helaas niet de oplossing voor het probleem. Het tegendeel is meestal waar verteld mevrouw Itard: "… doordat hernieuwbare energiebronnen en oude systemen niet correct op elkaar worden afgesteld wordt vaak volgens het 'LEGO' het energieverbruik van gebouwen minder efficiënt".
Om dit proberen te verhelpen heeft het 'LEGO' de volgende hoofdvraag opgesteld (Itard, 2016):

"Hoe kunnen gebouw- en stedelijke installaties bijdragen aan een duurzame energievoorziening en aan een gezond en comfortabel binnenmilieu?"

Om de bovenstaande onderzoeksvragen te beantwoorden richt het onderzoek van het lectoraat zich op de onderstaande deelvragen (Itard, 2016):

"1. Kunnen slimme diagnose- en regeltechnische methoden ontwikkeld worden voor het continu optimaliseren van energie- en binnenklimaatprestaties?"

"2. Hoe kunnen deze methoden bruikbaar gemaakt worden voor facility managers en voor het onderwijs?"

"3. Bieden smart DC-elektriciteitsnetten mogelijkheden voor energiebesparing en voor een betere integratie van duurzame bronnen in het huidige net?"

Het onderzoek dat de projectleden zullen uitvoeren, is opgezet om een bijdrage te leveren aan het beantwoorden van de eerste van de drie deelvragen. Hiervoor wordt in samenwerking met de lector, de kenniskring van het lectoraat en de docenten van de minor 'Applied Data Science' het onderzoek uitgevoerd in het schoolgebouw van 'De Haagse Hogeschool' te Delft.

1.1.3 Het gebouw

In 2009 is het gebouw van De Haagse Hogeschool geopend op de campus van de TU Delft. Bij de constructie van het gebouw is nagedacht over de duurzaamheid ervan (Haagse Hogeschool, 2018). Hierdoor is het in schooljaar 2010-2011 uitgeroepen als een van de duurzaamste gebouwen in Nederland met een GreenCalc+ score van 256 (Agentschap NL, 2010). Dit is een weergave van de impact die het gebouw uit op het milieu bij het bouwen en beheren.
In het gebouw zijn meerdere onderzoeken uitgevoerd, bijvoorbeeld naar het complexe HVAC systeem wat de temperatuur in het gebouw beheert (Salcedo, n.d.). De lokalen van de HHS worden verwarmd door middel van vloerverwarming en plafondpanelen waar water doorheen stroomt. Het gebouw heeft daarnaast een klimaatsysteem ook wel genaamd BMS (Building Management System). Dit systeem verzameld dat data afkomstig van de sensoren in het gebouw.

Tot slot worden deze gegevens opgeslagen in een database die beheerd wordt door de softwareleveranciers Octalix en Priva welke ook helpen bij het monitoren en analyseren van de data.

1.2. Aanleiding

De Haagse Hogeschool te Delft werkt volgens een automatisch klimaatregelsysteem. Deze werkt echter niet altijd naar wens. Bij de gebouwbeheerder zijn meerdere meldingen geweest van onder andere ongewenste klimaten, zonder dat het systeem dit overzichtelijk had gemeld. De voornaamste reden van ongewenste klimaten zijn defecte sensoren, maar het is echter niet altijd de oorzaak. Een voorbeeld hiervan is dat de buitentemperatuur te warm kan zijn, waardoor het systeem de gewenste temperatuur binnen het gebouw niet in stand kan houden. Dit voorbeeld is een van vele voorkomende anomalieën binnen het klimaatbeheersingssysteem.

Van de grofweg 12.000 sensoren in de vestiging Delft worden meetgegevens naar een database gestuurd. Het klimaatsysteem regelt hiermee zelf de instellingen van haar actuatoren, zoals luchtkleppen en verlichting (Itard, 2016).
Deze meetgegevens zijn alleen nog beschrijvend. Dat wil zeggen dat voor een foutief gemeten waarde, de oorzaak niet wordt aangegeven door het klimaatregelsysteem. Hierdoor worden fouten in het systeem pas gevonden wanneer gebruikers van het gebouw klachten melden. Ook geeft het klimaatregelsysteem de oorzaak van een foutieve meetwaarde niet aan. Dit is een voorbeeld van een melding via mail:

- Ruimte 0.033: luchtdebiet te laag (<80%) (1 uur)
- Ruimte 0.067: luchtdebiet te laag (<80%) (1 uur)
- Ruimte 0.075: luchtdebiet te laag (<80%) (1 uur)
- Ruimte 0.126: CO2 te hoog (5 uur)
- Ruimte 0.136: CO2 te hoog (2 uur)
- Ruimte 0.136: luchtdebiet te laag (<80%) (2 uur)
- Ruimte 1.015: CO2 te hoog (6 uur)
- Ruimte 1.015: luchtdebiet te laag (<80%) (8 uur)
- Ruimte 1.031: CO2 te hoog (5 uur)
- Ruimte 1.031: luchtdebiet te laag (<80%) (1 uur)
- Ruimte 1.032: CO2 te hoog (1 uur)
- Ruimte 1.032: luchtdebiet te laag (<80%) (1 uur)

Op dit moment komen er tussen de 35 en de 50 klachten binnen per jaar (Timp, 2018). In een ideale situatie is men de klachtenmelding voor.

Na onderzoeken en projecten in opdracht van het ‘LEGO’ zijn een tweetal programma’s opgeleverd door studenten waar handmatig data, in de vorm van een CSV-bestand, moet worden ingevoerd:

- Expert applicatie (Kortekaas & Vuuren, 2016):
  Data verwerkt door MonaVisa kan worden opgehaald als csv-bestand. Dit csv-bestand moet handmatig als input worden opgegeven waarna de data geanalyseerd wordt aan de hand van vooraf vastgestelde scenario’s.
- SAW (Ast, Scholte, & Wazir, 2016):
  De Sensor Application Wrapper, kort genoemd SAW, kan de sensordata ophalen uit

een mirror van de database en interpoleren waar nodig. De technieken uit eerdere projecten waarmee de gegevens kunnen worden geanalyseerd (Expert applicatie en SMILE) worden in deze applicatie geïntegreerd, zodat de gebruiker hiertussen kan kiezen. De analyseresultaten worden vervolgens op een overzichtelijke manier gepresenteerd.

Op het moment is 'LEGO' niet tevreden over de programma's en wilt het liefst een programma dat automatische analyses maakt en de defecten meldt aan bijvoorbeeld de gebouwbeheerder.

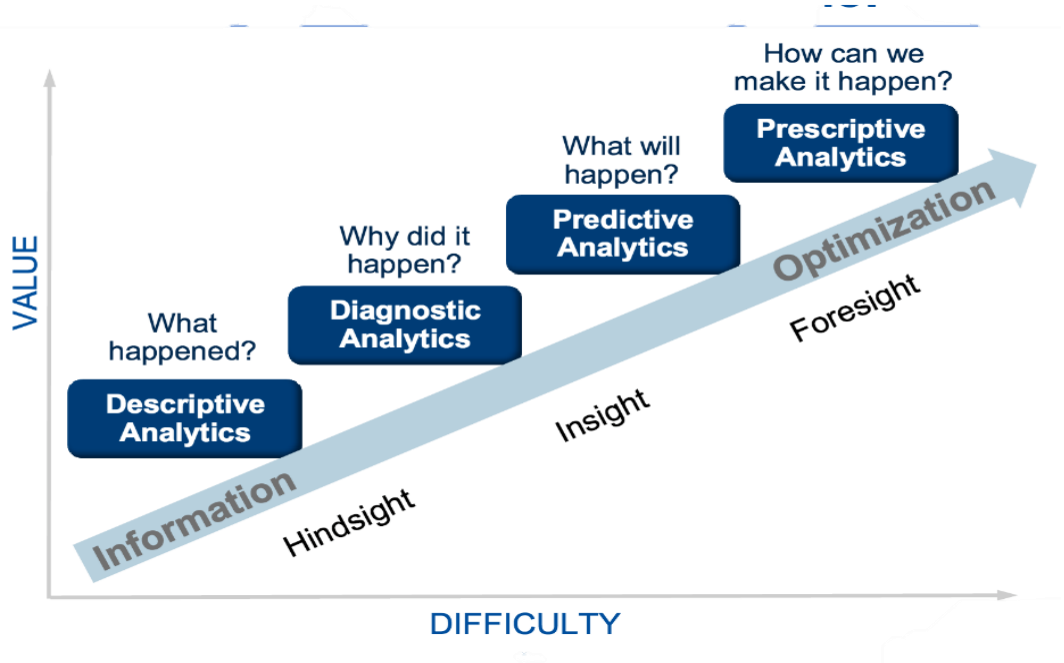
1.3 Onderzoeksopzet

1.3.1 Probleemstelling

Op dit moment wordt wel beschrijvende data over het klimaat in het schoolgebouw te Delft verzameld. Het klimaatregelsysteem geeft echter niet aan wanneer sensorwaardes afwijken of wat de achterliggende oorzaak is van de afwijkende sensorwaardes. Daarnaast worden deze waardes niet automatisch verzameld en gemeld. Hierdoor is het opsporen van defecten tijdrovend en niet energie-efficiënt.

1.3.2 Typering van het probleem

Zoals in de aanleiding beschreven is, wordt onnodig tijd verspilt aan het opsporen en melden van fouten in het klimaatbeheersingssysteem. Dit probleem bevindt zich in de scope van 'Data Maturity'. Het systeem zoals het nu aanwezig is, levert namelijk alleen maar beschrijvende data aan. Diagnostische data of voorspellende data wordt echter nog niet verzameld. In het 'Data Quality Maturity' model van Gartner bevindt de school zich dan ook pas in fase 1 van het model. Dit model is in figuur 1.3.1 weergeven. Dit model is door Gartner ontworpen om de verschillende fases in beeld te brengen, de groei in waarde die elke fase met zich meebrengt en de moeilijkheidsgraad om iedere fase te bereiken.



Figuur 1.3.1; Data Quality Maturity model

Het ‘LEGO’ is daarom onderhand druk bezig met het vergaren van informatie en data om fase 2 te bereiken. Dit onderzoek zal een bijdrage leveren om deze fase te kunnen bereiken en de eerste stappen zetten richting fase 3.

Dit document beschrijft het onderzoek naar de sensordata van het klimaatsysteem, om hieruit af te kunnen leiden wat de status van verschillende sensoren op een willekeurig moment is. Met deze data zou het namelijk mogelijk moeten zijn om aan te kunnen wijzen wat de oorzaak is van afwijkend of onverwacht gedrag in het klimaatregelsysteem (Ast et al., 2016)

1.3.3 Doelstelling

Het doel van het onderzoek is om afwijkende sensorwaardes, in het klimaatbeheersingssysteem van de Haagse Hogeschool te Delft, zichtbaar te maken inclusief de bijbehorende oorzaak van de afwijkende sensorwaardes. Ook zal op basis van het onderzoek en advies nodig zijn waarin vervolgstappen en aanbevelingen beschreven staan om fase drie in het ‘Data Quality Maturity’ model van Gartner te bereiken. Nadat de oplossing van de projectgroep is geïmplementeerd, zou het aantal klachtenmeldingen sterk vermindert moeten worden.

Het doel is om dit voor vijf Februari 2018 op te leveren in de vorm van een onderzoeksrapport en de resultaten van het onderzoeksrapport te publiceren.

1.3.4 Globale aanpak

Om het probleem van ongewenste klimaten en laat ontdekte defecten van sensoren te verhelpen, is besloten om een analyseprogramma te maken. Het programma moet het volgende kunnen:

- Automatisch data analyseren en daaruit anomalieën detecteren;
- Gevonden anomalieën overzichtelijk verklaren.

Het visueel weergeven van de data kan helpen bij het inzicht krijgen in de data en bij trendanalyses waarbij real-time inzicht wordt gegeven aan de eigenaar/gebouwbeheerder over het functioneren van het gebouwsysteem.

Voor het detecteren van anomalieën moet de inzet van algoritmes en programma’s onderzocht worden zodat meldingen over uitzonderlijke situaties overzichtelijk worden verklaard. Een mogelijkheid is het ontwerpen van een intelligent autonoom systeem, welk diagnoses en foutmeldingen maakt en automatisch componenten reset. Dit is echter niet de primaire focus van dit onderzoek.

Het doel is om afwijkende sensorwaardes zichtbaar te maken. Hier zijn veel verschillende methodes voor. Aan de hand van literatuuronderzoek zal een selectie worden gemaakt van vier verschillende methodes, deze worden vervolgens met elkaar vergeleken.
Omdat de data niet direct bruikbaar is gebruikt de projectgroep een stappenplan om tot het eindproduct te komen. Dit stappenplan is weergeven in figuur 1.3.2.

Figuur 1.3.2; Stappenplan tot stand komen van het eindproduct

1.3.5 Onderzoeksvragen

Hoofdvraag
Hoe kan sensordata van De Haagse Hogeschool te Delft automatisch worden geanalyseerd en gepresenteerd zodat anomalieën real time gemeld worden?

Dit is een afgeleide van de vraag die gesteld wordt door het 'LEGO' in de opdracht die ons gegeven is; "Hoe kan de sensordata zo worden geanalyseerd en gepresenteerd dat problemen in het systeem snel opgemerkt kunnen worden?", (Itard, 2016). Door de vraag verder te specificeren, wordt het makkelijker om het resultaat te toetsen.

Deelvraag 1
Welke anomalieën vinden nu plaats in De Haagse Hogeschool te Delft?

Om afwijkingen te kunnen detecteren zal eerst de huidige situatie onderzocht moeten worden. Omdat hierdoor de norm wordt vastgesteld wordt het hierdoor gemakkelijker om anomalieën te herkennen.

Deelvraag 2
Op welke manier kan sensordata gebruikt worden om anomalieën op te sporen?

Er zijn velen manieren mogelijk om data te analyseren, welke methodes hiervan relevant zijn voor de probleemstelling worden onderzocht. De accuraatheid van de data analyses zal per relevante methode getest worden, zodat een realistische aanbeveling gegeven kan worden.

Deelvraag 3

Op welke manier kunnen gevonden anomalieën gebruikt worden om defecten te melden?

Het laatste belangrijke aspect is het melden van de defecten, op dit moment worden alle afwijkende waarden per mail verstuurd. Het nadeel hiervan is dat het al snel bedorven kan worden onder alle andere mails en misschien niet zo overzichtelijk kan zijn.

1.4 Leeswijzer

Dit onderzoeksrapport is ingedeeld in acht hoofdstukken. In hoofdstuk één is de inleiding te vinden, bestaande uit de context, de aanleiding en de onderzoeksopzet. In de context is relevante informatie te vinden aangaande het lectoraat waarvoor het onderzoek wordt uitgevoerd. Ook wordt hierin relevante informatie over het schoolgebouw te Delft verstrekt. In de aanleiding is het probleem aangekaart. De onderzoeksopzet beschrijft de probleemstelling, de probleemtypering, de doelstelling, de globale aanpak en de onderzoeksvragen. In hoofdstuk 2 wordt het theoretisch kader gepresenteerd.

In het hoofdstuk ‘methode’ worden de gebruikte methoden tijdens het project en de redenering voor het gebruiken van deze methoden. Hierna volgt hoofdstuk vier waarin de resultaten per methode zijn uitgewerkt. Na de resultaten volgt de conclusie van het onderzoek. Vervolgens worden een aantal aanbevelingen beschreven waarvan de onderzoeksgroep acht dat zij relevant zijn voor vervolgonderzoeken. De laatste hoofdstukken bevatten de literatuurlijst en de bijlages.

2 THEORETISCH KADER

Om ervoor te zorgen dat het onderzoek duidelijk kan worden beschreven moet er research worden gedaan naar de aansluiting op en resultaten van voorgaande onderzoeken. Om deelvraag twee van het onderzoek te beantwoorden zijn vier verschillende methodes met elkaar vergeleken, deze methodes zijn: ‘Rule Based System’, ‘Bayesian Belief Network’, ‘Deep learning’ en ‘Cluster analyses’. Waarom de projectgroep deze methodes gebruikt wordt uitgelegd in het volgende hoofdstuk. Voordat dit gebeurd is het echter belangrijk dat men weet wat deze methodes inhouden.

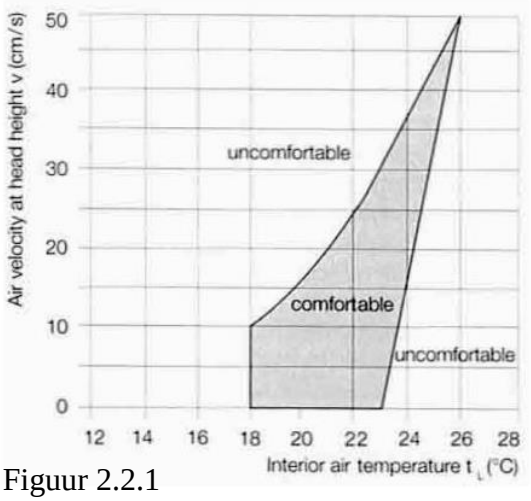
2.1 Aansluiting

In voorgaande onderzoeken zijn er onder andere het analyseren en toepassen van de verschillende data met behulp van twee analyse modellen onderzocht. Tevens is er onderzocht hoe één van deze twee modellen, het Bayesaanse model, kan worden gebruikt in het opsporen van problemen in het klimaat van een gebouw. In dit onderzoek wordt hierin aansluiting gegeven door te onderzoeken op welke manier deze modellen kunnen worden gebruikt voor het automatiseren van de analyse en detectie voor het klimaat in het gebouw.

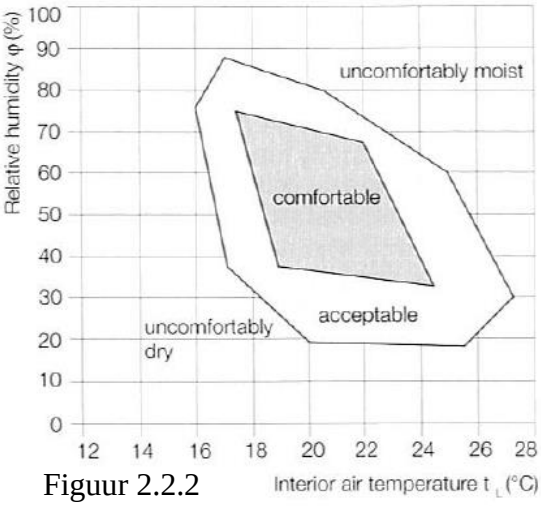
2.2 Comfort

Het comfort van het binnenklimaat van de Haagse Hogeschool wordt ervaren als drie verschillende type comfort (Marcopoloulos, 2014). Dit zijn thermisch comfort, visueel comfort en ventilatie comfort (Marcopoloulos, 2014). In voorgaand onderzoek, worden deze types als volgt gedefinieerd. Thermisch comfort: de gevoelsmatige temperatuur ervaren in een ruimte. Ventilatie comfort: de ervaren luchtkwaliteit in een ruimte. Visueel comfort: de hoeveelheid ervaren licht bestaande uit zowel kunstlicht en zonlicht.

Zoals uit de definities al blijkt, wordt comfort bepaald door degene die het ervaart. Comfort is dus voor ieder persoon anders. Echter zijn in voorgaande onderzoeken statistische gegevens verzameld om toch uitspraken te kunnen maken over het comfort. De uitkomst van deze onderzoeken is weergeven in figuur 2.2.1 en 2.2.2 (Riemer, 2017).Dit is enkel nog de uitkomst voor de thermische comfort bepaling en een gedeelte van de ventilatie comfort.



Figuur 2.2.1



Figuur 2.2.2

De ventilatie comfort wordt ook bepaald door de hoeveelheid co2 in de lucht. TNO heeft onderzoek gedaan naar maximale co2 waardes waaronder studenten nog steeds kunnen presteren. Zij raden aan om 1000 ppm co2 als limiet aan te houden (Jacobs, Dijken, & Boerstra, 2007). Het visueel comfort wordt simpel weg bepaald door de hoeveelheid daglicht en kunstlicht. 83% van de bezoekers van het schoolgebouw in Delft zijn tevreden over het daglicht dat aanwezig is (Marcopoloulos, 2014). Men heeft ook een eens gelijke tevredenheid over het kunstlicht (Marcopoloulos, 2014). Wanneer de aanwezigheidssensoren echter defect zijn is er vanzelfsprekend onvoldoende kunstlicht en ontvangt de gebouwbeheerder hier klachten over.

2.3 Klimaatregelsysteem

Onder klimaatregelsysteem wordt het energiesysteem verstaan dat de invoer van lucht, de temperatuur en het kunstlicht in het schoolgebouw regelt. Dit wordt uitgevoerd door het regelsysteem ‘Octalix’ (Goedhart, 2015). De sensorwaardes die door het klimaatregelsysteem gelezen woorden zijn aanwezigheidssensoren, infraroodsensoren, geluid- en dauwpuntsensoren, temperatuursensoren, lichtsensoren, CO2 sensoren en raamcontact sensoren.

2.4 Anomalieën

Volgens het onderzoek ‘Principled Sampling for Anomaly Detection’ bestaan anomalieën uit gemeten afwijkende waardes. In dit onderzoeksrapport zal dezelfde definitie gebruikt worden (Juba, 2017). Een belangrijk onderscheid dat wordt gemaakt in het genoemde onderzoeksrapport is het verschil tussen afwijkende waardes die overeenkomen met de werkelijke waardes en afwijkende waardes die niet overeenkomen met de daadwerkelijke waardes. In dit onderzoek zullen beide soorten opgespoord worden.

2.5 BBN Model

Het Bayesian Belief Network, hierna BBN, is aangeraden door de opdrachtgever, het ‘LEGO’. Door het lectoraat zijn verscheidene artikelen gepubliceerd over deze methode in combinatie met het analyseren van een gebouw.

In een eerder onderzoek is geschreven over het toepassen van ‘BBN’ in andere velden. Op Medisch gebied wordt het gebruikt om commerciële computerondersteunde diagnostische beslissingsondersteunende systemen te ontwikkelen. Op industrieel gebied hebben de op BBN gebaseerde diagnostische systemen veel belangstelling getrokken. Toepassingen zijn te vinden in kern-energiesystemen, vliegtuigmotoren, sensor-foutdetectie en -identificatie, halfgeleider-productiesystemen, enz. Door de conditionele kans heeft ‘BBN’ superieure prestaties getoond in vergelijking met neurale netwerken, support vector machines, beslissingsbomen, etc. Bayesian Network wordt een steeds belangrijker gebied van onderzoek en toepassing op het gebied van kunstmatige intelligentie (Ast et al., 2016). Er zijn echter weinig toepassingen van BBN op het gebied van HVAC.
In een andere studie wordt BBN wel toegepast, maar dan in combinatie met een expertsysteem (rule-based system) op het warmteopwekkingssysteem van De Haagse Hogeschool te Delft (A. Taal, L. Itard. Y. Zhao, 2015)

Voor het gebouw van de HHS Delft is een applicatie gemaakt waarin een rule-based system gebruikt wordt, verder is een eerste versie gemaakt van een BBN (Kortekaas & Vuuren, 2016). De applicatie is na dat onderzoek verder uitgewerkt, men heeft de keuze gekregen om te analyseren via rule-based of BBN (Ast et al., 2016). In een vervolgonderzoek wordt het

BBN gedeelte van de applicatie eruit gehaald, en verder uitgewerkt (Ast et al., 2016). Uit bovenstaande onderzoeken is geconcludeerd dat het mogelijk is om in een applicatie, een BBN model te importeren en data van een gebouw te analyseren.

2.6 Rule Based System

Rule Based Systems (RBS) is een manier van data opslaan en manipuleren zodat de gewenste resultaten naar boven komen. Het systeem werkt op basis van regels of statements die bepalen wat er moet gebeuren. Het voordeel hiervan is dat jijzelf de regel waarden bepaald en deze vast staan. Bij andere modellen als 'Machine Learning' en 'Bayesian Belief Network' kunnen deze bepaalde regel-waarden veranderen doordat het systeem leert uit eerdere ervaringen. Wanneer een afwijkende waarde erg vaak voorkomt, kan het systeem dit gaan zien als normaal. Daarnaast is het RBS relatief gemakkelijk in elkaar te zetten. Het vergt relatief weinig van een ontwikkelaar en de kosten voor het ontwikkelen van het systeem liggen lager ten opzichte van andere vergelijkbare systemen.

In een voorgaand onderzoek is er een systeem/applicatie ontwikkeld waarin de gegevens van de Haagse Hogeschool te Delft worden geanalyseerd door middel van een RBS en een Bayesian Belief Network (BBN) (Ast et al., 2016).

Het nadeel is dat voor elk nieuw onderdeel dat wordt toegevoegd er weer een aantal statements in de code moeten worden toegevoegd. Wanneer dit niet volgens een structuur gebeurd is het onoverkomelijk dat er fouten ontstaan en dat het systeem dus niet meer (optimaal) werkt. Daarnaast is een RBS minder goed in het voorspellen van de toekomst van andere modellen zoals 'Machine learning' en 'Bayesian Belief Network'. Wanneer er in de toekomst verschuivingen in waarden plaatsvinden kunnen de andere hiervoor genoemde modellen zich automatisch aanpassen, het RBS is hier niet toe instaat. Bij het wijzigen van het systeem is er daarnaast ook een expert nodig welke het afgestemde systeem kent en situaties kan toevoegen of wijzigen.

2.7 Deep Learning

Sinds 2006 wordt 'Deep learning' gezien als een sub-categorie binnen 'Machine learning' (Yu, Woradechjumroen, & Yu, 2013). 'Machine learning' is de mogelijkheid om met een algoritme uit ruwe data kennis op te doen. Het doel van machine learning is het mogelijk maken voor een computer om problemen op te lossen die subjectief lijken (Goodfellow, Bengio, & Courville, n.d.). Onder 'Deep learning' vallen alle bestaande 'Machine Learning' methodes, die gebaseerd zijn op algoritmes die meerdere levels binnen een model leren complexe verbanden in de gegeven data te herkennen (Deng & Yu, 2013). De categorie 'Deep learning' bestaat zelf weer uit drie subcategorieën Dit zijn de categorieën: 'supervised learning', 'semi-supervised learning' en 'unsupervised learning' (Deng & Yu, 2013). Alhoewel de drie categorieën van elkaar verschillen hebben zij allemaal de functie om data te classificeren. De manier waarop zij dit doen is wat het verschil maakt.

Neurale 'supervised learning' netwerken worden gebruikt voor het classificeren van data altijd "targetdata" met labels. Het netwerk krijgt voorbeeld data waarvan de classificatie al bekend is. Hierdoor leert het netwerk nieuwe data te classificeren.

Neurale 'semi-supervised learning' netwerken gebruiken een set gelabelde data en een set niet gelabelde data (Deng & Yu, 2013). Dit type netwerk wordt geprefereerd wanneer men het

classificeren van data niet volledig wil overlaten aan het algoritme. Doordat het algoritme geassisteerd wordt leidt dit in veel gevallen tot wenselijkere uitkomsten (Deng & Yu, 2013).

Neurale ‘unsupervised learning’ netwerken worden gebruikt wanneer er geen labels voor de “targetdata” beschikbaar zijn en sterke verbanden in de data geschetst dienen te worden om classificaties en data na te bootsen of te voorspellen (Deng & Yu, 2013).
De projectgroep onderzoekt verschillende ‘Deep learning’ methodes omdat de sensorwaardes van het klimaatsysteem geclassificeerd dienen te worden om afwijkende sensorwaardes te herkennen. Welke methode gebruikt kan worden zal beschreven worden in hoofdstuk vier van het onderzoeksrapport.

3 METHODEN

Om een zo breed mogelijke basis te kunnen onderzoeken is er besloten om onderzoek te doen naar een drietal methoden. Deze methoden zijn aangeraden door het ‘LEGO’. Allereerst is de methode van een ‘Rule Based System’ besproken in paragraaf 3.1. Een andere groep heeft zich beziggehouden met een ‘Bayesian Belief Network‘, zoals te lezen in paragraaf 3.2. De laatste groep was hield zich bezig met ‘Deep Learning’, hun methode staat beschreven in paragraaf 3.3. Tenslotte is er nog een klein onderzoek gedaan naar bepaalde clusteringsmethoden, te lezen in paragraaf 3.4. In elk van deze paragrafen wordt de methoden uitgelegd. En wordt antwoord gegeven op de vragen “Waarom is er voor deze methode gekozen?” en “Op welke manier wordt deze methode onderzocht?”.

3.1. Rule Based System

In een voorgaand onderzoek is een Rule Based System, hierna RBS, gebouwd om de beschikbare data uit het klimaatsysteem te analyseren. Een RBS wordt ontwikkeld door middel van het opstellen van regels waarbinnen de data moet blijven voor een ‘normale’ werking van het systeem. Wanneer de data niet binnen deze regels valt wijkt het af en wordt het dus niet meer als het ‘normaal’ gezien. Dit kan betekenen dat een sensor defect is of een andere oorzaak aan ten grondslag ligt. De ontwikkeling van het RBS diende vooral voor het verkrijgen van een validatie middel voor het lectoraat. Met behulp van het RBS konden nieuwere methoden worden gecontroleerd op de werking ervan.

Voor dit onderzoek zijn er aantal taken uitgevoerd op het gebied van RBS:

- Het werkend krijgen van het RBS in een nieuwe omgeving.
- Het analyseren en visualiseren van de eerder verkregen resultaten uit het RBS.
- Het uitbreiden van het RBS en toegankelijkheid verhogen.

3.2. Bayesian Belief Network

BBN wordt vooral gebruikt in de medische wereld. Het pand wordt daarom ook door het ‘LEGO’ vergeleken met een ziek lichaam; het heeft een aantal verschillende symptomen en uit daaruit is een kans aanwezig op een ziekte. Dit is ook toepasbaar voor een gebouw: een te hoge CO2 waarde, te hoge temperatuur, te lage airflow zijn bijvoorbeeld symptomen. Een kapotte ventilatie, CO2-meter, overbezetting, etc. zijn voorbeelden van resultaten/ziektes. Door de symptomen te combineren en gebruikt te maken van de stelling van Bayes, is de conditionele kans op een resultaat te berekenen. Door in plaats van een definitief resultaat, de kans erop neer te zetten kan ook nog rekening gehouden worden met andere mogelijke resultaten.

Er zijn een tweetal taken uitgevoerd door deze groep:

- Een BBN model maken en gebruiken in Python.
- Het bestaande analyseprogramma (smileApp) uitbreiden/verbeteren.

Beiden taken maken gebruik van het BBN model dat is gekregen van het ‘LEGO’.

3.3. Deep Learning

Uit de vier gekozen methodes is Deep Learning gekozen omdat Laure Itard de vraag stelde; ‘Is het mogelijk om een systeem dat anomalieën detecteert te ontwikkelen, zonder de voorkennis van een expert’. Waarbij de definitie van een expert is: iemand die zeer deskundig is op een bepaald gebied.

Deep Learning is een type Machine Learning welke geïnspireerd is door het menselijke brein. Binnen Deep Learning zijn er veel verschillende neurale netwerken bedacht, maar ze hebben allemaal de overeenkomst dat ze zijn opgebouwd uit zogenaamde neuronen. Een neuron doet, net als bij het menselijke brein, niet zo veel op zichzelf. Door deze neuronen op een bepaalde manier aan elkaar te koppelen wordt er een netwerk gecreëerd. Door gewichten aan bepaalde neuronen te geven kan dit netwerk zelf leren welke inputwaarden belangrijk zijn voor een bepaalde output. Hierdoor zou het zelf kunnen leren wat de sensorwaardes zouden moeten zijn in normale situatie. Zodra bekend is hoe sensoren zich normaal gedragen kan dit model gebruikt worden als basis. Alles wat een grote afwijking heeft van dit model zou een anomalie kunnen zijn. Bijvoorbeeld als de CO2 waarde normaal om 12:00 tussen de 600 en 800 ligt maar nu ineens 200 is.

Voor dit onderzoek zijn er een aantal taken uitgevoerd op gebied van Deep Learning:

- Onderzoek welk neuraal netwerk geschikt is voor ons probleem.
- Een neuraal netwerk maken die anomalieën in een tijdserie kan detecteren.

3.4. Clustermethoden

Clusteren is een techniek waarmee onderzoeksobjecten worden geclassificeerd. Denk hier bijvoorbeeld aan dierenpopulatie wat geclassificeerd wordt in reptielen, vissen, zoogdieren etc. Deze techniek is tevens binnen het gebied van data science zeer bruikbaar voor het classificeren van data. Tegenwoordig gebruiken bedrijven dit om hun online klanten te classificeren op basis van gebruikersgegevens waardoor gedrag van nieuwe klanten voorspeld kan worden op basis van bestaande klanten.

Voor dit project is een enorme hoeveelheid data beschikbaar waar weinig over bekend is wat het moeilijk maakt om concrete uitspraken te doen over het gedrag van het klimaatregelsysteem. Daarom is onderzoek gedaan naar de mogelijkheid om data te clusteren en vervolgens te classificeren.

Hiervoor zijn de volgende taken uitgevoerd:

- Onderzoek naar toepasbaarheid Factoranalyse op de beschikbare data.
- Onderzoek naar verschillende clusters herkenning algoritmen.

4 RESULTATEN

In dit hoofdstuk worden de resultaten van RBS, BBN, 'Deep Learning' en Clusteranalyse gepresenteerd. Ten eerste zullen in paragraaf 4.1. de resultaten van het RBS besproken worden. Als tweede zullen de resultaten van het Bayesiaanse model gegeven worden, deze zijn te vinden in paragraaf 4.2. Ten derde zal in paragraaf 4.3. de resultaten van het Deep Learning onderzoek worden weergegeven. Tenslotte worden in paragraaf 4.4. de resultaten van de clustermethode worden besproken.

4.1. Rule Based System

Het RBS is, in een eerder onderzoek, ontwikkeld in een samengestelde applicatie genaamd: SAW (Sensor Application Wrapper). Deze applicatie heeft voorheen correct gewerkt, wat resulteerde in een analyse door de SAW applicatie van het RBS onderdeel. Onfortuinlijk genoeg was het niet mogelijk om het RBS onderdeel opnieuw werkend te krijgen. De applicatie vertoonde op basis van de opgehaalde data geen output of resultaten. Nadat de applicatie meerdere malen onderzocht was, is er besloten om door te gaan met de al aanwezige resultaten van het RBS uit het voorgaande onderzoek waarin het systeem was ontwikkeld. (zie figuur 4.1.1)

		Normal situation	Empty classroom	Lots of people present	TEMP sensor probably broken	PIR sensor probably broken	AIR sensor probably broken	CO2 sensor probably broken	Normal situation	Empty classroom	Lots of people present	TEMP sensor probably broken	PIR sensor probably broken	AIR sensor probably broken	CO2 sensor probably broken	Normal situation	Empty classroom	Lots of people present	TEMP sensor probably broken	PIR sensor probably broken	AIR sensor probably broken	CO2 sensor probably broken	Normal situation	Empty classroom	Lots of people present	TEMP sensor probably broken	PIR sensor probably broken	AIR sensor probably broken	CO2 sensor probably broken	Normal situation	Empty classroom	Lots of people present
0.005 (B4.01) Studielandschap								761							713							760							737			
0.014 (B3.03a) Praktijkruimte								761							713							760							737			
0.014 (B3.03b B3.04) Praktijkruimte								761							713							760							737			
0.031 (E07) Facilitair Kantoor								761							713							760							736			
0.033 (E05) Schoonmaak Kantoor								761							713							760							737			
0.035 (E02) Opslag Facilitair bedrijf								761							713							760							737			
0.051 (A1.04a) Praktijkruimte		164	574	3		8	12		243	470						202	558						257	480						486	249	25
0.052 (A1.01) Studielandschap		227	515	4		5	10		324	386	3					242	449	39			1	29	308	423	3			3		507	202	47
0.067 (A1.03a) Praktijkruimte		41	550	123		17	30		78	460	134			41		23	545	124		6	62		10	478	130		1	118			246	457
0.073 (A1.03b) Praktijkruimte		99	576	66		9	11		167	472	59			15		73	527	111			20	29	78	488	103			68		28	250	446
0.075 (A1.03c) Praktijkruimte		13	657	47		6	38		20	558	50		1	84		4	641	64			51		1	558	58			120			691	42
0.081 (A1.04c) Praktijkruimte								761							713							748							737			
0.087 (B4.03) Praktijkruimte								761							713							760							737			
0.097 (B4.04) Praktijkruimte								761							713							754							737			
0.112 (C11/12) Regio Regisseur								761							713							759							737			
0.114 (D07/08) Flexplek		136	523	73		16	13		179	382	92		13	47		164	480	101		2	13		134	451	85		3	64		64	219	467
0.118 (C1.10) EHBO								761							713							760							737			
0.126 (C04) Zaal groot								761							713							760							737			
0.136 (C03) Zaal								761							713							760							737			
1.013 [A2.04b] Praktijkruimte								761							713							760							737			
1.015 (A2.04a) Praktijkruimte																																
1.021 (B3.02) Instructieruimte								761							713							760							737			
1.031 (B3.08a) Werkkamer		88	645	17		6	5		207	428	68		1	9		139	555	59		1	6		124	536	59			18		39	676	31
1.032 (B3.06a) Spreekkamer								761							713							760							737			
1.033 (B3.08b) Werkkamer								761							713							760							737			
1.035 (B3.11) Werkkamer		10	542	120		24	66		6	451	106		15	136		7	505	106		37	105		3	465	101		7	161		2	237	463
1.037 (B3.10a) Stafkamer		84	549	103		18	6		136	473	91		11	7		115	510	96		9	4	29	194	468	68		2	6		180	233	343
1.039 (B3.10b) Stafkamer		89	591	57		22	2		139	472	94		2	6		93	486	87		66		28	112	452	65		89	19		14	241	488

Binnen één categorie kunnen meerdere regels/situaties vallen. Figuur 4.1.1

Normal situation

PIR light air temp co2

	Is functioning	MaxDeviation	MaxDeviationViolated	hasConstantValue	maxAcceptable	maxAcceptableViolated
co2 sensor	True	400	false	false	1400	false
Tempature	True				22	False
Airflow	True	50	false			
Lightstate	True					
PIRsensor	True					

Over het algemeen zijn er tussen de zes en tien situaties per categorie. Het nadeel van de resultaten is echter dat er niet kan worden afgeleid welke situatie er exact speelt in de ruimte op dat moment. Per maand zijn het aantal fouten per categorie bij elkaar opgeteld. Op basis van het hoogst aantal getelde 'fout categorieën' kan worden afgeleid of er iets aan de hand is en wat er aan de hand is. Dit zijn de volgende 'hoofdcategorieën':

- Normale situatie
- Lege ruimte
- (Te)Veel mensen aanwezig
- Temperatuur sensor is waarschijnlijk kapot
- Licht sensor is waarschijnlijk kapot
- Luchtventilatie sensor is waarschijnlijk kapot
- CO_2 sensor is waarschijnlijk kapot

In de bovenstaande categorieën is er een onderscheid te maken tussen 'goede' situaties en 'foute' situaties. In dit geval zijn de eerste twee situaties ('Normale situatie' en 'Lege ruimte') als 'goed' te bestempelen. Hierin kan worden aangenomen dat dit geen daadwerkelijke afwijking is op het normale patroon binnen een onderwijsinstelling. De overige categorieën zijn als afwijking te bestempelen, aangezien deze situaties niet voor mogen komen in verband met de juiste normen voor een comfortabel binnenklimaat.

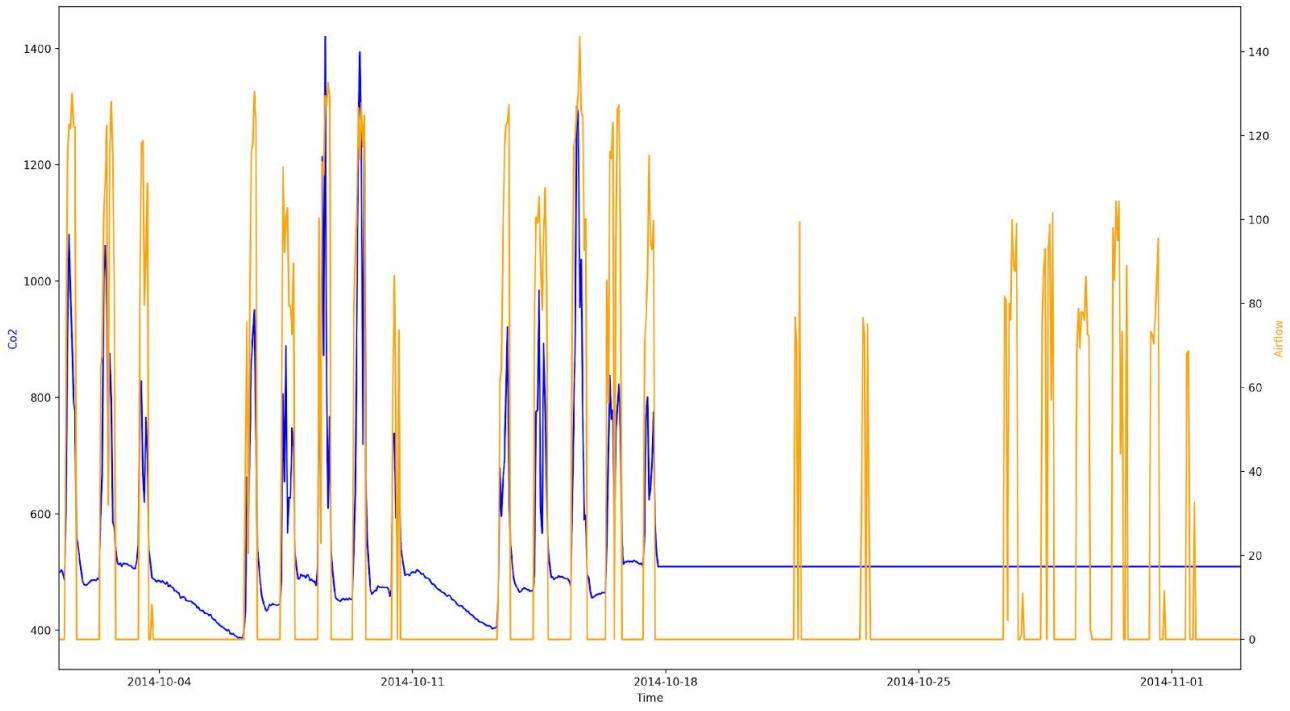
Om duidelijk te krijgen hoe het RBS in de reeds genoemde SAW applicatie zijn taken uitvoert is er onderzoek gedaan naar de werking van de applicatie. Zoals gebruikelijk en nodig voor een RBS werkt de applicatie op basis van een aantal vooraf bepaalde situaties. Deze situaties, of ook wel 'regels', zijn opgesteld door een expert op het gebied van het creëren van een juist binnenklimaat. De gemeten gegevens worden vervolgens door de applicatie geanalyseerd, per variabele, op of het voldoet aan de situatie welke vooraf zijn opgesteld. Wanneer de variabelen aan de situatie voldoen wordt dit opgeslagen en geregistreerd. De registratie verloopt per de hierboven genoemde hoofdcategorieën.

Vanuit het RBS systeem zijn er na analyse resultaten gevisualiseerd in een schematisch overzicht. Op basis van dit overzicht hebben wij het schema uit figuur 4.1.1 overzichtelijker gemaakt zodat eventuele anomalieën beter zichtbaar zijn. Daarnaast hebben wij alle situaties welke als een 'XML-format' in het RBS staan, in tabellen gezet, zodat er per situatie een beeld geschetst kan worden met de waardes die bij deze situatie passen. De tabellen van deze situaties zijn te vinden in bijlage 4.1 . Aan de hand van het verbeterde schema en de tabellen zijn vervolgens alle mogelijke anomalieën onderzocht en, na controle, genoteerd als anomalie indien dit het geval was (zie figuur 4.1.2, volledigheid bijlage 4.2).

Lokaal	Vanaf wanneer	Tot wanneer	Situatie	Verandering	Opmerkingen
0.005 (B4.01) Studielandschap	1-01-12	-	C02 sensor probably broken	T/m de periode 01-07-2013 - 01-08-2013 wordt er alleen de genoteerde situatie berekend. Hierna komen er "AIR sensor.." situaties bij.	Hier lijkt het goed mogelijk dat de sensor v vraag is tevens waar dit lokaal gesitueerd i deuren naar buiten e.d.?
0.014 (B3.03a) Praktijkruimte	1-01-12	1-08-2013	C02 sensor probably broken	Vanaf de omschakeling komen er meerdere andere situaties bij, waaronder normale situaties e.d.	
0.014 (B3.03b B3.04) Praktijkruimte	1-01-12	1-08-2013	C02 sensor probably broken	""	
0.031 (E07) Facilitair Kantoor	1-01-12	1-08-2013	C02 sensor probably broken	""	
0.033 (E05) Schoonmaak Kantoor	1-01-12	1-08-2013	C02 sensor probably broken	""	

Gevonden anomalieën, Figuur 4.1.2

Op basis van deze anomalieën is de rauwe data bekeken per anomalie. Hierin is bekeken of het hier gaat om een daadwerkelijke anomalie of dat het achteraf toch ruis bleek te zijn. Deze anomalieën (zie figuur 4.1.3) zijn vervolgens gevisualiseerd, zodat dit kan bijdragen aan het duidelijk maken waar een anomalie zich bevindt en tevens hoe het systeem zich in de loop er naartoe gedraagt. Dit kon niet alleen voor het eigen onderzoek naar de RBS-methode nuttig zijn, maar tevens voor de overige onderzoeken naar de andere twee methoden.



Figuur 4.1.3 Visualisatie CO_2 anomalie

Na het analyseren van de rauwe data zijn de anomalieën overgebleven welke zich in de periode van de data hebben voorgedaan. In deze periode zijn er in totaal 66 ruimtes, van de 119, waar zich defecten/afwijkingen voordoen. In het merendeel van de gevallen gaat dit om een langere periode van een aantal maanden, waarna de resultaten zich normaliseren en de

detectie van een anomalie verdwijnt. In 64 van de 66 gevallen gaat het om de situatie waarin de CO_2 sensor waarschijnlijk defect is.

Uit bovenstaande onderzoeken bleek dat het ontwikkelde en ontworpen RBS systeem niet langer voldeed aan de eisen, aangezien de applicatie niet werkend gekregen kon worden op eigen lokale omgevingen, noch op de lokale omgeving in Delft (de oude server). Doordat het systeem niet langer voldeed is door het onderzoeksteam gezocht naar een alternatief. Dit alternatief is gevonden in het ontwikkelen van een web-applicatie welke hierdoor dus vanaf elke plek bereikbaar zal zijn, maar tevens ook uitkomst zou bieden in het werkend krijgen van de huidige software.

Voor de ontwikkeling van de web-applicatie waren twee delen vereist. Dit betreft de back-end applicatie, waarin de gegevens worden opgehaald en geanalyseerd, en een frontend applicatie waar de gebruiker uiteindelijk gebruik van maakt via zijn/haar browser. Voor het ontwikkelen van de backend applicatie was het noodzakelijk om de oude SAW applicatie, welke geschreven was in C++, om te zetten naar een door ons beheerste taal en geschikt is voor het ontwikkelen van een web-applicatie. Hierin werd C# als uitkomst gevonden wat resulteerde in het omzetten van de C++ applicatie naar de nieuwe C# applicatie. De frontend applicatie is opgezet door middel van Angular, welke een Javascript fundering gebruikt samen met HTML, SASS en Typescript als webtalen voor de lay-out van de applicatie. (Figuur 4.1.4)

Naast de omzetting van de applicatie naar een nieuwe taal is er tevens zorg gedragen voor een kleine verbetering in de uitbreidbaarheid van de applicatie. Het nieuwe ontwerp heeft enkele voordelen welke in de huidige taal (C++) niet goed te verwezenlijken zijn. Dit leverde een compacter ontwerp op waarin enkel de Analyser klasse niet kon worden ontworpen. Dit laatste heeft te maken met het feit dat voor het omzetten van de volledige applicatie onfortuinlijk genoeg geen tijd meer beschikbaar was. Hierdoor zijn de onderdelen, welke nog niet zijn omgezet en waar niet voldoende informatie over is, nog niet verder uitgewerkt in het ontwerp. (Bijlage 4.3)

Mockup's web-applicatie 'Dashboard', Figuur 4.1.4

4.2. Bayesian Belief Network

4.2.1 Analyse in een C++ applicatie

Met voorgaande onderzoeken is geprobeerd BBN te gebruiken. Het resultaat hiervan is een applicatie waarin een BBN model geïmporteerd en geanalyseerd kan worden. Vanuit de opdrachtgever waren echter nog een aantal wensen die mogelijk tijdens dit project doorgevoerd konden worden:

1. Het programma accepteert alleen komma gesepareerde bestanden;
2. De laatste regel data wordt niet meegenomen in de analyse;
3. De database is niet direct te benaderen;
4. Het crasht erg veel zonder duidelijke foutmelding (error afhandeling);
5. De resultaten moeten overzichtelijker weergeven worden (anomalieën duidelijker laten uitspringen);
6. De analyse moet geautomatiseerd worden.

Vanwege complexiteit en tijd zijn alleen de eerste drie punten geïmplementeerd.

4.2.2 Meerdere delimiters accepteren

Het accepteren van verschillende soorten separatoren is een redelijk simpele oplossing waarmee direct de gebruiksvriendelijkheid verhoogt wordt. Afhankelijk van een aantal instellingen kan het zijn dat verschillende mensen, verschillende separatoren gebruiken. Een duidelijk voorbeeld is de USA/UK variant, dat gebruik maakt van de komma als scheidingsteken en een punt als decimaalteken. Waartegen het Europese format de puntkomma als scheidingsteken en de komma als decimaalteken gebruikt.

De eerste oplossing was het gebruik van een regularp expression, deze split de data op elk voorkomende separator (komma, puntkomma of tab). Het probleem wat naar voren kwam, was dat de door komma gescheiden getallen ook gesplitst werden (dus bij bestanden volgens het Europese format). Om dit te voorkomen wordt op de eerste regel van een bestand gezocht naar het eerst voorkomende separator (komma, puntkomma of tab). Deze wordt dan als delimiter voor de overige regels gebruikt.

Het viel op dat na analyse in de ‘smileApp’, de laatste regel niet terugkwam in het resultaat. Om de data compleet te houden, is begonnen met het analyseren van de code. Daardoor is in de functie waar een CSV-bestand wordt ingelezen, een lijn code ontdekt dat altijd de laatst ingelezen regel verwijderd. De reden achter dit stuk code is, jammer genoeg, niet bekend.

4.2.4 De database is niet direct te benaderen

Op dit moment kunnen alleen CSV bestanden geïmporteerd worden in de ‘smileApp’, dus men moet nu via een omweg data van de database analyseren. Verder is een ‘directe’ connectie’ nodig voor het automatiseren; door de verbinding kan de ‘smileApp’ waardes opvragen bij een database, zonder dit via een omweg te moeten doen.
Er zijn twee databases waar sensorwaardes van De Haagse Hogeschool te Delft worden opgeslagen, namelijk een oudere dat beheerd werd door MonaVisa en een nieuwere dat beheerd wordt door Priva. In de oude database is data opgeslagen tot april 2016, de nieuwe start dus ongeveer waar de vorige database ophoudt.

Omdat de nieuwe database pas bruikbaar was tegen het einde van de minor, is de focus gelegd op de oude database. Het voordeel hiervan was dat in de SAW applicatie (rule

based), al een connectie was gemaakt met de oude database. In principe was het dus een kwestie van de juiste functies samenvoegen.

Het eerste probleem dat naar voren kwam, was dat een aantal sensoren niet goed werden opgehaald. Dit gebeurde ook in de SAW applicatie. Voor sensoren wordt gezocht naar het fieldtype en een deel van de omschrijving. Voor CO_2 was de omschrijving niet kloppend, en voor de temperatuur (5) was de fieldtype van airflow (48) ingevuld.
Een ander probleem was dat de (originele) data niet meer geëxporteerd kon worden nadat het was opgehaald uit de database, dit was voorheen wel een optie. Dit is opgelost door een aparte functie te maken voor het opslaan van data dat uit een database komt.
De opgehaalde data moet geanalyseerd worden, om dat te kunnen doen moeten eerst de waardes omgezet worden naar getallen die horen bij de symptomen van het BBN model. Het kan bijvoorbeeld dat het symptoom CO_2 als opties laag, middelmatig of hoog heeft en symptoom airflow laag of hoog. De originele waardes moeten dan geclassificeerd worden in respectievelijk nul, één of twee voor CO_2 en nul of twee voor airflow. Deze 'geconverteerde' metingen worden ook wel 'rule values' genoemd.
In het huidige BBN model is het anders opgedeeld, hier zijn bijvoorbeeld hoge en lage CO_2 aparte symptomen en krijgen dus allebei de waarde 0 (normale situatie) of 1 (hoge/lage CO_2), afhankelijk van de limieten. Dit zijn een paar voorbeelden van vier vorige grenswaardes:

- Lage CO_2 = 418
- Hoge CO_2 = 630
- Lage airflow = 7
- Hoge airflow = 70

Bovenstaande waardes zijn bepaald door middel van de 25e percentiel en 75e percentiel op de dataset van lokaal 2.008: 2014-2015 (per uur).
De hoge en lage CO_2 blijven hetzelfde voor elk lokaal, maar de airflow grenzen niet. Verschillende lokalen hebben namelijk verschillende waardes voor hoge en lage airflow. Op dit moment worden ze bepaalt door 75% en 10% te nemen van de hoogst voorkomende airflow, nadat de grootste outliers zijn verwijderd. Voordat dit gedaan wordt, wordt data tussen 7 en 23 uur genomen omdat hiertussen het gebouw 'actief' is. De hoge en lage airflow van lokaal 2.008 veranderen hierdoor van 70 en 7, naar 83 en 11. Alle overige waardes zijn te vinden in bijlage 4.4.

Nadat alle relevante waardes zijn omgezet, wordt door middel van een pop-up venster gevraagd of de nieuwe waardes opgeslagen moeten worden. Dit is de tweede mogelijkheid tot het exporteren van de data. De laatste is nadat de analyse gedaan is, zodat het resultaat ook opgeslagen kan worden.

4.2.5 Testen van het model

Nadat bovenstaande handelingen verwerkt zijn, moeten het model en de applicatie getoetst worden. Hierbij is gebruik gemaakt van testdata waarbij duidelijk te zien is dat een sensor defect is. Hiervoor is tussen 28 april 2015 6:00 en 1 juni 2015 23:00, data van de Airflow CO_2 en PIR sensor opgehaald uit de oude database via de applicatie. Onderstaande tabel is een deel van die data.

Datum en tijd (uur)	CO_2	PIR	Airflow
30-4-2015 08:00	448,157	0	17,309
30-4-2015 09:00	499	0	4,47758

30-4-2015 10:00	863	0	4,87582
30-4-2015 11:00	924,995	0	5,20573
30-4-2015 12:00	1233,67	0	7,89801
30-4-2015 13:00	733,654	0	9,549
30-4-2015 14:00	847	0	8,39731
30-4-2015 15:00	794,167	0	11,5692
30-4-2015 16:00	1427,59	0	5,893
30-4-2015 17:00	702,1	0	4,88038
30-4-2015 18:00	519,425	0	4,86786
30-4-2015 19:00	488,161	0	7,96258
30-4-2015 20:00	477	0	14,1579

Het is duidelijk te zien dat de PIR-sensor kapot is, omdat de CO_2 waarde erg fluctueert en dit duidt op aanwezigheid. Airflow gaat meestal aan als er beweging wordt gedetecteerd, daarom zijn desbetreffende waardes niet zoals verwacht. De data moet worden omgezet naar ‘rule values’ voor analyse, in de tabel in bijlage 4.5 is bovenstaande data geconverteerd

Een ‘1’ betekent een normale situatie, voor ‘0’ is het symptoom (kolomnaam) actief. Alle kolommen zien eruit zoals verwacht; er is geen “High_flow” (grenswaarde van 83 m^3/h), maar “Low_flow” (grenswaarde van 11 m^3/h), “High_CO2” komt een aantal keer voor (grenswaarde van 630 ppm) en er is “Flow_while_PIR_0” om maar een paar voorbeelden te noemen. Het resultaat ziet er als volgt uit:

Datum en tijd (uur)	Air_flow_sensor	CO2_sensor	Damper	Fan	Occupancy_not_too_high	PIR_sensor
30-4-2015 08:00	1	1	0,944488	1	0,999722	1
30-4-2015 09:00	1	1	0	1	0,999722	1
30-4-2015 10:00	1	1	0,233167	0,710572	0,990786	1
30-4-2015 11:00	1	1	0,233167	0,710572	0,990786	1

30-4-2015 12:00	1	1	0,233167	0,710572	0,990786	1
30-4-2015 13:00	1	1	0,233167	0,710572	0,990786	1
30-4-2015 14:00	1	1	0,233167	0,710572	0,990786	1
30-4-2015 15:00	1	1	0,09575	0,982583	0,910067	1
30-4-2015 16:00	1	1	0,233167	0,710572	0,990786	1
30-4-2015 17:00	1	1	0,233167	0,710572	0,990786	1
30-4-2015 18:00	1	1	0	1	0,999722	1
30-4-2015 19:00	1	1	0	1	0,999722	1
30-4-2015 20:00	1	1	0,944488	1	0,999722	1

In de tabel zijn de kansen weergegeven op het **NIET** gebeuren van het resultaat (kolomnaam). Wat opvalt is dat het model aangeeft dat niet de PIR-sensor kapot is, maar juist de Damper (‘luchtklep’). Dit komt waarschijnlijk doordat in het BBN model, de Damper verbonden is met “Low_flow”, “Flow_while_PIR_0” en “High_CO2”. Dit toont drie van de vijf symptomen aan, en dus wordt de Damper als kapot wordt beschouwd.
Het merkwaardige is echter dat twee van de vier symptomen die verbonden zijn met de PIR, in hetzelfde rijtje staan (namelijk Low_flow” en “High_CO2”).
Vervolgens is onderzoek gedaan naar wanneer de PIR volgens dit model kapot is. De mogelijke combinaties van symptomen die verbonden zijn met de PIR worden als input gebruikt, dit is te zien in bijlage 4.6.

Het volgende resultaat (kans op **NIET** gebeurd/kapot) na analyse met de smileApp, is te vinden in bijlage 4.7.

Wat opvalt is dat het patroon van “PIR_NaN” gevolgd wordt. Dit is een juiste indicatie dat de sensor kapot is, maar de andere symptomen lijken geen invloed te hebben. Dit is dus iets wat bijgewerkt moet worden in het BBN model.

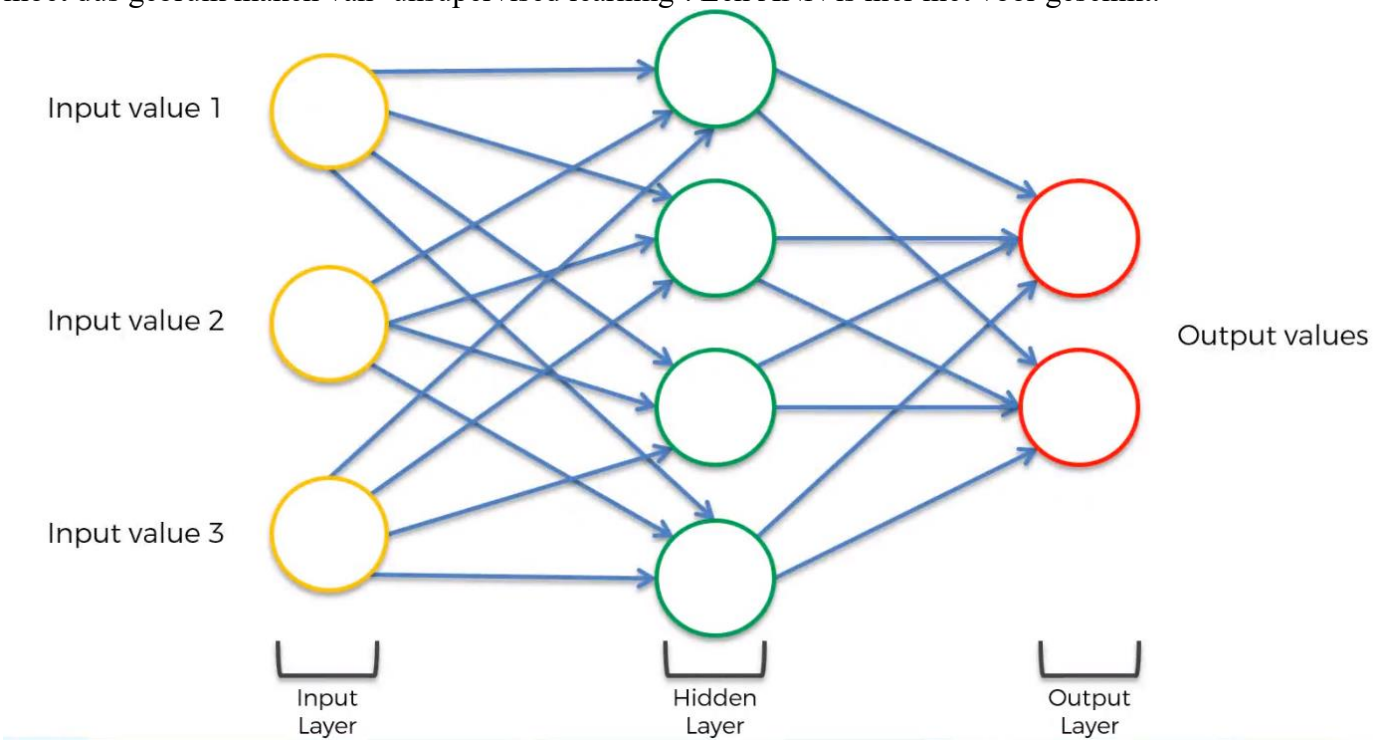
<< Stuk van Andre >>
Instructies;
Per kopje kies je voor stijl Kop3
Na afmaken van stuk, zorg ervoor dat de pagina's mooi uitkomen.
Als je kopjes hebt gebruikt moet je de inhoudsopageve updaten.
Nadat je dit gedaan hebt maak je de Kop 1 hoofstukken, de grote koppen dus, groen met de groene kleur die er al in dit bestand.

4.3. Deep Learning

Deze paragraaf zal de resultaten van het onderzoek naar Deep Learning weergeven. Allereerst is onderzoek gedaan naar verschillende types neurale netwerken, de uitkomsten hiervan kunt u vinden in paragraaf 4.3.1. Nadat een geschikt type netwerk gevonden is, zijn deze geprogrammeerd en op de dataset getraind. De resultaten daarvan zijn te vinden in paragraaf 4.3.2.

4.3.1. Neurale netwerken

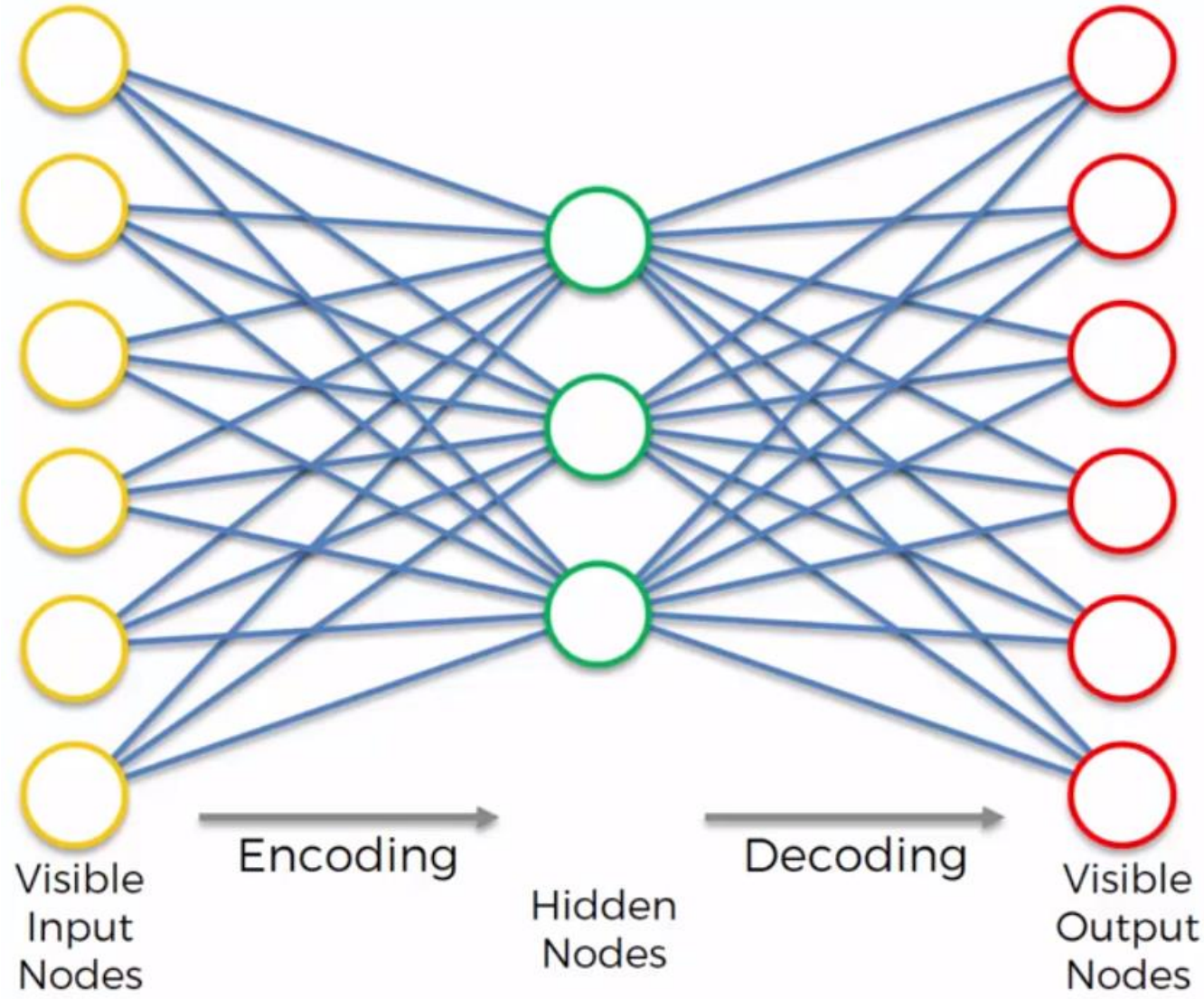
Door neuronen op een bepaalde manier te schakelen kunnen verschillende types netwerken ontstaan. Deze hebben ieder bepaalde eigenschappen die wel of niet toepasbaar zijn op een bepaald probleem. Allereerst is er een simpel artificieel neuraal netwerk (ANN) zoals te zien op figuur 4.3.1. Om dit netwerk te trainen moet een dataset ingeladen worden met ook de output waarde daarbij. Het gaat dan om een zogenaamd Supervised Learning probleem omdat de data gelabeld is. Een voorbeeld daarvan is dat er van een bepaalde ruimte de temperatuur, CO_2-waarde en de airflow bekend zijn. In de output staat of het wél of juist geen anomalie is. Het netwerk kan bepaalde gewichten aan de neuronen geven zodat deze in de toekomst, bij nieuwe data, een voorspelling kan doen. Bij een ANN is het verplicht om gelabelde data te hebben. De data die door de sensoren wordt opgeleverd is niet gelabeld. Het is dus niet bekend of data geclassificeerd kan worden als anomalie of niet. Het type neuraal netwerk moet dus gebruik maken van ‘unsupervised learning’. Een ANN is hier niet voor geschikt.



Figuur 4.3.1 Artificial Neural Network (ANN)

Een van de oplossingen op het gebied van ‘Unsupervised Learning’ is een zogenaamde ‘Auto Encoder (AE)’, te zien in figuur 4.3.2. Doordat de hidden layers minder neuronen bevatten dan inputs wordt de data ‘gecomprimeerd’. Vervolgens worden de outputs bepaald vanuit het gecomprimeerde model, waarbij informatie verloren gaat. Normaal gesproken is dat geen gewenst gedrag, maar in dit geval wordt gezocht naar een gegeneraliseerd model. Het gegeneraliseerde model kan vervolgens vergeleken worden met nieuwe data. Wanneer een sterk verschil tussen de nieuwe data en het model optreed, zou dat kunnen duiden op een anomalie. Het nadeel van een ‘AE’ is dat dit netwerk geen rekening houdt met wat er in het

verleden gebeurt is. Dit houdt in dat het netwerk eenmalig een model genereert door middel van de gegeven data. Omdat de data van de sensorwaardes seizoensgebonden zijn, is het model niet bruikbaar wanneer het hier geen rekening mee houdt.

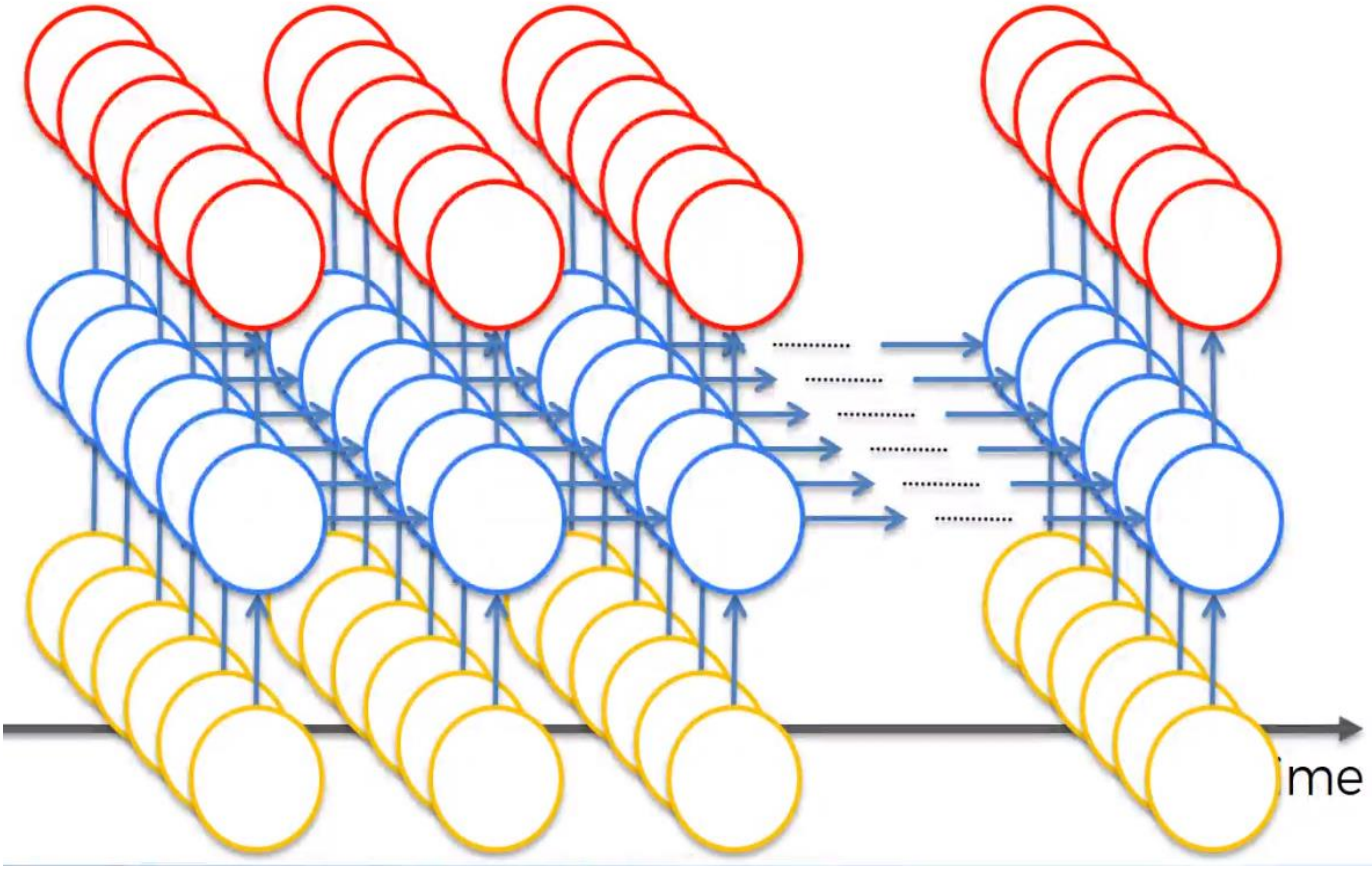


Figuur 4.3.2 Auto Encoder (AE)

Een oplossing hiervoor is een zogenaamd Recurrent Neural Network (RNN), zoals te zien in figuur 4.3.3. Een ‘RNN’ heeft behalve de standaard outputs, ook een output die als terugkoppeling gebruikt wordt voor de input. Op deze manier kan een tijdserie van data per tijdsinterval geclassificeerd worden. Hierdoor wordt duidelijk hoe verbanden tussen sensoren zich in de verloop van tijd gedragen.

Echter hebben ‘RNN's’ last van een zogenaamd ‘Vanishing Gradient Problem’(Hochreiter, 1997). Dit houdt in dat als data verder in het verleden ligt, het minder meegenomen wordt in de voorspelling. Ook duurt het trainen van een ‘RNN’ exponentieel langer bij elk nieuw tijdsinterval, omdat de eerste lagen van het ‘RNN’ nooit tot een optimum kunnen komen (Hochreiter,1997).

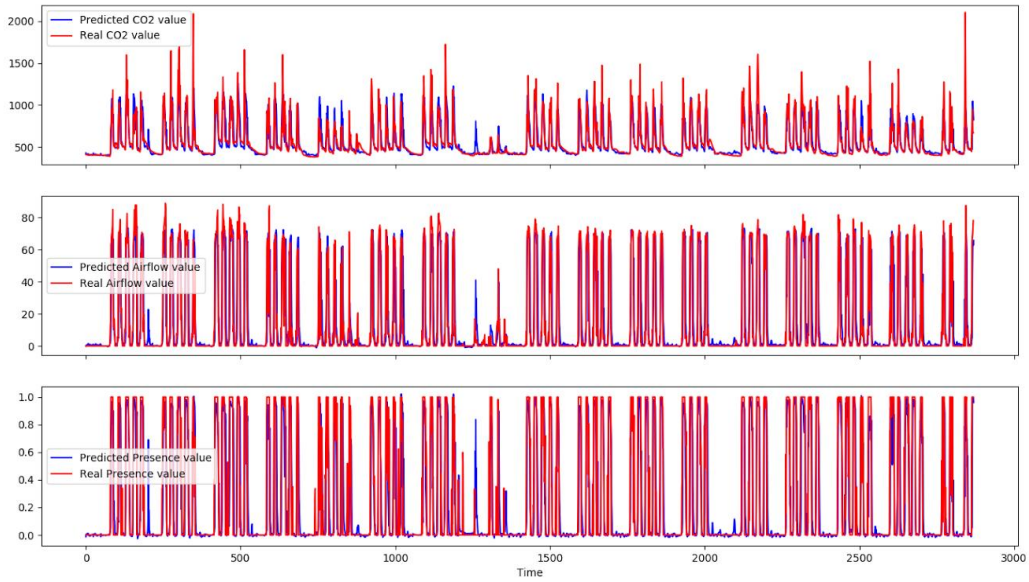
Een ‘Long Short Term Memory’, dat in de volgende paragraaf beschreven wordt, heeft hier geen last van(Hochreiter,1997).



Figuur 4.3.3 Recurrent Neural Network (RNN)

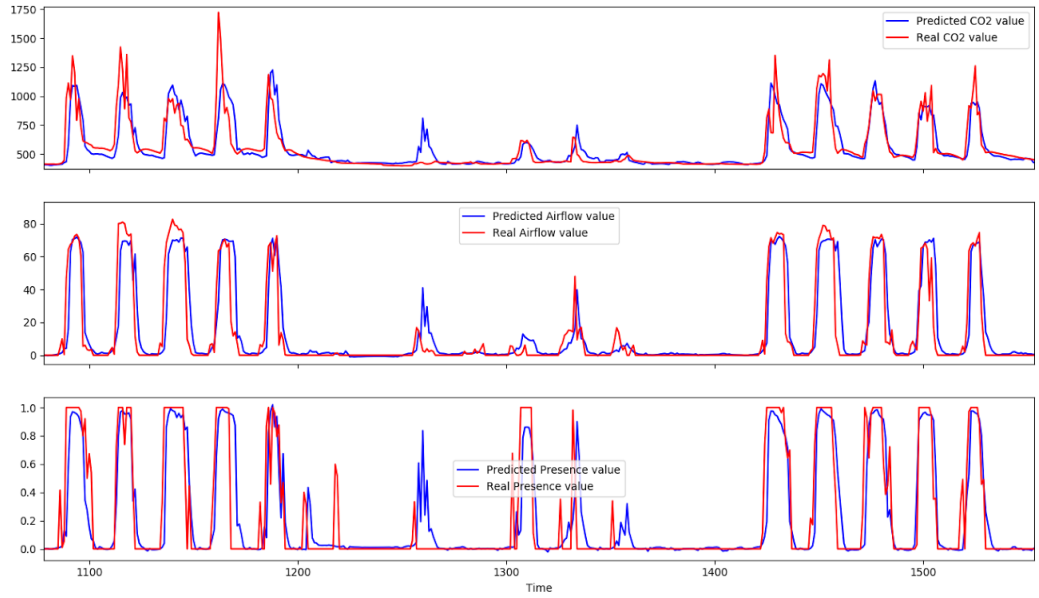
4.3.2. LSTM

Een ‘Long Short Term Memory (LSTM) heeft geen last van het ‘Vanishing Gradient Problem’ doordat het gebruik maakt van zogenaamde ‘Memory Cells’. De ‘LSTM’ die gemaakt is, bestaat uit 4 hidden layers, met 50 neuronen per layer. De input die gebruikt is, is alle sensordata over een periode van 24 uur in het verleden. Met deze data wordt het eerstvolgende uur voorspeld. Als dit elke keer een tijdstip later gedaan wordt, komt er een voorspelling voor een aantal weken, in dit geval 18, zoals te zien is in figuur 4.3.4.



Figuur 4.3.4 Voorspelling met LSTM

De periode tussen 1050 en 1550 is de grafiek zoals te zien is in figuur 4.3.5. In dit figuur zien we een normale schoolweek, een vakantie van een week en vervolgens weer een schoolweek. Zoals te zien is wordt er op de maandag in de vakantieweek aanwezigheid gedetecteerd waarna het 'LSTM' voorspelt dat er weer een normale schooldag aankomt. De voorspelling gaat hierbij fout omdat er over een periode van 24 uur is getraind.



Figuur 4.3.5. Twee normale schoolweken met hiertussen een vakantie

Om het model beter te maken is vervolgens een window-size genomen van 168 uur, zodat het 'LSTM' een week als periode ziet. Elke week wordt opnieuw voorspelt met data van vorige weken. Hierdoor worden weekenden ook goed herkend.

Als er een gegeneraliseerd model is kan vervolgens gekeken worden in hoeverre de werkelijke waardes afwijken van de voorspelde waardes. Als de afwijking dermate groot is kan worden aangenomen dat er een anomalie is. Op dit moment geen concrete anomalieën gedetecteerd zijn met de getoetste 'deep learning' methodes. Het 'LSTM' netwerk heeft van alle neurale netwerken, de beste potentie getoond om anomalieën te detecteren in een tijdserie dataset.

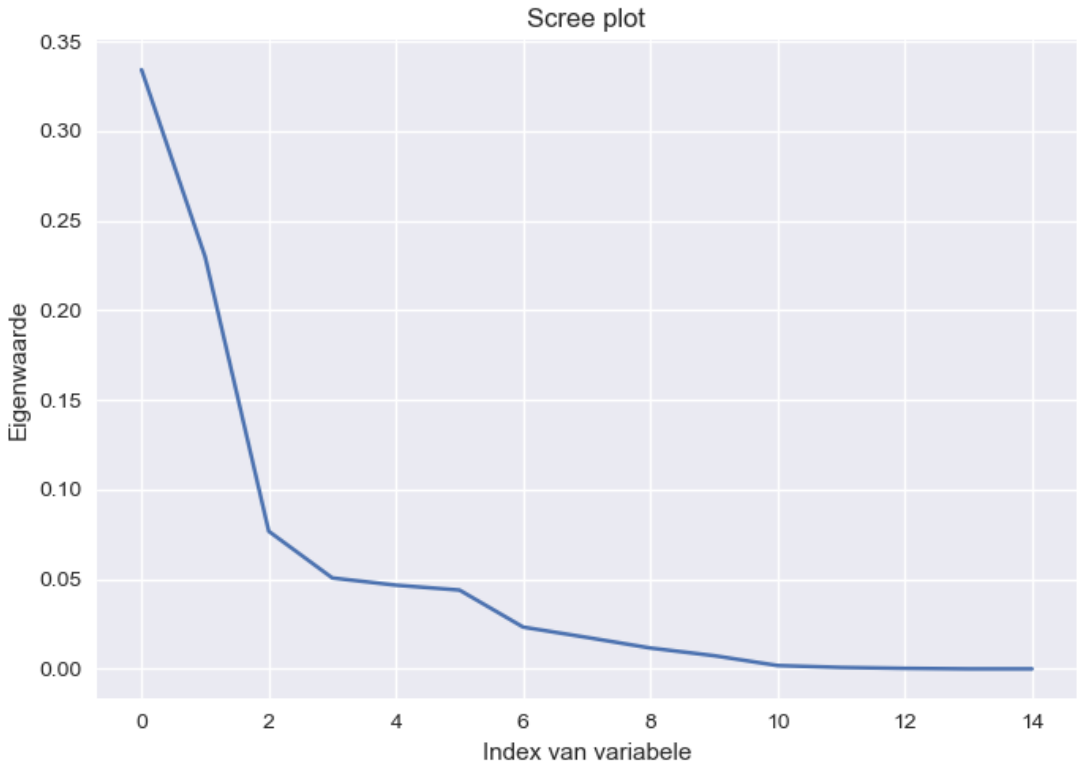
4.4. Clusteranalyse

In deze paragraaf, zijn de verkregen resultaten van het onderzoek, naar clustermethoden beproken. In eerste instantie staat de werking van de factoranalyse beschreven. Hierna wordt het onderzoeksproces van de factoranalyse, samen met de bijbehorende resultaten daarvan besproken. Ten slotte, zijn in het onderzoek verschillende cluster-herkenningsmethoden, beschreven met het bijbehorend resultaat.

4.4.1 Factoranalyse

Voor het clusteren is gekozen om de data om te zetten door middel van factoranalyse. In het kort houdt de factoranalyse in dat veel verschillende variabelen/kolommen worden samengevoegd tot een nieuwe variabele. Denk bijvoorbeeld aan aanwezigheid. Op basis van een combinatie tussen meerdere variabelen kan beweert worden dat mensen aanwezig zijn, zoals met een combinatie tussen CO_2 gehalte en energie verbruik. Dit komt voor het project goed van pas, sinds van bepaalde ruimten binnen de HHS tientallen variabelen worden bijgehouden wat niet duidelijk geclusterd kan worden. In hoofdlijnen werkt de factoranalyse als volgt:

- Bereken het *n* aantal factoren waar de dataset in gesplitst kan worden. Gedurende het project is gebruik gemaakt van '*scree plots*'. Hiermee worden de eigenwaarden van de dataset berekend en weergegeven zoals is afgebeeld in figuur 4.4.1. In dergelijke figuren zijn buigpunten aanwezig die een indicatie geven voor hoeveel factoren gebruikt moeten worden. De gemaakte analyses zijn op basis van $n = 5$, mits anders vermeld.
- Factor analyse berekent de maximale waarschijnlijkheid uit voor de zogeheten 'gewichten' matrix. Dit is de transformatie van de onzichtbare variabelen naar de waargenomen variabelen, door middel van 'expectation-maximization' (Barber, 2012). Voor dit project is de ingebouwde factor analyse vanuit python gebruikt.
- De resultaten van de factoranalyse zijn in figuur **4.4.2** te zien in de vorm van een matrix (*n x m*) met daarin de gewichten van elk variabele op elk gevonden factor. Deze matrix moet vervolgens worden vermenigvuldigd met elke rij aan data zodat voor elke rij uit de dataframe, de factorwaarde per factor wordt berekend. Daar komt vervolgens een dataframe zoals die in figuur **4.4.3** is te vinden.



*Figuur **4.4.1**: Voorbeeld van een 'Screeplot'*

	bedrijfsstatus	temperature0	temperature1	energysupplycool	energysupplyheat	measairflow	lampenergy	objecttemp	waterflow	ambienttemp
0	0.658164	0.899263	0.855051	1.285067	1.052003	-0.960220	0.952178	0.797765	-0.960220	1.282541
1	1.307428	-0.701642	-0.529106	-1.133841	-1.085076	-0.311685	-1.067076	-0.656289	-0.311685	-0.914267
2	-0.355574	-1.165285	-1.319538	1.765645	1.582292	-0.238114	1.808512	-0.674953	-0.238114	-0.691141
3	1.622666	0.060455	-0.069398	-1.548718	-1.334981	-0.291846	-1.615808	0.804418	-0.291846	0.935943
4	2.517507	-0.617367	-0.231135	1.112940	0.771766	-0.304410	1.056131	-0.644765	-0.304410	-1.060661

*Figuur **4.4.2**: Voorbeeld uitkomst factoranalyse. Hierin is elke rij een factor en de bijbehorende waarden zijn de gewichten.*

	0	1	2	3	4
0					
2013-08-20 12:35:00	4.745090	1.051471	-1.045883	6.161641	1.349352
2013-08-20 12:40:00	4.750289	1.046577	-1.047357	6.174706	1.344032
2013-08-20 12:45:00	4.823112	1.182509	-1.087425	6.349219	1.609078
2013-08-20 12:50:00	4.986716	1.468739	-1.176468	6.730367	2.170196
2013-08-20 12:55:00	5.018972	1.443747	-1.201426	6.756145	2.145089
2013-08-20 13:00:00	5.048003	1.422104	-1.225699	6.774662	2.123369
2013-08-20 13:05:00	5.062278	1.412218	-1.228337	6.791924	2.109433
2013-08-20 13:10:00	5.483013	1.555780	-1.009933	5.409152	-0.075384
2013-08-20 13:15:00	5.215613	1.438620	-1.179967	6.383828	1.425379
2013-08-20 13:20:00	5.111551	1.378507	-1.263099	6.821138	2.071580
2013-08-20 13:25:00	4.836496	0.396007	-1.117932	7.226929	2.135158
2013-08-20 13:30:00	4.618137	-0.337109	-0.994789	7.529925	2.192536

*Figuur **4.4.3**: Door middel van factoranalyse getransformeerde data.*

4.4.2 Beginfase

Als uitgangspunt is een jaar aan data gebruikt van een ruimte waarvan bekend was dat tussen 20 april en 1 juni 2015 de aanwezigheidssensor defect is geweest, D2.008. Dit is een kleine spreekkamer, zonder ramen en voor twee á drie personen bedoeld. Hierdoor kan het CO2 niveau snel stijgen in de ruimte. In combinatie met een defecte aanwezigheidssensor zorgt dit voor een onaangenaam klimaat.
Voor de eerste poging van factoranalyse is de dataset, zoals die in figuur **4.4.4** is te zien, gebruikt. Dit leverde echter niet het gezochte resultaat op, namelijk een cluster van de defecte aanwezigheidssensor. Het bleek te komen doordat verscheidene variabelen op verschillende schalen zijn bijgehouden, waardoor variabelen die in miljoentallen werden bijgehouden een hogere invloed kregen dan de variabelen in bijvoorbeeld tientallen. Aan de hand hiervan is gekozen om Min/Max normalisatie toe te passen op alle variabelen, zodat alle waarden tussen nul en één liggen. Dit leverde logischere resultaten op, want niet elke gevonden factor werd door dezelfde variabelen het meest beïnvloedt. De variabelen waren echter niet klaar om gebruikt te worden. Een aantal variabelen, vooral energieverbruik, zijn cumulatief bijgehouden, waardoor deze op de lange termijn meer de rol van tijd speelt dan een waargenomen waarde. Om dit tegen te gaan is voorgaande het normaliseren gekozen om de cumulatieve waarden om te zetten in absolute waarden. Dit leverde betere resultaten op doordat tijdsverloop nu geen invloed speelde op de factoren. Deze twee stappen zijn bij alle hierna te volgen factoranalyses toegepast.

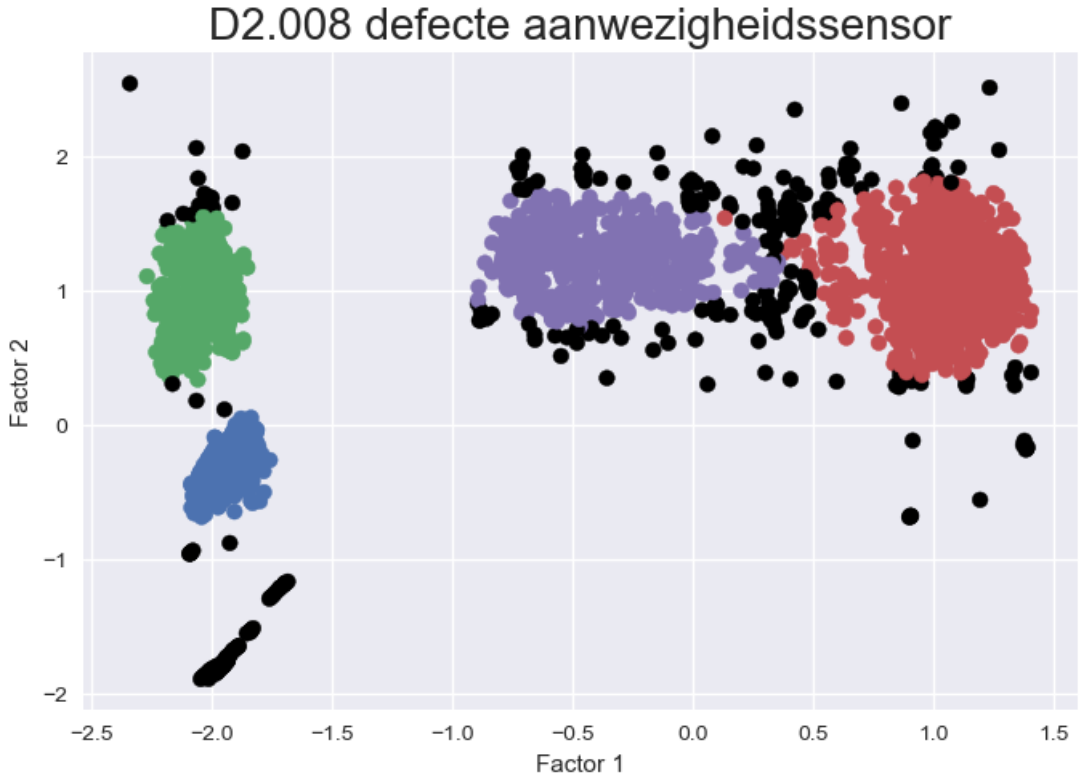
0	8265 IrtempObjectTemp OG097D	8266 IrtempAmbientTemp OG097D	10250 LampEnergy OG138D	16548 CO2 Level OG0B5C	17232 LampEnergy OG12AA	17785 Valve Actual Position OG0830	18696 LampEnergy OG11D9	19447 Valve Actual Position OG0872	19517 EnergySupplyHeat OG0872
2013-10-05 09:00:00	21.908750	22.190000	1.035165e+09	507.000000	1.319241e+09	0.000000	1.330308e+09	90.000000	1.557149e+08
2013-10-05 09:05:00	21.935714	22.190000	1.035175e+09	511.071429	1.319252e+09	0.000000	1.330319e+09	90.000000	1.557149e+08
2013-10-05 09:10:00	21.986000	22.192667	1.035192e+09	515.200000	1.319270e+09	0.000000	1.330337e+09	90.000000	1.557149e+08

*Figuur **4.4.4**: Voorbeeld van de gebruikte data zoals het eerst werd gebruikt.*

Met behulp van de hiervoor genoemde transformaties is weer de factoranalyse toegepast, maar dit leverde opnieuw niet de gezochte resultaten. Bij nader inzien kwam dit door de grote hoeveelheid combinaties tussen sensoren die aanwezigheid aantonen. Hierdoor kregen bepaalde variabelen zoals ‘Light dim state’ (mate van licht toevoer) een sterke invloed op een factor terwijl dit in de praktijk geen invloed heeft op het klimaatregelsysteem. Daarom zijn voor de factoranalyse niet alle beschikbare variabelen gebruikt, maar slechts de variabelen die de groep als belangrijk ziet. De gebruikte variabelen luiden als volgt:

- Bedrijfsstatus (Waarden tussen 1 en 4 die aangeven in hoeverre het klimaatregelsysteem aan staat.
- Temperature 0 (Temperatuur van binnenkomend water in de vloer)
- Temperature 1 (Temperatuur van naar buiten gaand water in de vloer)
- Energy supply cool (Verbruikt energie om water af te koelen)
- Energy supply heat (Verbruikt energie om water te verwarmen)
- Meas airflow (hoeveelheid luchttoevoer in m^3 per uur)
- Lamp energy (Verbruikt energy van lichten)
- Object temp (Gemeten wand temperatuur)
- Ambient temp (Gemeten lucht temperatuur)
- Water flow (Een 1 of een 0 wat respectievelijk aangeeft of de watertoevoer aanstaat of niet)
- Presence (Een 1 of een 0 wat respectievelijk aangeeft of aanwezigheid is gemeten of niet)
- CO_2 (Het CO_2 gehalte)
- Air flow pressure difference (Het verschil tussen luchthoeveelheid start ventilatie en binnenkomst ruimte)
- Actual airflow (Gemeten hoeveelheid luchttoevoer in m^3 per uur)
- Valve actual position (Waarde tussen 0 en 100 die de mate van openheid voor kleppen aangeeft)

Op basis van deze variabelen zijn de eerste bruikbare resultaten geproduceerd waarin duidelijk te zien was dat de aanwezigheidssensor defect was. In figuur **4.4.5** is een ‘scatterplot’ afgebeeld waarin dit zichtbaar is. Het gaat hier om de blauwe cluster aan de linkerhand in de figuur. Deze ‘scatterplot’ bevat echter alleen data waarbij de CO_2 niveau boven een bepaalde grens uitkomt, want als alle datapunten worden gebruikt zijn de clusters niet duidelijk zichtbaar. Dit geldt ook voor alle andere scatterplots, opnieuw met de bedoeling om alles duidelijk zichtbaar te houden.



*Figuur **4.4.5**: Scatterplot tussen twee factoren waarmee de defecte aanwezigheidssensor duidelijk zichtbaar is in het blauw.*

In de bovenstaande figuur zijn twee factoren tegenover elkaar afgebeeld waartussen een aantal clusters zijn gevonden. Deze clusters zijn gevonden door middel van de 'Hdbscan' algoritme wat in paragraaf 4.4.5 verder staat uitgelegd. Van deze twee factoren is niet bekend welke variabelen daar stekt invloed op hebben, dus naar ons weten zijn het twee vrijwel willekeurig gemaakte factoren die samen toch een situatie kunnen schetsen waarin het klimaatregelsysteem niet naar wens werkte.

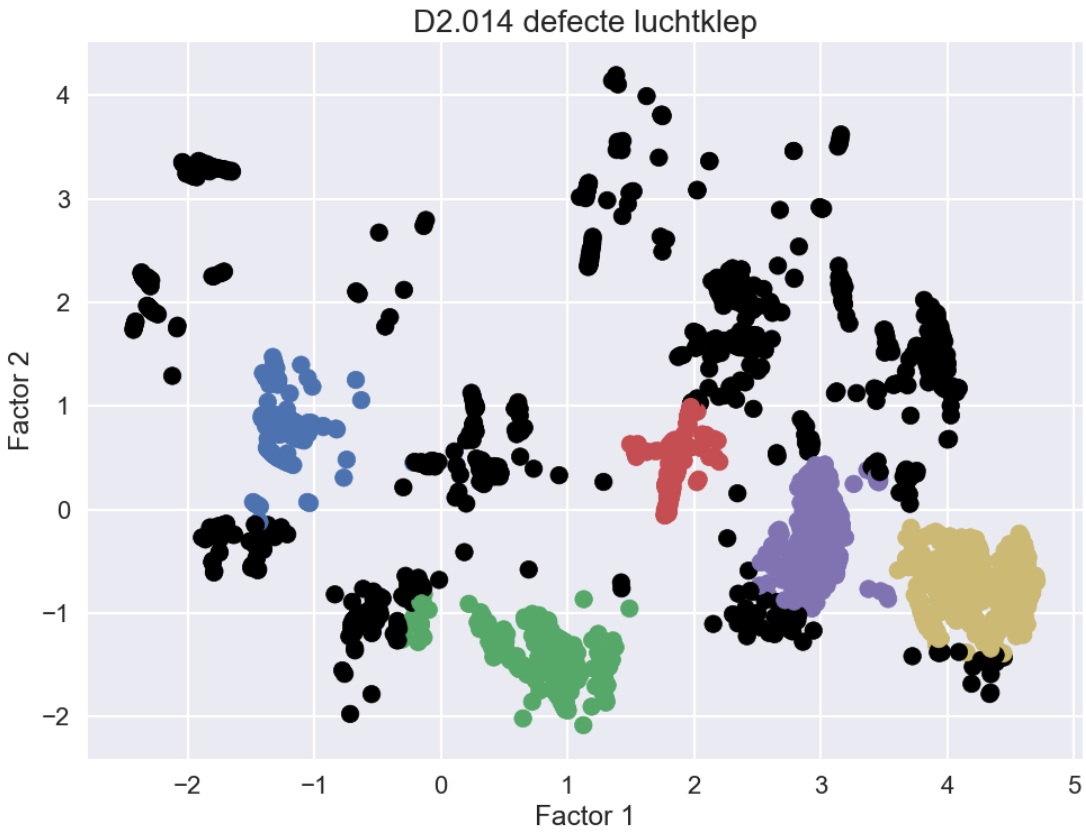
4.4.3 Doorontwikkeling factoranalyse

Aan de hand van het voorbeeld in de voorgaande paragraaf blijkt dat clusteren mogelijk is, maar dit is slechts één voorbeeld waarin een dergelijke situatie is gevonden. Om te controleren of dit bruikbaar is voor meerdere situaties is een tweede dataset van een andere ruimte getest, D2.014. Dit is een wat grotere ruimte waarvan bekend is dat een luchtklep gedurende een aantal maanden, in het kalenderjaar 2014, niet functioneerde waardoor de ruimte niet geventileerd werd. Voor deze controle is opnieuw één jaar aan data gebruikt.

Om mee te beginnen is de dataset door alle hiervoor genoemde stappen gehaald, dus zowel het voorbereidend werk als het selecteren en normaliseren van de data. Op basis van de voorgaande paragraaf zou dit moeten werken, maar dit viel tegen. Zoals eerder vermeld is deze ruimte wat groter dan de eerste, dat houdt in dat soms meerdere sensoren van één bepaald type aanwezig zijn. Denk hier bijvoorbeeld aan de 'lamp energy' sensoren die per lamp in de ruimte wordt bijgehouden welke ten opzichte van de eerste ruimte één keer werd bijgehouden, maar in de nieuwe ruimte vier keer. Dit zorgt ervoor dat een aantal gevonden factoren sterk overeenkomen met elkaar, maar dat ze elk door een andere sensor van dezelfde type worden beïnvloedt.

Dit probleem is opgelost door het gemiddelde van alle sensoren van dezelfde type te nemen zodat elk type sensor slechts éénmaal voorkomt. Met de data in een bruikbaar format is de factoranalyse toegepast met als doel het vinden van de defecte luchtklep. Dit leverde nog niet het gezochte defect op, daarom is uitgebreid gekeken naar de sensor van de luchtklep, 'valve actual position'. Hieruit bleek dat niet alleen luchtkleppen onder deze type sensor werden opgeslagen, maar ook de waterkleppen. Dit betekent dat de gebruikte 'valve actual position' een gemiddelde was van twee verschillende soorten kleppen, waardoor de defecte luchtklep onopgemerkt bleef. Daarom is gekozen om niet het gemiddelde te nemen van de 'valve actual position' sensoren, maar deze apart te houden. Dit leverde opnieuw het hiervoor behandelde probleem van meerdere sensoren van dezelfde type, maar dat is op een dusdanig klein niveau dat hier toch voor is gekozen. Dit is een alternatief op de ideale oplossing, namelijk achterhalen welke 'valve actual position' de luchtklep is. Helaas is dit niet te automatiseren door gebrek aan gegevens, sinds slechts van een aantal ruimten bekend is welke 'valve' de juiste is.

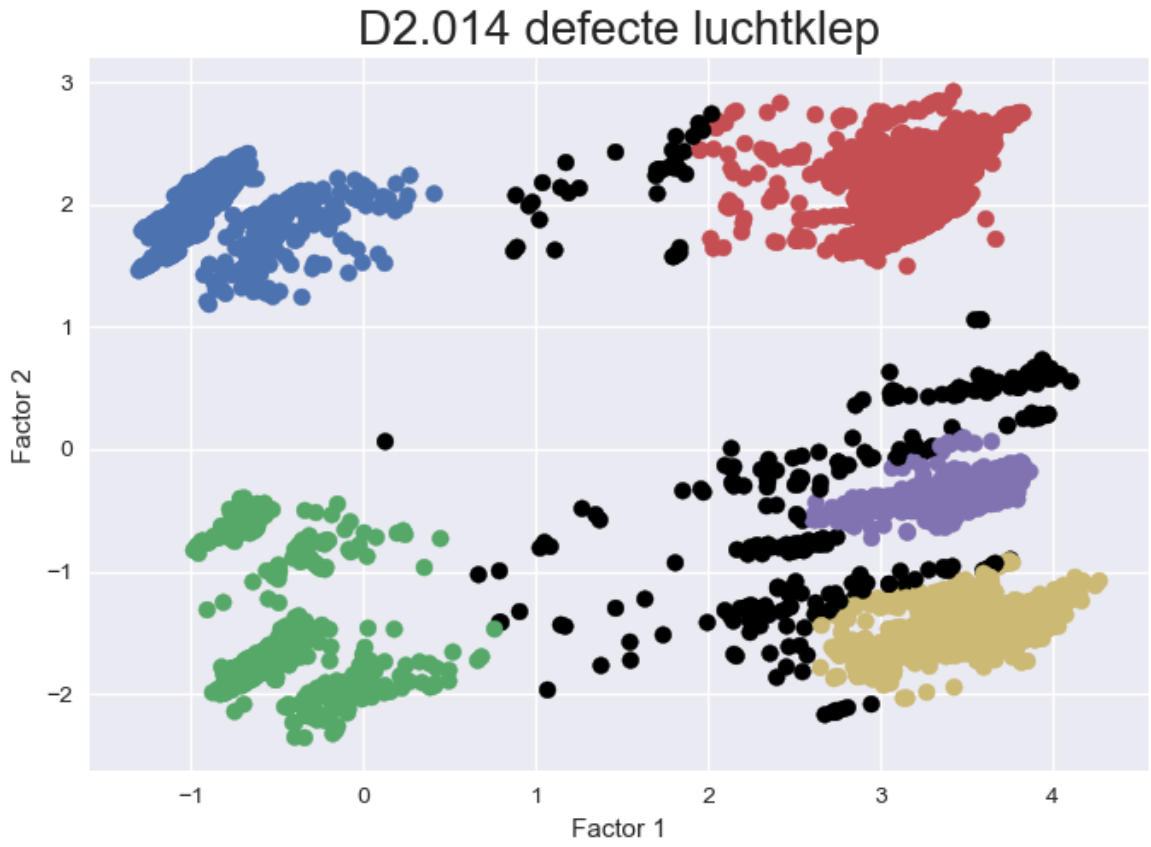
Aan de hand van de aangepaste dataset is de factor analyse nogmaals uitgevoerd, met wisselvallig resultaat. Dit leverde namelijk een factor op welk sterk door de luchtklep werd beïnvloedt, maar ook voor een deel door één van de waterkleppen. Hierdoor is het niet duidelijk zichtbaar wanneer de luchtklep defect was. In figuur **4.4.6** zijn de gevonden clusters te zien, met in het bruingeel de cluster waarin de defecte luchtklep ook toe behoort.



*Figuur **4.4.6**: Scatterplot tussen twee gevonden factoren waar de bruingele cluster onder andere de defecte luchtklep bevat.*

Dit resultaat is een goede stap in de juiste richting, maar het is de bedoeling dat alleen de defecten tot de cluster behoren. Gelukkig is ruimte D2.014 één van de ruimten waarvan bekend is welke klep de luchtklep is. Daarom is de factor analyse nog een keer uitgevoerd zonder de waterkleppen mee te nemen, hier zijn de clusters in figuur **4.4.7** op de volgende bladzijde uitgekomen. Hierin zijn twee clusters gevonden die iets vertellen over de defecte luchtklep, namelijk de paarse en bruingele clusters. De bruingele cluster geeft aan dat de luchtklep dicht staat, en de paarse geeft aan dat de luchtklep een beetje open is. Dit bewijst dat als het bekend is welke klep waartoe behoort, het wel degelijk mogelijk is om defecten te vinden.

Aan de hand van de gevonden defecten uit beide ruimten is geconcludeerd dat de factor analyse bruikbaar is voor defecten opsporen. Dit is tot nu toe grotendeels handmatig gebeurd en met slechts één jaar aan data ter beschikking. Dit moet echter algemener kunnen als het grootschalig gebruikt gaat worden. De stappen om te automatiseren staan in de volgende paragraaf beschreven, samen met in hoeverre het geautomatiseerde model werkt.

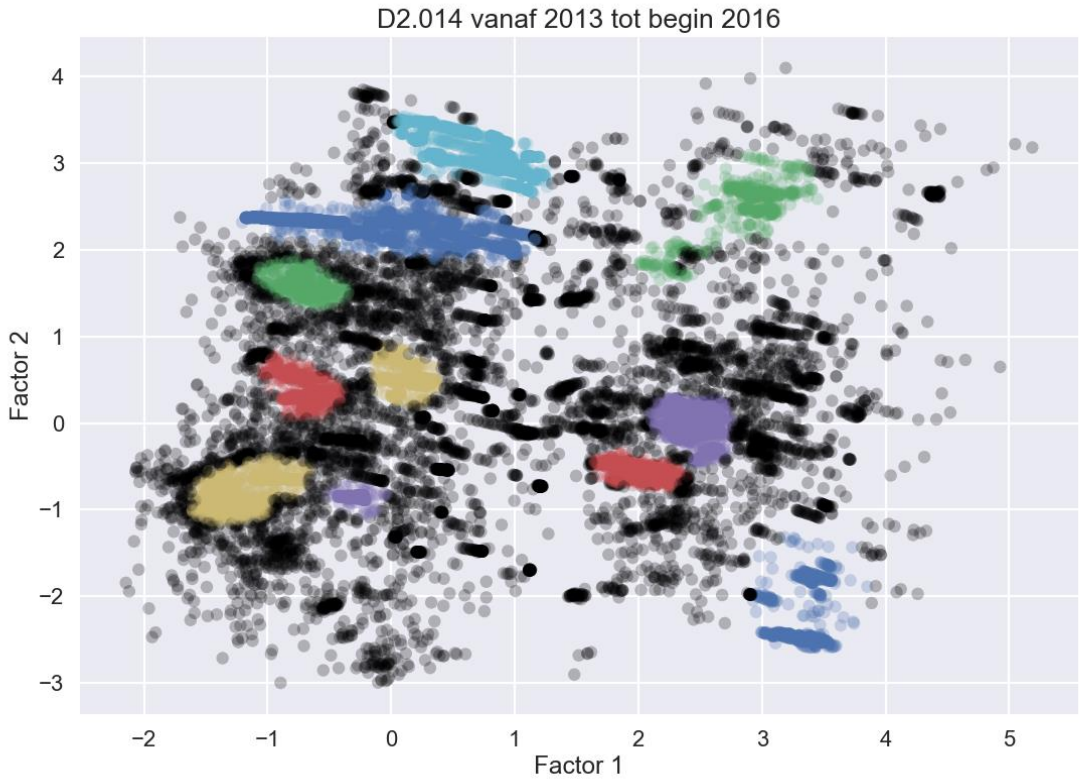


*Figuur **4.4.7**: Scatterplot tussen twee gevonden factoren waarvan de bruingele en paarse clusters de defecte luchtklep aangeven.*

4.4.4 Eindfase

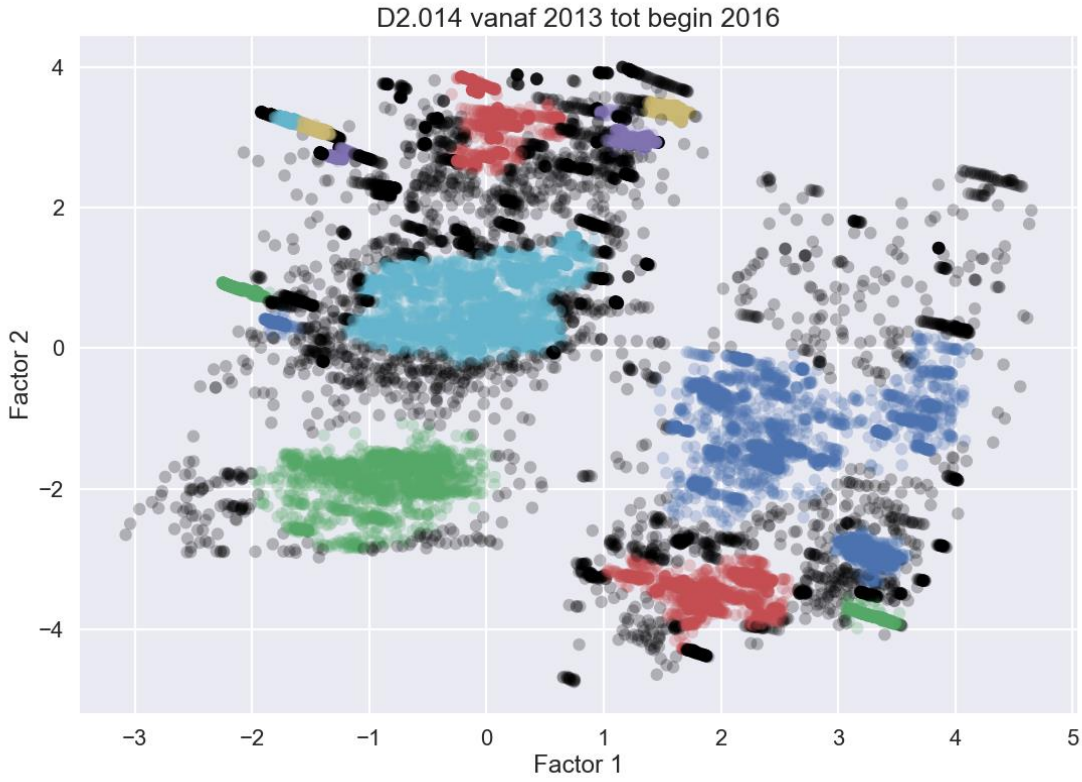
In de tussentijd is bekend dat factoranalyse toepasbaar is voor defecten opsporen aan de hand van bepaalde datasets. In deze paragraaf is verder gekeken naar de bruikbaarheid hiervan, in het bijzonder hoe het reageert op meerdere jaren aan data.

Voor deze controle is gekozen om nogmaals ruimte D2.014 te onderzoeken, maar nu op basis van alle data tussen oktober 2013 tot begin 2016. Hieruit is figuur **4.4.8** gevonden met daarin een paarse cluster aan de rechterhand. Deze cluster bevat een groot deel van de periode waarin de luchtklep defect was, maar niet allemaal. Verder bevat de cluster een aantal waarden waarin die nog niet defect was, dus een onbekende combinatie tussen variabelen speelt nog een rol.



*Figuur **4.4.8**: Scatterplot tussen twee factoren van alle beschikbare data van ruimte D2.014. De paarse cluster aan de rechterhand bevat de defecte luchtklep.*

Sinds dit geen cluster heeft opgeleverd waar concreet te zien is of een defect is of niet, is een nieuwe poging gewaagd met *n=10* factoren i.p.v. *n=5*. De gedachte hierachter is om de onbekende uit het vorige resultaat te vermijden door meer mogelijke factoren toe te voegen. Dit leverde helaas niet veel op zoals in figuur **4.4.9** is te zien. De donkerblauwe cluster rechts onderaan is een cluster wat overeenkomt met gevonden cluster in figuur **4.4.8**. Dit heeft dus geen verbetering opgeleverd.



*Figuur **4.4.9**: Scatterplot tussen twee factoren van alle beschikbare data van ruimte D2.014. De donkerblauwe cluster rechts onderaan bevat de defecte luchtklep.*

Hieruit is geconcludeerd dat factoranalyse wel degelijk toepasbaar is om ongewenst klimaatregelsysteem gedrag te detecteren. Dit werkt echter het beste met databestanden van één jaar in plaats van databestanden over meerdere jaren. De precieze reden hiervoor is niet bekend, maar hoogstwaarschijnlijk heeft dit te maken met de omvang van het bestand. Hoe groter de omvang hoe drukker de scatterplot 's worden en vallen eerst gevonden clusters weg.

4.4.5 Cluster herkenning

Voor cluster herkenning bestaat een flink scala aan mogelijke methoden, zoals het K-means algoritme. Binnen dit onderzoek is gekeken naar de toepasbaarheid van een aantal methoden op het vinden van clusters van dit onderzoek.
K-means: Als eerste is de K-means clustermethode bekeken. Dit is een methode wat in de praktijk veel wordt gebruikt sinds dit een snelle en simpele methode is om te implementeren. Deze is op één van de gevonden scatterplots uitgevoerd, maar dit leverde niet de gewenste clusters op. Één van de principes van K-means is namelijk dat elk datapunt aan een cluster wordt toegevoegd, ongeacht of die in de buurt is van een andere cluster of niet. Dit is niet hetgeen waar de onderzoeksgroep naar zocht, want datapunten die buiten clusters vallen kunnen ook interessant zijn. Verder is één van de argumenten voor de K-means het aantal clusters waaruit het bestaat, waardoor het lastig is om dit te automatiseren voor verschillende databestanden. Daarom is er dus niet voor K-means gekozen.

Affinity propagation: Als tweede is 'Affinity propagation' bekeken. Dit is ten opzichte van K-means een betere keuze door het feit dat je niet het aantal clusters als argument hoeft te geven. Het grote nadeel hieraan is echter dat elk datapunt aan een cluster wordt toegevoegd,

ongeacht of deze daar echt bij hoort of niet. Daarom is ook niet voor de ‘Affinity propagation’ techniek gekozen.
Spectral clustering: Dit is een uitbreiding op de K-means algoritme. Dit houdt in dat de data in eerste instantie door bepaalde algoritmen wordt getransformeerd waardoor de afstanden tussen punten veranderd. Na deze transformatie wordt weer de K-means toegepast. Dit houdt dus in dat het aantal clusters waar het uit bestaat bekend moet zijn. Hier is dus ook niet voor gekozen.
Agglomerative: Deze algoritme gaat er in eerste instantie vanuit dat elk datapunt een apart cluster vormt. Vervolgens voegt deze steeds clusters samen om zo steeds grotere clusters te vormen. Aan deze methode zijn twee al eerder besproken nadelen, namelijk dat je het aantal clusters moet aangeven en alle punten worden aan een cluster gevoegd. Deze methode is dus ook niet gekozen.
DBscan: Dit is een methode die, in tegenstelling tot de meeste clustermethoden, niet naar de afstand tussen punten kijkt maar naar de dichtheid tussen punten. Zo worden in eerste instantie de punten die niet super dichtbij andere punten liggen nog verder weg gezet zodat deze niet in clusters worden meegenomen. Vervolgens worden de overgebleven punten op basis van de dichtheid geclusterd. Het grote voordeel aan deze methode is dat niet alle punten aan clusters worden toegevoegd. Het nadeel van deze methode is dat een parameter voor de mate van dichtheid als argument moet worden gegeven. Deze parameter is anders voor zowat elke dataset, dus dit valt weer niet te automatiseren voor meerdere datasets. Daarom is niet voor deze methode gekozen.
HDBscan: Dit is een uitbreiding op de DBscan. In feite werkt de HDBscan precies hetzelfde, dus ook in eerste instantie de outliers verder weg doen en de rest vervolgens clusteren. Het enige verschil hiertussen is het argument wat aan het algoritme wordt gegeven. Bij de DBscan was dit een parameter voor de dichtheid, maar voor de HDBscan is het minimum aantal punten per cluster nodig. Dit levert een groot voordeel ten opzichte van alle voorgaande clustermethoden, want het minimum aantal punten is iets wat niet per dataset anders hoeft te zijn. Daarom is uiteindelijk gekozen voor de HDBscan als cluster herkenningsmethode.

5 CONCLUSIES

In dit onderzoek is gekeken naar verschillende manier waarop sensordata van De Haagse Hogeschool te Delft automatisch geanalyseerd en gepresenteerd kan worden, zodat anomalieën real time gemeld worden. Uit de resultaten is gebleken dat de clustermethode en deep learning beiden het beste resultaat hebben opgeleverd. Bij BBN dient het model zoals is aangeleverd van het 'LEGO' aangepast te worden voordat het een vergelijkbaar resultaat kan neerzetten. Voor BBN in Python x. RBS kwam in de haak met de bestaande applicatie, en kan daardoor uiteindelijk geen valide resultaat leveren.

Een ander vraagstuk was welke anomalieën er real-time plaatsvinden in De Haagse Hogeschool te Delft. Vanuit de methode RBS zijn over de periode van 01-01-2012 t/m 23-06-2016, 66 anomalieën gevonden. Hiervan zijn 64 anomalieën in dezelfde categorie: "CO2 sensor is waarschijnlijk kapot". De kans dat RBS de foutmelding "CO2 sensor is waarschijnlijk kapot" toont, is aanzienlijk groter dan de kans op andere foutmeldingen.
Door BBN in combinatie met de aangepaste applicatie te gebruiken zijn alle resultaten gevonden die in het aangeleverde BBN model voorkomen; overbezetting, luchtklep, ventilatie, PIR -, CO_2 - en Airflow sensor. Als men andere resultaten/defecten wil vinden, dan moet het huidige BBN model worden aangepast.

<<Stuk voor Andrè>> Hier alleen je conclusie!

De methode deep learning is hier niet mee bezig geweest, omdat de focus vooral lag op een voorspellend model. Verder zijn eerdere anomalieën niet geclassificeerd, daardoor kan het model ook niet leren en trainen.
Met behulp van clusteranalyse zijn een tweetal bekende anomalieën gevonden ter validatie van de bruikbaarheid van clusteranalyse. Ten eerste een defecte aanwezigheidssensor in ruimte D2.008 gedurende de periode tussen 20 april en 1 juni in 2015. Ten tweede is een defecte luchtklep in ruimte D2.014 gedurende een aantal maanden in 2014 gedetecteerd.

In de tweede deelvraag werd de vraag gesteld op welke manier de sensordata gebruikt kan worden om een anomalie op te sporen. Hiervoor zijn vier methoden voorgesteld en getest, namelijk RBS, BBN, deep learning en clustermethode.
Vanuit de methode RBS kan sensordata gebruikt worden om anomalieën op te sporen. Dit kan gebeuren op alle data van elke periode. Wanneer de data binnenkomt moet deze data worden geanalyseerd, hieruit komt vervolgens een resultaat. Het resultaat is onderverdeeld in één van de in paragraaf 4.1 beschreven categorieën. Echter kunnen bestaande anomalieën niet gebruikt worden voor eventuele analyses.

Voor BBN kan men door gebruik van de aangepaste applicatie om data op te halen vanuit de oude database, of invoeren als CSV bestand en in combinatie met het BBN model kan het geanalyseerd worden. Door middel van de SMILE engine kunnen de voorwaardelijke kansen volgens de waardes van het BBN model worden berekend in de applicatie.
BBN in Python is een vorm van supervised learning, en zal calculaties maken om een voorspellend model te bouwen en zo inzage te geven over het gewenste gedrag. Het model simplificeert deze calculaties door de waarschijnlijkheid uit te rekenen per attribuut, los van de andere attributen. Dit resulteert in een snelle en effectieve methode. De kans dat een bepaalde attribuut een bepaalde waarde geeft kan op deze manier uitgerekend worden. Vervolgens kun je alle waardes van de attributen vermenigvuldigen om zo een class te definiëren. Indien je meerdere klassen gedefinieerd hebt dan kun je dus aan de hand van alle

attributen bepalen wat de grootste kans is dat de waardes bij een specifieke class horen. De voordelen van deze methode zijn; gemakkelijk en snel te bepalen bij welke klasse het probleem hoort, het kan voorspellingen maken op real time data, het model wat de hoogste waarschijnlijkheid voor een specifiek probleem, er is een lijst met problemen (gesorteerd op voorkomen, dit wordt ook meegenomen bij de calculatie)

Ook vanuit deep learning kan sensordata gebruikt worden om anomalieën te detecteren. Dit kan door middel van een LSTM. De LSTM maakt een model waarin een gegeneraliseerde tijdserie voorspeld wordt. Als een nieuwe sensorwaarde een grote afwijking heeft van de voorspelde waarde kan het geclassificeerd worden als anomalie. Het is dan nog niet duidelijk wat voor anomalie het dan is. Zodra er een database gemaakt wordt welke anomalieën er optreden bij welke sensordata, kan er een nieuw neuraal netwerk gemaakt worden. Dat nieuwe neurale netwerk kan dan de verschillende anomalieën classificeren.
Tenslotte is het gebruik van clusteranalyse ook een mogelijkheid voor anomalieën detecteren. Dit werkt echter alleen op datasets van een bepaalde grootte. Dit komt doordat clusters minder duidelijk te voorschijn komen bij het analyseren van grote datasets. Hierdoor kunnen eerder bekende anomalieën onopgemerkt blijven. Verder blijven weinig voorkomende anomalieën onopgemerkt, sinds deze buiten de clusters vallen.

Voor de derde, en tevens, laatste deelvraag is uitgezocht op welke manier de gevonden anomalieën gebruikt worden om defecten te melden. Dit is te doen via de web-applicatie. Gebruikers kunnen hierdoor in de toekomst geïnformeerd worden over eventuele afwijkingen of defecten in het klimaatsysteem. Op dit moment werkt het alleen vanuit RBS, maar het is mogelijk om meerdere systemen te koppelen. Hierdoor kan men op verscheidene manieren anomalieën doorgeven vanaf welke methode dit dan ook afkomstig mag zijn.

5.1 Hoofdvraag

Hoe kan sensordata van De Haagse Hogeschool te Delft, automatisch worden geanalyseerd en gepresenteerd, zodat anomalieën real time gemeld worden?

Voordat de hoofdvraag beantwoord kan worden, moeten de antwoorden op de deelvragen in overweging genomen worden. Hieraan is onmiskenbaar verbonden dat elk van de drie onderzochte methoden zijn eigen voor- en nadelen heeft. Een samenstelling van deze methoden is dan ook onvermijdelijk.

Voor het analyseren van de gegevens is het, vanwege de uitbreidbaarheid, niet haalbaar om een RBS te gebruiken, noch een BBN. Dit omdat beide modellen op dit moment niet goed aansluiten in de wens tot real-time analyse van de gegevens. Daarnaast sluiten LSTM en clustermethode het best aan op de wensen die ‘LEGO’ gesteld heeft, dit was een zelf regulerend/automatische klimaat regelsysteem voor het gebruik van LSTM en cluster methode zijn geen experts nodig die situaties moeten schetsen of aannames moeten doen. Bij zowel BBN als RBS is er ten aller tijden een expert nodig die het systeem sturing biedt. Om deze reden is het dan ook beter om te kiezen voor een LSTM welke dan wel of niet in samenwerking met een Cluster methode kan worden ingezet.

Bovenstaande geeft echter nog geen antwoord op de vraag hoe deze gevisualiseerd kan worden, dan wel real-time beschikbaar te stellen aan de eindgebruiker(s). Hierom is een soortgelijk systeem, zoals het ontworpen dashboard van de RBS methode, de manier voor het doorgeven van real-time analyse resultaten. Hierin moet gedacht worden aan een soortgelijk

systeem waarin de Backend applicatie de logica zal moeten omvatten voor het analyseren en eventueel visualiseren van de data.

Hieraan willen wij toevoegen dat het gebruik van een web-applicatie de toegankelijkheid van de applicatie ten goede komt. Dit vanwege het feit dat een gebouwbeheerder uiteindelijk met een mobile device door de gang kan lopen terwijl hij het gebouw analyseert. Het enige wat hij hiervoor nodig heeft is een mobiel apparaat welke toegang heeft tot het netwerk zodat hij de web-applicatie kan gebruiken. De mogelijkheden van een web-applicatie zijn uiteraard in nog veel verdere maten uitbreidbaar tot aan het melden van klachten e.d. aan toe.
Kortom, de ideale situatie zou een soortgelijke web-applicatie zijn zoals het RBS Dashboard waaraan een analyse applicatie gekoppeld zit op basis van een LSTM met een eventuele Cluster toevoeging. In de aanbeveling wordt hier dieper op ingegaan.

6 DISCUSSIE

In dit onderzoek zijn er vier verschillende methoden onderzocht, waarbij elke methode is getoetst op de haalbaarheid voor het automatisch kunnen analyseren en presenteren van anomalieën van het klimaatregelsysteem. Uit het onderzoek is gebleken dat:

"De ideale situatie zou een soortgelijke web-applicatie zijn zoals het RBS Dashboard waaraan een analyse applicatie gekoppeld zit op basis van een LSTM met een eventuele Cluster toevoeging."

In dit hoofdstuk wordt ter reflectie nader ingegaan op de totstandkoming van het antwoord op de hoofd- en deelvragen.

6.1 Rule Based System

In het onderzoek naar de uitbreiding en verbetering van het reeds onderzochte RBS kwam snel naar voren dat het gebruikte systeem SAW (Sensor Application Wrapper) niet door te gebruiken was. Dit ligt mede ten grondslag aan een aantal defecten binnenin de applicatie welke niet goed naar voren gekomen zijn en niet konden worden verholpen. Hierdoor werkten de benodigde functionaliteiten van de applicatie niet naar behoren.

Het grootste nadeel hiervan is dat de validatiemethode voor andere methoden hiermee direct wegvalt. Dit doordat het RBS in voorgaande onderzoeken was ontworpen om de anomalieën te detecteren, hierdoor was het RBS een ideale methode om andere methodieken te valideren.

Door het wegvallen van deze methode is er geprobeerd dit te ondersteunen met het handmatig zoeken naar anomalieën op basis van eerder verkregen resultaten. Hieruit kan een solide conclusie worden getrokken, echter kan dit niet meer worden gevalideerd of de verkregen resultaten op het moment nog altijd dezelfde waren als verkregen.

In het ontwikkelproces van de Dashboard applicatie is naar voren gekomen dat het omzetten van de applicatie meer tijd in beslag nam dan van te voren was ingeschat. Hierdoor is er maar een gedeelte van de applicatie omgezet. De twee onderdelen welke nog niet volledig zijn omgezet zouden voor een volledige werking nog omgezet moeten worden. Daarentegen zijn de koppelingen voor de connectie met de web-applicatie al opgezet voor het merendeel van de deling van de informatie.

6.2 BBN en Deep Learning

Er is onderzocht in hoeverre het BBN en een Deep Learning methode gebruikt kunnen worden voor anomalie detectie. Hieruit is gebleken dat binnen de Deep Learning mogelijkheden een Long Short Term Memory (LSTM) de beste mogelijkheid is voor het voorspellen van de situatie. Door de voorspelling van het LSTM kunnen anomalieën worden gedetecteerd. Echter moet worden meegenomen dat voor het onderzoek naar de mogelijkheden van Deep Learning. Dit zonder enige voorkennis of ervaringen met een Deep Learning methode. Hierdoor moet worden opgemerkt dat de mogelijkheid bestaat dat het resultaat significant anders had kunnen zijn wanneer deze kennis reeds aanwezig was geweest. Hierdoor zou er tevens meer ruimte zijn geweest om de mogelijkheden binnen de Deep Learning verder te verkennen waardoor het resultaat op een bredere kennisbasis zou zijn gebaseerd.

Voor het BBN model was er reeds, door vooronderzoek uitgewezen, kennis beschikbaar binnen het lectoraat 'LEGO' waardoor dit qua kennisbasis geen problemen op mocht leveren. Echter was er tot op heden nog niet eerder getracht anomalieën te kunnen detecteren op basis van het BBN. Hierdoor ontstond de situatie dat er op onontgonnen terrein moest worden gewerkt, waardoor de resultaten afhankelijk zijn van de gekozen tussenoplossingen.

Bij zowel BBN als Deep Learning was het vrijwel niet mogelijk om een juiste validatie methode op te zetten. Dit komt doordat de tot dit onderzoek gebruikte validatiemethode, het RBS, niet kan omgaan met de nieuwe gegevens uit de nieuwste database van het huidige klimaatregelsysteem noch op dit moment de mogelijkheid heeft om gegeven te analyseren. Dit laatste doordat de aangedragen applicatie fouten bevat welke niet oplosbaar waren voor de huidige onderzoeksgroep. Beide modellen zijn hierdoor nu geverifieerd op basis van reeds bekende anomalieën, echter is deze validatie te klein om een gedegen solide acceptatie te garanderen.

Voor een volgend onderzoek raden wij dan ook aan om er zorg voor te dragen dat de kennis voor zowel BBN als Deep Learning verder op orde is waardoor de resultaten uit dit onderzoek kunnen worden verfijnend, evenals de gebruikte methoden. Tevens is het van het grootst mogelijke belang dat er een eenduidige validatiemethode komt waarmee beide methoden gevalideerd kunnen worden.

Voor een volgend onderzoek adviseren wij dan ook om door te gaan met het ontwikkelen van de web-applicatie in combinatie met het RBS en tevens in combinatie met een Backend applicatie voor een LSTM en/of een BBN.

7 AANBEVELINGEN

Het is niet wenselijk om een medewerker constant de resultaten van ‘Machine learning’ te laten visualiseren, daarom zou in een ideale situatie een dashboard middels een back-end systeem de uitkomst bieden. Er kunnen dan anomalieën real-time via de LSTM of clustermethode geanalyseerd en gepresenteerd worden. Op deze manier kunnen stakeholders op een toegankelijke manier inzicht krijgen zonder dat zij in detail hoeven te weten hoe het systeem functioneert. Ook is geen expert nodig die het systeem constant in de gaten moet houden (Exclusief onderhoud), het systeem leert immers zichzelf. Deze onderzoeksgroep raadt dan ook toekomstige onderzoekers aan het dashboard verder uit te werken.

De onderzoeksgroep heeft niet alle geplande doelen kunnen bereiken. De niet bereikte doelen zijn het toetsen van de validiteit van de vier methodes en het vinden van een toetsmethode waarmee de vier methodes eenduidig getoetst kunnen worden.

De onderzoeksgroep raadt aan om een eenduidige toetsmethode te ontwikkelen, voordat een vervolgonderzoek van start gaat. De reden hiervoor is dat het garandeert dat de methodes die onderzocht worden, gevalideerd en vergeleken kunnen worden.

Ook wordt aanbevolen om een vervolgonderzoek pas te starten wanneer een dataset verzameld is, waarvan minimaal 10% van de dataset bestaat uit gelabelde anomalieën. Deze gelabelde anomalieën dienen een representatie te zijn voor de verhouding van de verschillende type anomalieën.

Tot slot raadt de onderzoeksgroep, de toekomstige onderzoeksgroep, aan om ervoor te zorgen dat de database in orde is. Hiermee wordt bedoeld dat alle kolommen logische benamingen krijgen, die aansluiten op het desbetreffende lokaal, waardoor de data in de toekomst op een efficiënte manier verkregen kan worden.
Wanneer er data wordt opgehaald moet dit eerst gemanipuleerd en schoongemaakt worden zodat het gewenste onderzoek uitgevoerd kan worden. Deze problemen zijn niet gelijklopende (tijds)intervallen tot het ontbreken van waardes. In de toekomst kan er een hoop tijd bespaard worden door deze database zo optimaal mogelijk in te richten, zodat de data op een efficiënte manier gebruikt kan worden.

Bibliografie

Deng, L., & Yu, D. (2013). *Foundations and Trends in Signal Processing.* Redmond: Microsoft Research.

Itard, L. (2016). *Management samenvatting Jaarverslag 2015- Jaarplan 2016 Lectoraat Energie van de Gebouwde Omgeving.*

Jacobs, P., Dijken, F. v., & Boerstra, A. (2007). *Prestatie-eisen ventilatie in klaslokalen.*

Juba, B. (2017). *Principled Sampling for Anomaly Detection.* St. Louis: Washington University.

Riemer, H. (2017). *Factors of Comfort*. Opgehaald van educate-sustainability: https://www.educate-sustainability.eu/kb/content/factors-comfort

Timp, K. (2018). Klachtregistratie. (B. Tuynman, Interviewer)

A. Taal, L. Itard. Y. Zhao, W. Z. (2015). diagnose en foutencorrectie, 2–9.

Agentschap NL. (2010). Duurzame haagse hogeschool delft populair bij studenten.

Ast, T. Van, Scholte, J., & Wazir, F. (2016). SAW - Sensor Application Basics.

Goedhart, L. (2015). Klimaatsysteem Delft Gebruikerskant.

Goodfellow, I., Bengio, Y., & Courville, A. (n.d.). Deep Learning.

Kortekaas, C., & Vuuren, B. (2016). Foutdetectie van sensoren in het klimaatregelsysteem.

Marcopoloulos, J. (2014). Onderzoeksrapport hhs delft], 0–91.

Salcedo, T. B. (n.d.). *Intro to HVAC*.

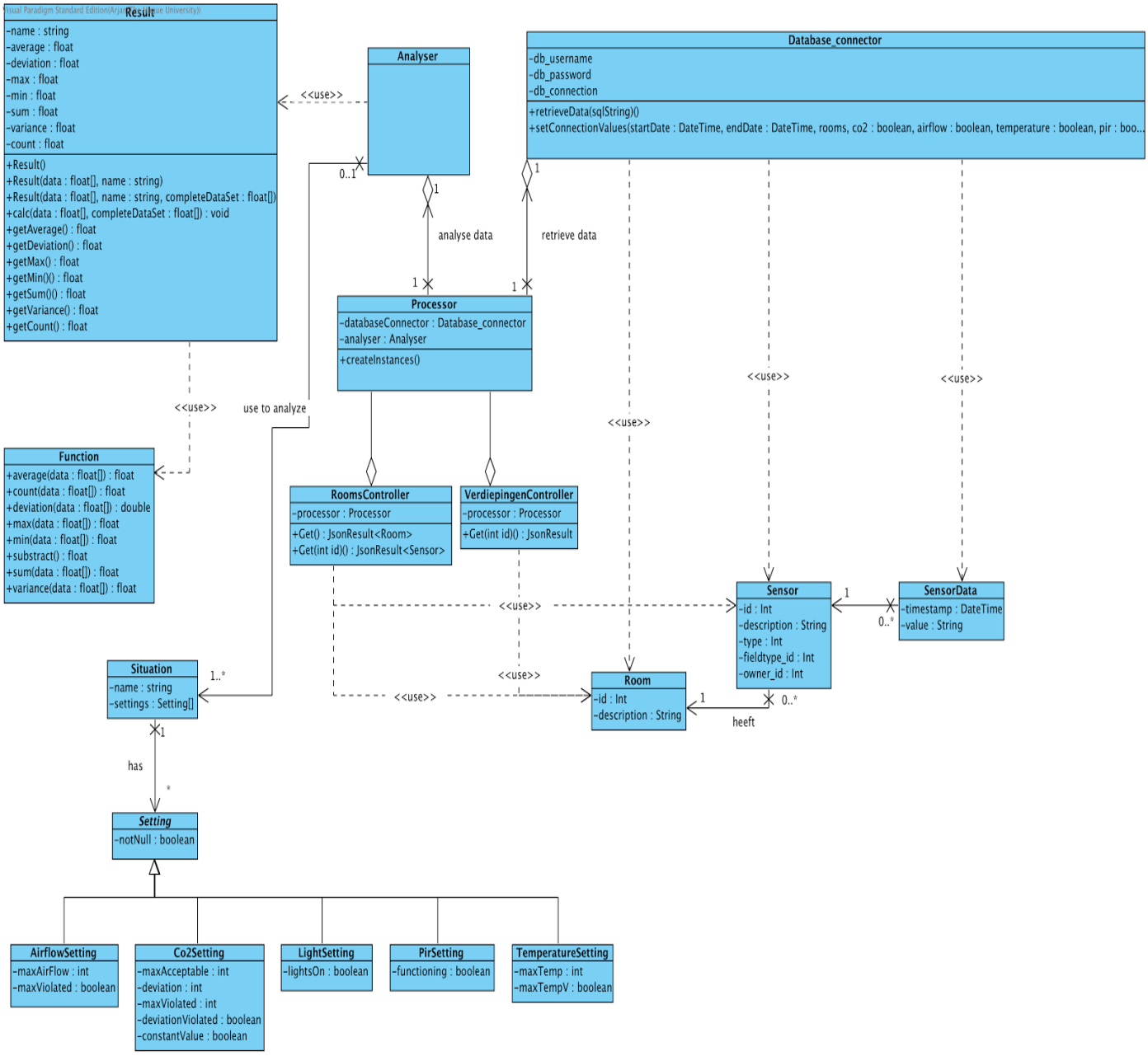
Yu, Y., Woradechjumroen, D., & Yu, D. (2013). A Review of Fault Detection and Diagnosis Methodologies on Air-Handling Units.

BIJLAGEN

4.1 XML-Format
https://docs.google.com/document/d/1DIp1L27s7zg96N58YBQQjgcOj64gqa9hemjrZvN9MAQ/edit?usp=sharing

4.2 Lijst met anomalieën
https://docs.google.com/spreadsheets/d/1R1ExH0SDwoFCKbwXbaE6WDLgA4RPWbhPs9AsQZEMUkI/edit?usp=sharing

4.3



Hier bijlage 4.4 t/m 4.7 van BBN

4.4 Hoge en lage Airflow grenswaardes (https://drive.google.com/file/d/1hTXD8Ttx1ZRExcERdn0dChb01p7U1a4J/view?usp=sharing)

4.5

Datum en tijd (uur)	High_ flow	Low_ flow	Flow_frozen	Flow_while_PIR_0	High_ CO2	Low_ CO2	CO2_frozen	CO2_neigh-bours_non_identical	PIR_ NaN
30-4-2015 08:00	1	1	1	0	1	1	1	1	1
30-4-2015 09:00	1	0	1	0	1	1	1	0	1
30-4-2015 10:00	1	0	1	0	0	1	1	0	1
30-4-2015 11:00	1	0	1	0	0	1	1	1	1
30-4-2015 12:00	1	0	1	0	0	1	1	0	1
30-4-2015 13:00	1	0	1	0	0	1	1	0	1
30-4-2015 14:00	1	0	1	0	0	1	1	0	1

30-4-2015 15:00	1	1	1	0	0	1	1	1	1
30-4-2015 16:00	1	0	1	0	0	1	1	0	1
30-4-2015 17:00	1	0	1	0	0	1	1	0	1
30-4-2015 18:00	1	0	1	0	1	1	1	0	1
30-4-2015 19:00	1	0	1	0	1	1	1	1	1
30-4-2015 20:00	1	1	1	0	1	1	1	1	1

4.6

High_ flow	Low_ flow	Flow_f rozen	Flow_while _PIR_0	High_ CO2	Low_ CO2	CO2_fr ozen	CO2_neighbours_ non_identical	PIR_ NaN
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	0
1	1	1	1	1	0	1	1	0
1	1	1	1	1	0	1	1	1
1	1	1	1	0	0	1	1	1
1	1	1	1	0	0	1	1	0
1	1	1	1	0	1	1	1	0
1	1	1	1	0	1	1	1	1
1	0	1	1	0	1	1	1	1
1	0	1	1	0	1	1	1	0
1	0	1	1	0	0	1	1	0
1	0	1	1	0	0	1	1	1
1	0	1	1	1	0	1	1	1
1	0	1	1	1	0	1	1	0
1	0	1	1	1	1	1	1	0
1	0	1	1	1	1	1	1	1

4.7

Air_flow_sensor	CO2_sensor	Damper	Fan	Occupancy_not_too_high	PIR_sensor
1	1	0,999976	1	0,999722	1
1	1	0,999976	1	0,999722	0
1	1	0,99995	1	0,999722	0
1	1	0	1	0,999722	1
1	1	0	0,997974	0,989293	1
1	1	0,999894	0,998868	0,993896	0
1	1	0,999948	0,998868	0,993896	0
1	1	0,996082	0,83786	0,165112	1
1	1	0,998632	0,001268	0,994469	1
1	1	0,999894	0,976758	0,993909	0
1	1	0,999783	0,976758	0,993909	0
1	1	0	0,926633	0,989664	1
1	1	0	1	0,999722	1
1	1	0,999897	1	0,999722	0
1	1	0,99995	1	0,999722	0
1	1	0	1	0,999722	1

B1. Additioneel effect op het gebouw

Het hoofdonderzoek richt zich volledig op het binnenklimaat van De Haagse Hogeschool te Delft. Er zijn externe effecten die invloed kunnen hebben op dit binnenklimaat. Er is onderzoek gedaan naar één van deze factoren. Dit onderzoek betreft de CO_2 waarde van de naar binnen gezogen lucht. De kwaliteit van deze lucht kan invloed hebben op het binnenklimaat.

1.1 Luchtkwaliteit

Mensen genereren CO_2 wat de lucht vervuild. Er is schone lucht nodig om alle gegenereerde CO_2 te vervangen. Deze lucht wordt door middel van twee luchtschachten het gebouw ingezogen. De mate van zuiverheid verschilt per tijdstip, om hier meer grip op te krijgen is onderzoek gedaan naar de kwaliteit van de lucht die het gebouw binnen komt. Dit onderzoek liep simultaan met de overige werkzaamheden.

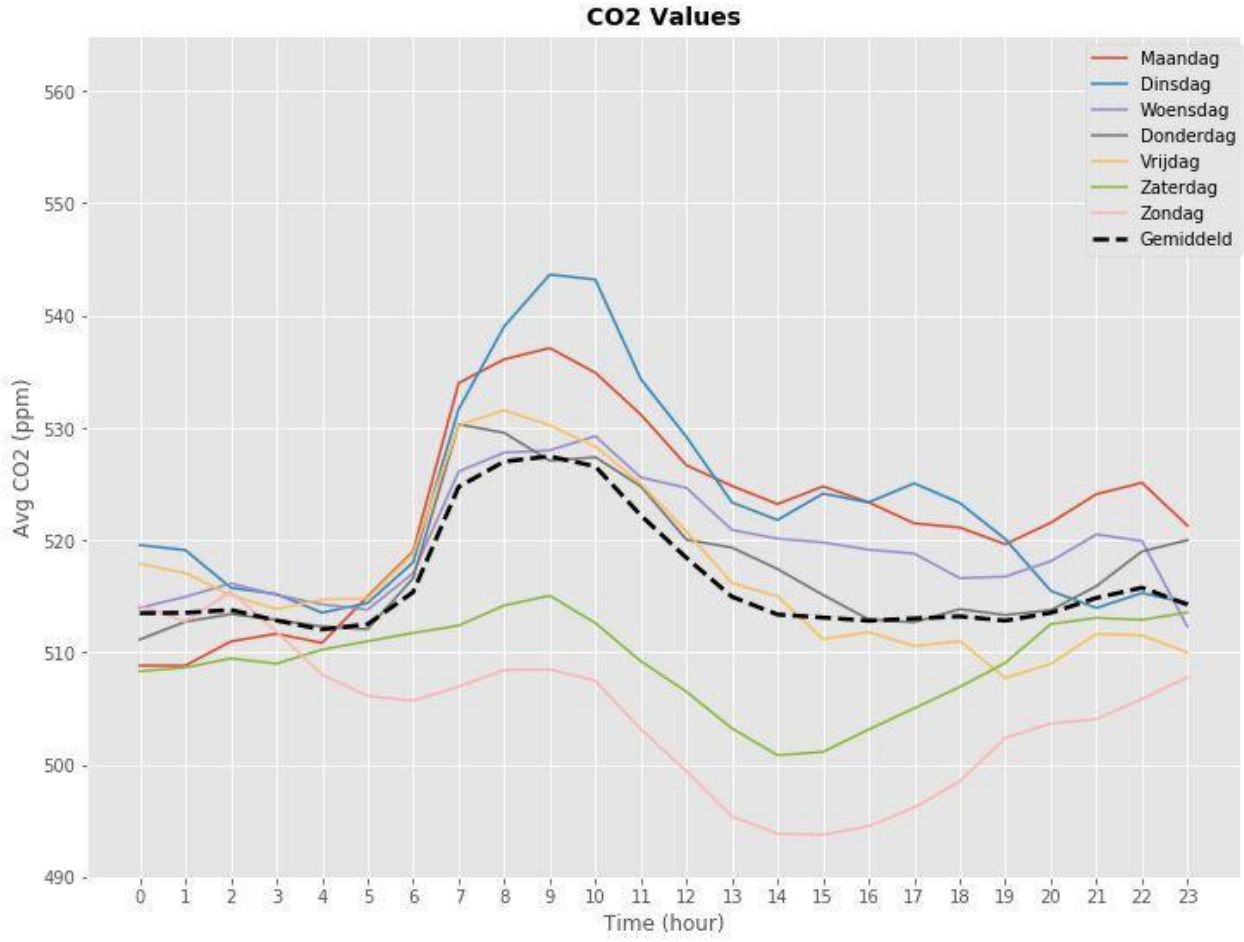
Er waren meerdere redenen om dit vraagstuk bij het hoofdonderzoek te betrekken. De grootste reden was om te kijken wat de invloed van het verkeer was op het CO_2 niveau van de naar binnen gezogen lucht. Afhankelijk van de uitkomst van dit onderzoek, kan men op bepaalde momenten meer of juist minder lucht vervangen. Hierdoor kan een hogere kwaliteit aan lucht binnen het gebouw gewaarborgd worden. De tweede reden is dat de sensoren geverifieerd moeten worden. Er is data van twee sensoren, alleen geen kennis welke data bij welke sensor past. Aangezien het zeker is dat aan een kant van het gebouw een hogere CO_2 waarde is. Kan op basis van de data beslist worden welke sensor bij welke data past. Het probleem hierbij is dat er geen zekerheid is over de kalibratie van een CO_2 sensor. Een oplossing hiervoor is om te kijken naar de delta tussen het hoogste en het laatste punt.

1.2 Situatie

De situatie is dat er twee luchtinlaten zijn. Elke inlaat heeft een CO_2 sensor die data waarneemt. De sensoren zijn gelabeld als sensor 87 en 94, deze labels zullen in dit onderzoek ook gebruikt worden. Voor dit onderzoek is data gebruikt van 31-10-2016 tot 16-06-2017. De analyse is uitgevoerd door middel van python.

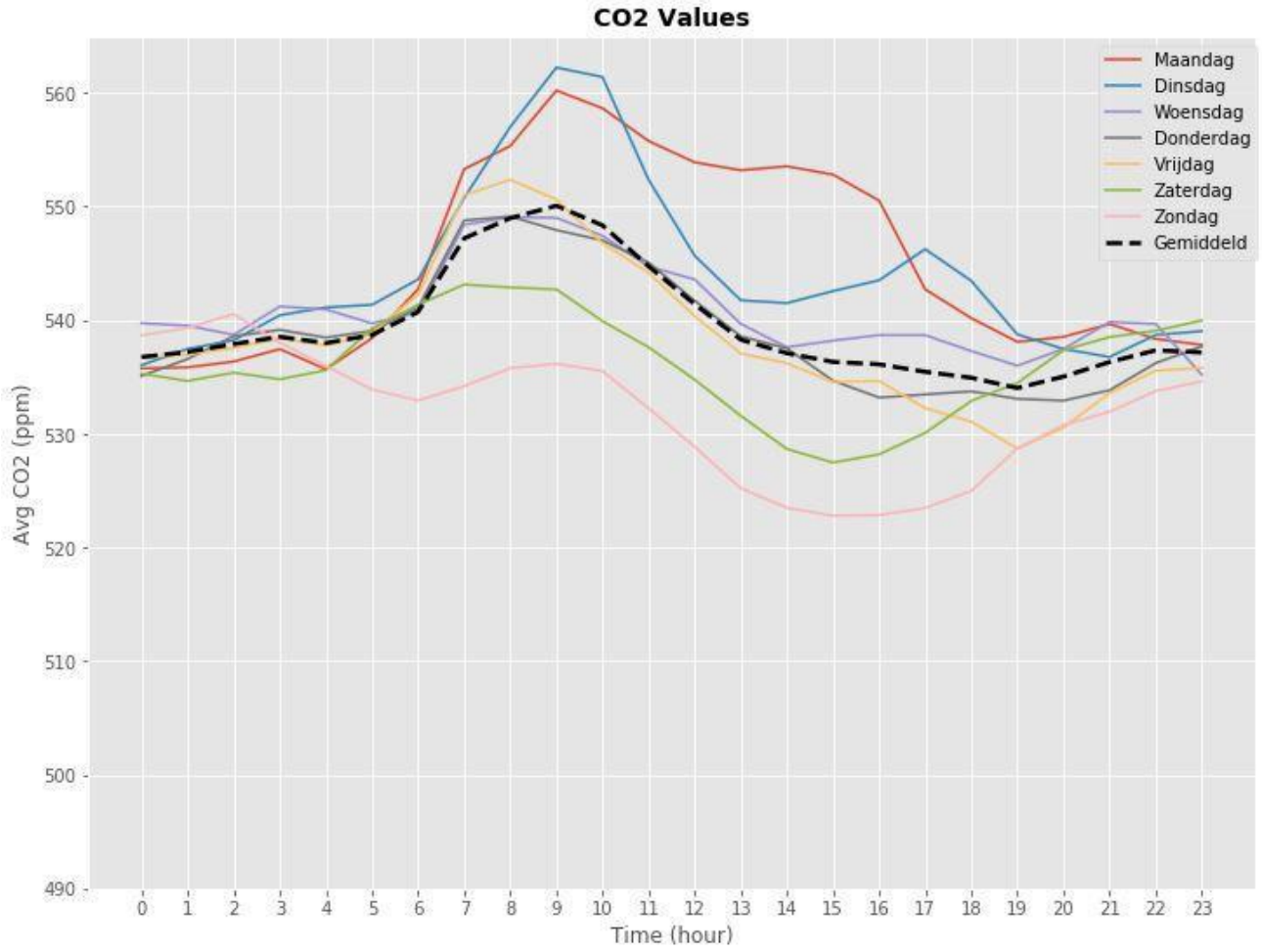
1.3 Onderzoek sensoren

Het hoofddoel is om te kijken of het verkeer invloed heeft op de CO_2 niveaus. We weten dat in de spits meer auto’s rijden. Door de data per uur te analyseren kan nagegaan worden of er meer CO_2 bij de luchtinlaten aanwezig is tijdens de spits. Een ander onderdeel is dat in het weekend een ander verkeerspatroon aanwezig is. Daarom zijn ook de dagen opgesplitst om een duidelijker beeld te hebben bij de impact van het verkeer. Voor deze analyse is gebruik gemaakt van een lijngrafiek. De grafiek geeft de CO_2 waarde weer ten opzichte van de uren in een dag. Dagen zijn in verschillende kleuren weergegeven.



Figuur 1: Sensor 87

Op basis van figuur 1 kunnen verscheidene conclusies getrokken worden. Duidelijk is een verhoogde CO_2 waarde tijdens de ochtendspits. Ten tweede is de luchtkwaliteit wat betreft de CO_2 in het weekend beter. Deze grafiek geeft de indicatie dat de verkeers hypothese klopt. Wat betreft de plaatsing van de sensoren zal er een vergelijking plaatsvinden in hoofdstuk 1.5.



Figuur 2: Sensor 94

Op basis van figuur 2 kunnen de volgende conclusies getrokken worden:

- Tijdens de ochtendspits zijn de CO_2 waarde verhoogd;
- Wat betreft het weekend kan hier weer gesproken worden van een alternatief patroon.

1.4 Vergelijking resultaten

De resultaten van de twee grafieken zijn vergelijkbaar. Beide tonen ze de impact van het verkeer. Wat betreft het koppelen van de sensor en de data is gekeken naar de spreiding. Hierbij heeft sensor 87 een spreiding van 50 CO_2 , en sensor 94 geeft een spreiding van 40 CO_2 . Omdat verkeer de grootste invloed heeft op deze schommeling wordt verwacht dat sensor 87 aan de drukke zuidkant van het gebouw zit.

1.5 Conclusie

Er is aangetoond dat het verkeer invloed heeft op de CO2 waarden. Er is vervolgonderzoek nodig hoe deze bevinding kan worden geëxploiteerd. Wat betreft de plaatsing van de sensoren zal gekalibreerde data nodig zijn om een definitieve conclusie te kunnen trekken.