INTRODUCTIE IN MACHINE LEARNING EN DATA SCIENCE VOOR TOEPASSING BINNEN DE BOUWFYSICA

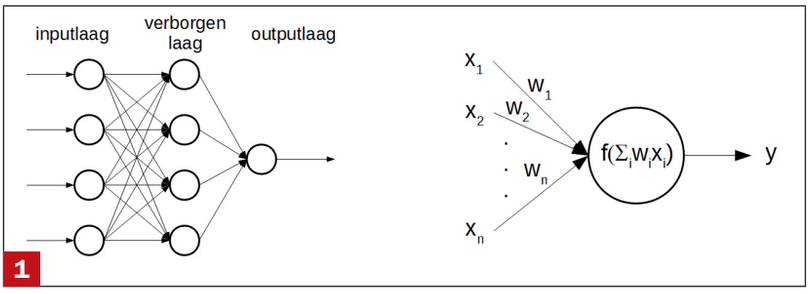
In de huidige maatschappij worden machine learning en data science steeds relevanter. Hoewel dit nog in minder mate zichtbaar is, zal dit ook binnen de sector bouwfysica steeds meer duidelijk worden. In dit artikel wordt een introductie in machine learning en data science gegeven en worden een aantal belangrijke methodieken inzichtelijk gemaakt. Tevens wordt aan de hand van een voorbeeld de kracht ervan geïllustreerd.

Wie de actualiteiten volgt zal wellicht weten dat er een transitie gaande is naar een datagedreven maatschappij [1]. Door toepassing van sensoren en dataopslag komen steeds meer digitale gegevens beschikbaar. De totale hoeveelheid data in de wereld betreft inmiddels miljarden gigabytes. Elke 2 tot 3 jaar verdubbelt de hoeveelheid data. Tevens is er duidelijk sprake van een significante toename van data-utilisatie. Door de toenemende beschikbaarheid van een grote variëteit aan data en de toepassing van slimme machine learning algoritmes zijn veel processen, producten en/of methodes te optimaliseren. Hier kan bijvoorbeeld worden gedacht aan het optimaliseren van logistieke processen wat leidt tot kostenreductie. Een ander voorbeeld is het diagnosticeren van ziektes en het voorspellen van de meest effectieve behandelmethodes. Weer een ander voorbeeld heeft betrekking op het ontwikkelen van staal met een zeer hoge kwaliteit. En de lijst van toepassingen wordt langer en langer. Interessant genoeg is de toepassing van machine learning en data science binnen de sector bouwfysica vooralsnog beperkt. Opvallend, gezien het juist een sector is waar veel draait om informatie en kennis. In dit artikel zal een inkijkje worden gegeven in de wereld van machine learning en wordt een beeld van de potentie die machine learning voor de sector bouwfysica biedt geschetst.

dr. ir. R.J. (Robbert-Jan) Dikken, Peutz bv, Zoetermeer

ARTIFICIËLE NEURALE NETWERKEN, BACKPROPAGATION EN GENETISCHE ALGORITMES

Het principe van machine learning is dat algoritmes op basis van data slimmer worden waardoor hun voorspellende kracht verbetert [2]. Veel machine learning-toepassingen maken gebruik van artificiële neurale netwerken. Maar wat is een artificieel neuraal netwerk (ANN) nu precies? Het principe is gebaseerd op de informatieverwerking in biologische neurale netwerken, oftewel hersenen. Het is een netwerk van neuronen, wat eenheden zijn die relatief eenvoudige mathematische activatiefuncties representeren. Elk neuron geeft afhankelijk van de input van dat specifieke neuron en de verbindingen (synapsen) tussen de neuronen een outputsignaal. Het collectief van de verwerking van inputsignalen van alle neuronen representeert een bepaalde mathematische functie die voor elk specifiek probleem anders is. In figuur 1 is dit schematisch weergegeven.

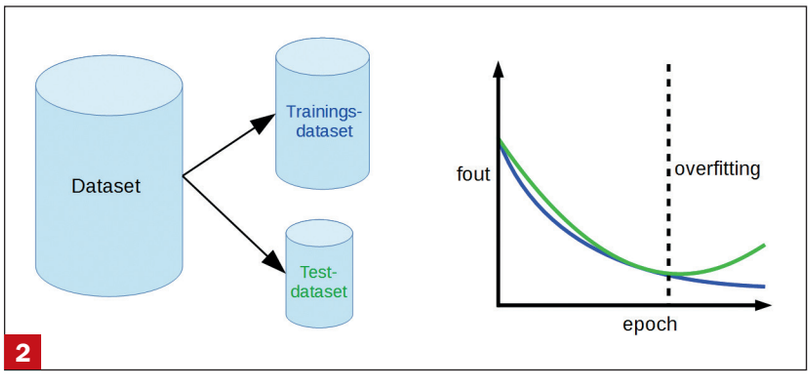


1

Schematische weergave van een feedforward artificieel neuraal netwerk en van een individueel neuron

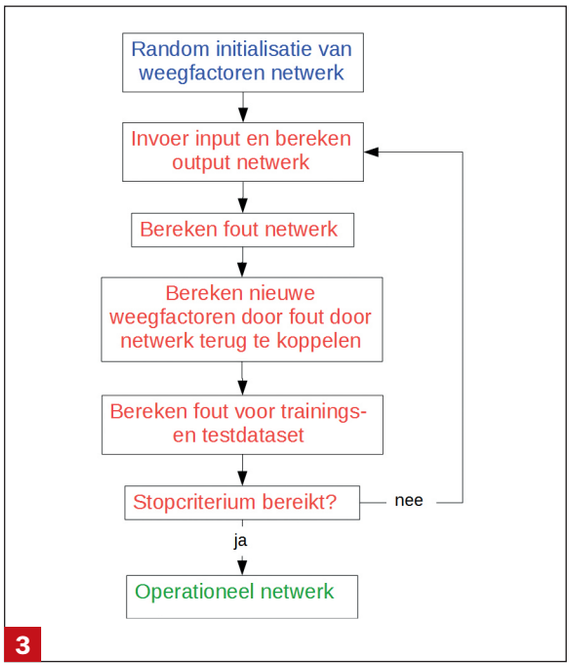
Er zijn verschillende aspecten van belang bij het ontwikkelen van toepassingen die gebruikmaken van neurale netwerken. Zo dient bijvoorbeeld gedacht te worden aan het aantal neuronen waaruit het neuraal netwerk dient te bestaan. Dit is zeer belangrijk. Een te klein netwerk betekent dat niet alle relaties tussen input en output kunnen worden gerepresenteerd. Een te groot netwerk (bij een te kleine / incomplete dataset) betekent dat het risico van overfitting bestaat, en dat de output van het netwerk dus geen generiek karakter heeft. Om dit te voorkomen wordt een dataset altijd in minstens twee delen opgedeeld (figuur 2). Eén deel wordt gebruikt om het netwerk de relaties tussen input en output te laten leren, de trainingsdataset, en een tweede deel wordt gebruikt om te testen hoe generiek het netwerk presteert, de testdataset. Helaas bestaat er geen gouden regel wat precies de meest optimale grootte van een neuraal netwerk is. Dit komt vaak neer op gebruikmaken van bestaande kennis opgedaan uit eerdere ervaringen. Het is wel mogelijk routines te ontwikkelen die zelfstandig de optimale vorm van het netwerk zoeken. De keuze van activatiefuncties is tevens een belangrijk aandachtspunt. In het geval dat de parameter(s) aan de outputzijde zowel negatief als positief kunnen zijn, dient hier een bijbehorende functie bij gekozen te worden, bijvoorbeeld de hyperbolische tangens. Wanneer waarden uitsluitend positief zijn wordt bijvoorbeeld een logistische functie veel toegepast.

Een neuraal netwerk wordt met bestaande data getraind, wat inhoudt dat het netwerk op basis van de data de relatie tussen input en output leert. Voor het trainen van een neuraal netwerk zijn meerdere methoden beschikbaar. Eén van de meest toegepaste methoden om een ANN te trainen is het backpropagation algoritme. Het ANN leert door de fout tussen de output van het netwerk en de werkelijke waarden in de trainingsdataset in combinatie met de afgeleiden van de activatiefuncties door het netwerk terug te koppelen, waarbij de weegfactoren tussen de neuronen aangepast worden. Effectief wordt hiermee beoogd het globaal minimum van het netwerk te vinden om het netwerk een generiek karakter te geven. Dit proces is in figuur 3 schematisch weergegeven.



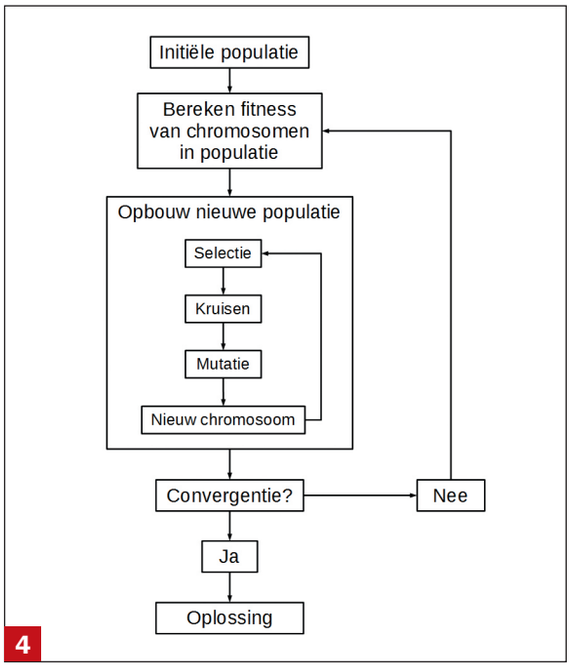
Opdeling dataset in een trainings- en testdataset

Een andere methode om een neuraal netwerk te trainen is met behulp van een genetisch algoritme (GA). Een genetisch algoritme is een krachtig zoek- en optimalisatiealgoritme dat voornamelijk effectief is voor complexe problemen waarbij de oplossingsruimte non-convex is waardoor klassieke regressiemethodes vaak niet toepasbaar zijn. Het principe van genetische algoritmes is gebaseerd op biologische evolutie. Een initieel geheel willekeurige populatie van oplossingen (chromosomen) ondergaat genetische operaties zoals selectie, kruising en mutatie waardoor de populatie convergeert naar de oplossing die het meest compatibel is met het probleem. Elke oplossing wordt een bepaalde fitness toegekend die definieert hoe compatibel de oplossing met het probleem is. Vervolgens worden willekeurige oplossingen geselecteerd, waarbij oplossingen met een hogere fitness een grotere kans op selectie hebben. Twee geselecteerde oplossingen worden vervolgens gekruist, wat inhoudt dat 50% van de ene en 50% van de andere oplossing gecombineerd worden tot een nieuwe oplossing. Met een bepaalde waarschijnlijkheid wordt dan nog een deel van de oplossing willekeurig gemuteerd. Dit wordt herhaald tot een nieuwe populatie is opgebouwd. Er worden nieuwe populaties gegenereerd totdat een bepaald convergentiecriterium is behaald. In figuur 4 is het concept van genetische algoritmes weergegeven. Bij de toepassing zijn verschillende aspecten waar rekening mee gehouden dient te worden. Zo is het voor het vinden van de juiste (globale) oplossing essentieel dat het aantal permutaties in de populatie groot genoeg is, wat inhoudt dat de populatie groot genoeg moet zijn. Dit is ook het nut van muteren, het variëren in de oplossingsruimte, terwijl het selecteren op basis van fitness en kruisen juist convergentie drijft.



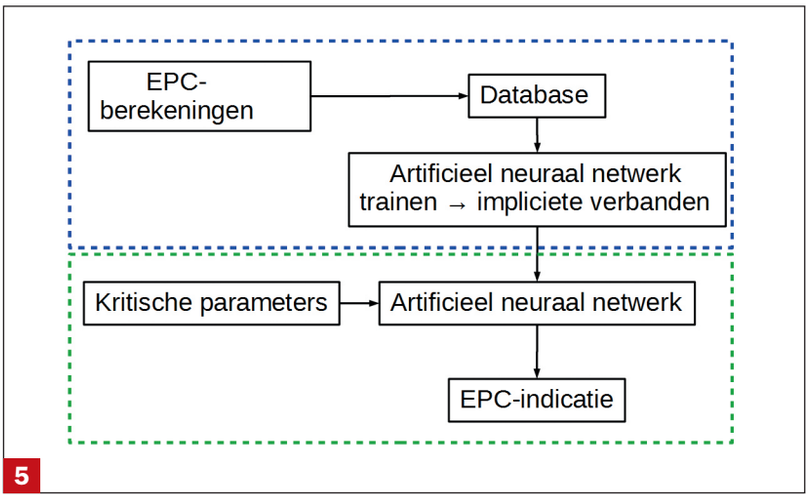
Schematische weergave van het backpropagation algoritme

In het geval van het gebruik van een genetisch algoritme voor het trainen van een neuraal netwerk betekent dit dat de chromosomen de weegfactoren tussen de neuronen representeren. Het algoritme zoekt dus een configuratie van weegfactoren voor het neuraal netwerk. Hierbij is de fitnessfunctie gedefinieerd op basis van de trainingsdataset. Tijdens de evolutie wordt voor de gehele populatie de prestatie van de oplossingen voor zowel de trainingsdataset als testdataset berekend. Op basis hiervan wordt uiteindelijk een configuratie gevonden die het meest compatibel met de data is.



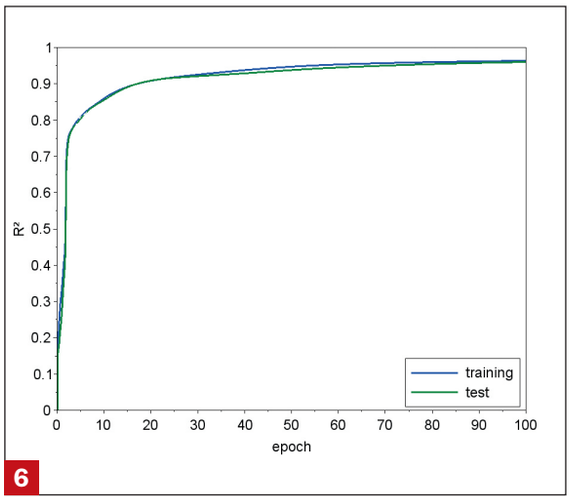
Schematische weergave van het concept van genetische algoritmes

In het vervolg zal een voorbeeld gegeven worden van een toepassing van machine learning binnen de bouwfysica.



5

Schematische weergave van een EPC-tool gebruikmakend van neurale netwerken



6

De determinatiecoëfficiënt als functie van trainingsiteratie voor zowel de trainingsdataset als de testdataset

EPC-INDICATIE OP BASIS VAN DATA EN NEURALE NETWERKEN

De energieprestatiecoëfficiënt (EPC) is de afgelopen jaren het uitgangspunt geweest om de prestatie van gebouwen met betrekking tot energiegebruik te beoordelen. De berekening van de EPC vindt plaats aan de hand van de NEN 7120, een boekwerk van zo'n 500 pagina's die een berekening op basis van circa 150 parameters beschrijft. Uiteindelijk zijn er echter slechts enkele parameters die dominant zijn in de bepaling van de EPC. Het is dan ook handig in bijvoorbeeld het ontwerptraject van een gebouw niet de volledige EPC-berekening te hoeven doen, maar om hiervoor een tool te hebben die als input de kritische parameters heeft en eenvoudig en snel een indicatie geeft. Door jarenlange ervaring kunnen deze kritische parameters op basis van expertkennis bepaald worden:

- Grondoppervlakte;
- Gebruiksoppervlakte;
- Dakoppervlakte;
- Geveloppervlakte;
- Totaal PV-piekvermogen;
- Aanwezigheid warmtepomp;
- Aanwezigheid zonneboiler;
- Glasoppervlakte;
- Rc-waarden;
- U-waarden glas;
- Aanwezigheid douche-WTW;
- Aanwezigheid externe warmtelevering en rendement;
- Type ventilatiesysteem (C/D).

Uiteraard is het nog niet evident hoe de combinatie van de parameters leidt tot een indicatie van de werkelijke EPC van een gebouw. Machine learning, en specifiek de toepassing van artificiële neurale netwerken, biedt hier een oplossing. Een database met de door de jaren heen uitgevoerde EPC-berekeningen kan gebruikt worden om een neuraal netwerk de relatie te leren tussen de genoemde kritische parameters en de bijbehorende EPC. Dit is schematisch weergegeven in figuur 5. Het blauw-omkaderde deel bevat het trainingsproces en het groen-omkaderde deel bevat de voorspelling van de EPC van nieuwe gebouwen.

Door de jaren heen zijn er duizenden EPC-berekeningen gedaan. Met deze data kan een artificieel netwerk (ANN) getraind worden dat snel een EPC-indicatie geeft op basis van enkele gebouw- en installatieparameters. In theorie is een ANN dat bestaat uit twee lagen in staat elke willekeurige non-lineaire functie te representeren, mits de benodigde data beschikbaar is en de mathematische methodiek die het trainen van het netwerk mogelijk maakt goed wordt toegepast. Hierbij moet over verschillende aspecten worden nagedacht. Zo kan de dataset met alle EPC-berekeningen bijvoorbeeld wel heel groot zijn, maar dat houdt niet per definitie in dat deze direct optimaal bruikbaar is. De dataset kan bijvoorbeeld sterk non-uniform zijn, doordat bijvoorbeeld de eis voor nieuwbouw lange tijd 0,4 is geweest en er dus veel berekeningen zijn gedaan waarbij de gebouw- en de installatieparameters richting de EPC-eis zijn ontworpen. Het non-uniform zijn van datasets bemoeilijkt het trainen van een neuraal netwerk. Daarom is er in dit geval gekozen voor de toepassing van een genetisch algoritme om het neuraal netwerk te trainen, omdat deze methodiek het beste om kan gaan met "gaten" in de dataset. De dataset is zo goed mogelijk uniform gemaakt door uniforme random sampling.

In figuur 6 is de determinatiecoëfficiënt R^2 weergegeven tijdens de training met een GA voor zowel de trainingsdataset als de testdataset. R^2 kwantificeert de accuraatheid van een fit ofwel een statistisch model. $R^2 = 1$ representeert een perfecte fit. De determinatiecoëfficiënt neemt tijdens het trainen voor zowel de trainings- als de testdataset toe richting $R^2 = 0{,}96$. Dit betekent dat de output van het neuraal netwerk bij de input van de kritische parameters de werkelijke EPC zeer goed benaderd, binnen de grenzen van de dataset. Buiten de grenzen van de dataset is het neuraal netwerk beperkt in staat te extrapoleren, maar is de voorspellende kracht aanzienlijk kleiner. In figuur 7 is de door het getrainde neuraal netwerk voorspelde EPC uitgezet als functie van de werkelijke EPC na 118 trainingsiteraties, wat aan het einde van de curves in figuur 6 is.

VOORSPELLEN BENG-INDICATOREN MET NEURALE NETWERKEN EN DATA

Machine learning en data science biedt ook uitkomst bij de introductie van de nieuwe BENG-eisen die de EPC gaan vervangen. BENG staat voor Bijna EnergieNeutrale Gebouwen [3]. BENG 1 geeft de maximale energiebehoefte in kWh per m² gebruiksoppervlakte per jaar. BENG 2 geeft het maximale primair fossiel energiegebruik in kWh per m² gebruiksoppervlakte per jaar. En BENG 3 geeft het minimale aandeel hernieuwbare energie in procenten. Met de introductie van deze eisen kan enige onzekerheid ontstaan in ontwerptrajecten of voldaan kan worden aan deze nieuwe energieprestatie-eisen.

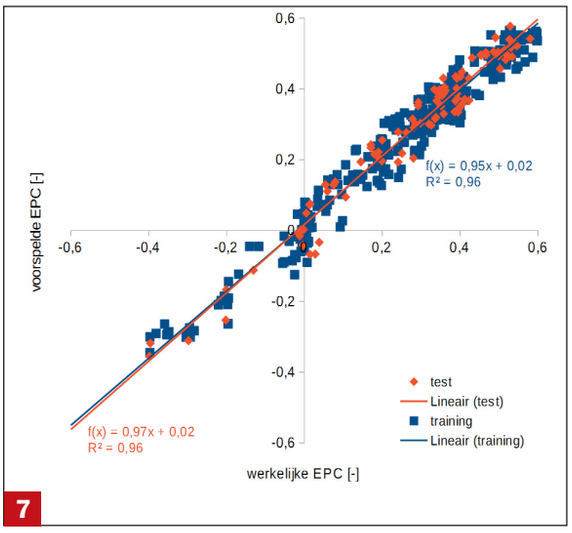
Analoog aan de methodiek om de EPC-indicatie te voorspellen, zo kunnen neurale netwerken ook uitkomst bieden om inzicht te krijgen in de verwachte BENG-indicatoren van een gebouw. In figuur 8 is dit weergegeven voor alle drie de BENG-indicatoren. Dezelfde kritische parameters liggen hieraan ten grondslag als voor de EPC. In dit geval is ervoor gekozen, uit oogpunt van eenvoud, om een enkel neuraal netwerk te trainen op alle drie de BENG-indicatoren tegelijk. Het is uiteraard tevens mogelijk om drie afzonderlijke netwerken te trainen voor de drie BENG-indicatoren. Zoals eerder voor de EPC ook reeds is aangegeven, de grenzen van de database geven ook de grenzen waarbinnen het getrainde netwerk de BENG-indicatoren van een gebouw kan voorspellen.

BREDE TOEPASSING

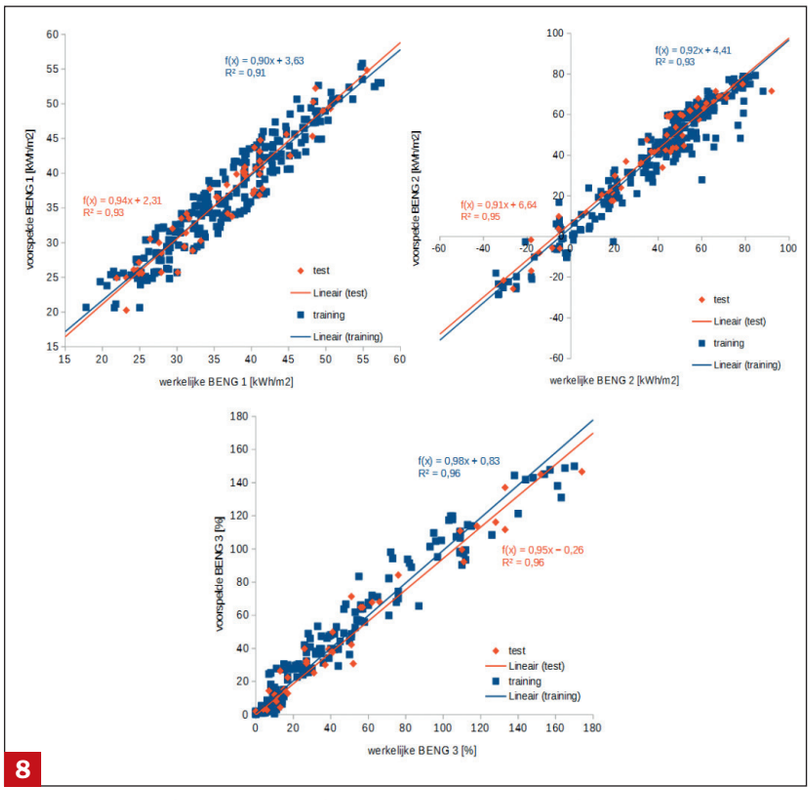
In dit artikel is een introductie in machine learning en data science gegeven en aan de hand van een relatief eenvoudig voorbeeld binnen de bouwfysica de kracht ervan geïllustreerd. Door het abstracte en conceptuele karakter van machine learning-algoritmes is het toepassingsveld zeer breed. In de lijn van de tool om de EPC en BENG-indicatoren te voorspellen is een andere toepassing een tool om het benodigd PV-vermogen te voorspellen om te voldoen aan een bepaalde energieprestatie-eis. De ontwikkeling van een dergelijke tool is analoog aan de genoemde tools en is gebaseerd op dezelfde data.

Een ander voorbeeld dat voor de bouwfysicasector interessant is, is de ontwikkeling van zelflerende regelingen van klimaatinstallaties die optimaliseren op energiegebruik en comfort. Deze regelingen kunnen worden ontwikkeld door gebruik te maken van dynamische energiemodellen van gebouwen.

Zo zijn er talloze toepassingen binnen vakgebieden van geluid en trillingen, tot duurzaamheid en comfort, tot wind en milieu, waar machine learning en data science een toegevoegde waarde kunnen bieden. Adviesbureau Peutz heeft het afgelopen jaar uitgebreid onderzoek gedaan naar ontwikkelingen en toepassingen van machine learning en data science breed over alle vakgebieden waarin het bureau actief is. Duidelijk is gebleken dat er een significante meerwaarde zit in de integrale benadering van machine learning, fysische modellen en data. ■



7 Voorspelde EPC als functie van werkelijke EPC



8 De voorspelde BENG-indicatoren als functie van de werkelijke BENG-indicatoren

BRONNEN

▶ **[1]** L. Kool, J. Timmer en R. van Est, De datagedreven samenleving - Achtergrondstudie, Den Haag, Rathenau Instituut 2015

▶ **[2]** T.M. Mitchell, Machine Learning, The McGraw-Hill Companies, Inc. 1997

▶ **[3]** https://www.rvo.nl/onderwerpen/duurzaam-ondernemen/gebouwen/wetten-en-regels-gebouwen/nieuwbouw/energieprestatie-beng/wettelijke-eisen-beng