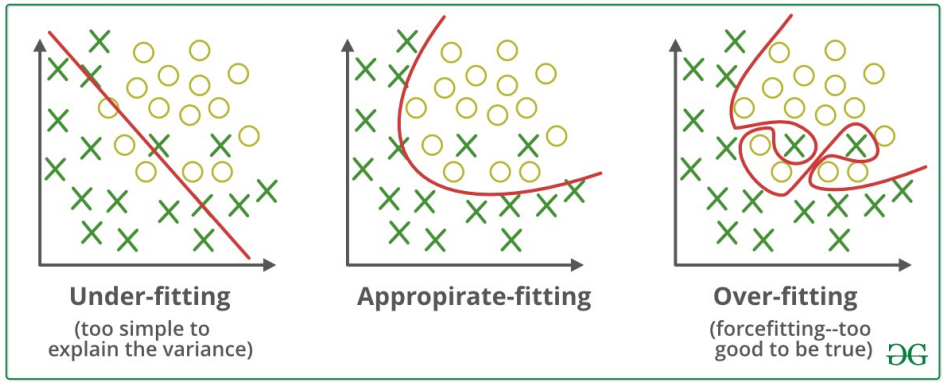
Data-Science 1

overfitting



Karel de Grote Hogeschool

24/04/2024

Inhoud

- overfitting
- training en test dataset

Overfitting

Wat is het probleem?

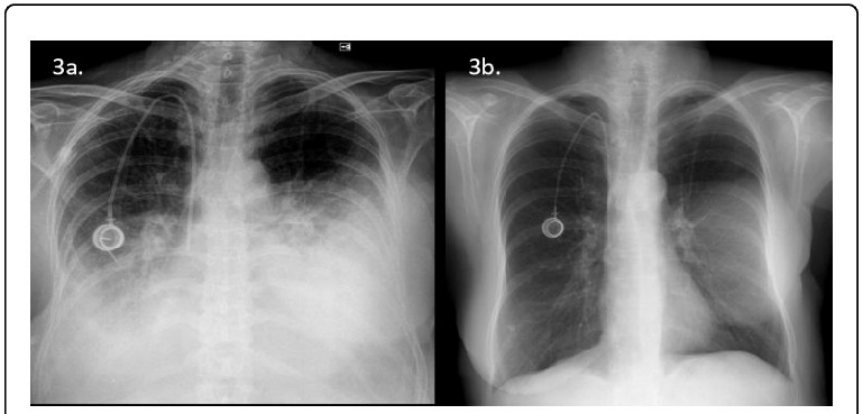
- voorbeeld1:
  - beslissingsboom met Simpsons
  - configureer naam als “categorical” en bekijk het resultaat

Wat is het probleem?

- voorbeeld2:
  - beslissingsboom met leninggegevens
  - zet diepte op onbeperkt
  - bekijk de boom
  - zet de diepte nu op 1
  - bekijk de boom
  - welke boom is het beste in de praktijk?
- opmerking: de diepte van de boom is een “hyperparameter”

Voorbeeld

- borstkanker herkennen op X-ray foto’s
- studie aan de UA in de 90’s
- neuraal netwerk wordt getraind
- werkt bijna perfect op de dataset
- werkt totaal niet in de praktijk...



Voorbeeld

- wat is er gebeurd?
- neuraal netwerk is heel intelligent
- het zag dat alle foto’s met kanker iets donkerder waren...
- dus: donkere foto -> kanker

Overfitting en underfitting

- als het algoritme “te goed” leert: dataset wordt vanbuiten geleerd, kan niet veralgemenen
  - dit heet **overfitting**
- als het algoritme “te slecht” leert: alles wordt veralgemeend, geen goed onderscheid meer
  - dit heet **underfitting**

Overfitting en underfitting

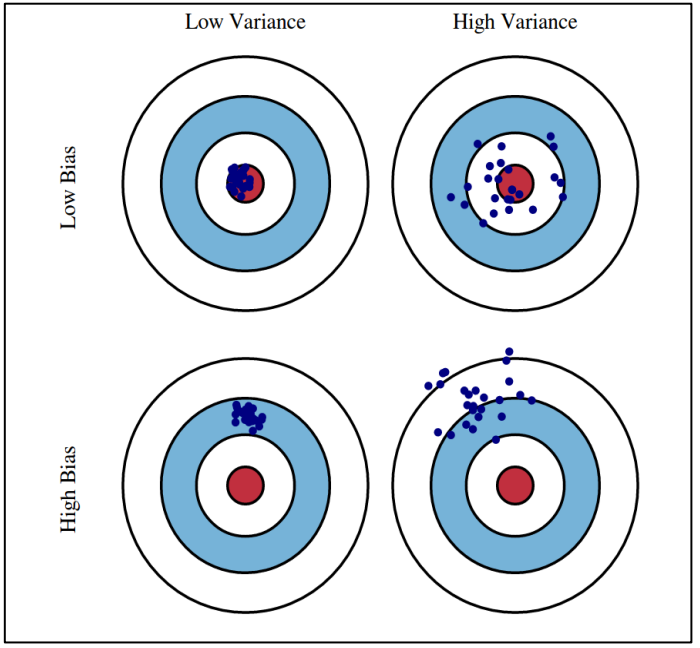
- dus: hoe intelligenter een algoritme, hoe meer fouten het maakt...
- hoe kunnen we weten of er sprake is van overfitting?

Oplossing 1: gebruik meerdere modellen

- bijvoorbeeld: maak verschillende beslissingsbomen met andere diepte
- **Bias** = het verschil tussen de gemiddelde voorspelling van alle modellen en de echte waarde
- **Variance** = het (gekwadrateerde) verschil tussen de voorspelling van 1 model en de gemiddelde voorspelling van alle modellen
- doel: bias en variance zo klein mogelijk

Gebruik meerdere modellen

- iedere stip stelt een model voor
- het midden is de werkelijke waarde



Balanceren

Lagere complexiteit

Lagere variance → geen overfitting

Het model is minder gevoelig voor kleine veranderingen in de data, wat leidt tot een betere generalisatie naar nieuwe data.

Hogere bias → wel underfitting

Het model kan de complexiteit van de data niet volledig modelleren, wat leidt tot minder nauwkeurige voorspellingen.

Hogere complexiteit

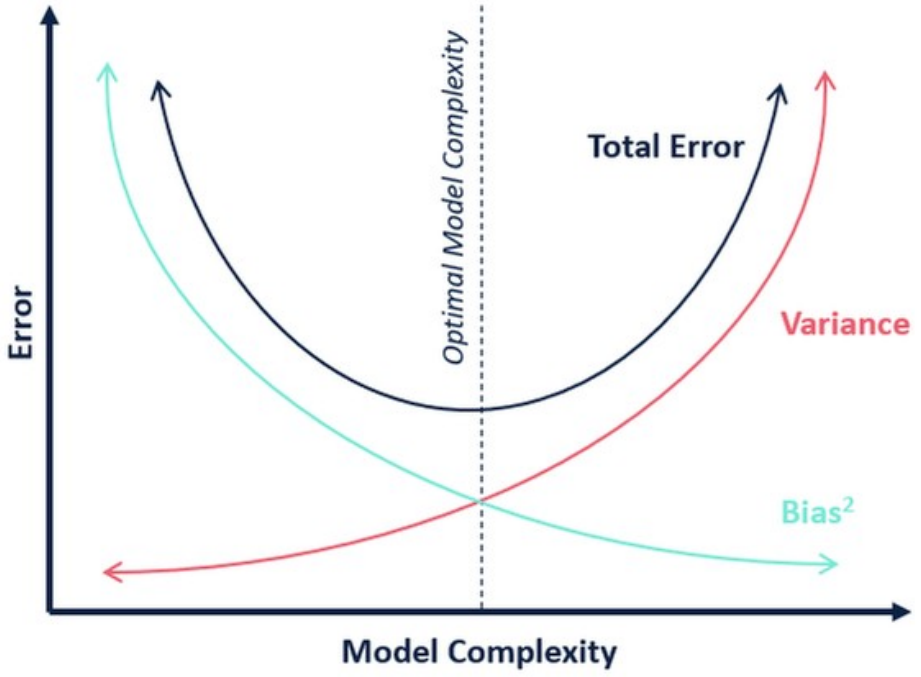
Hogere variance → wel overfitting

Het model is gevoeliger voor kleine veranderingen in de data, wat kan leiden tot overfitting.

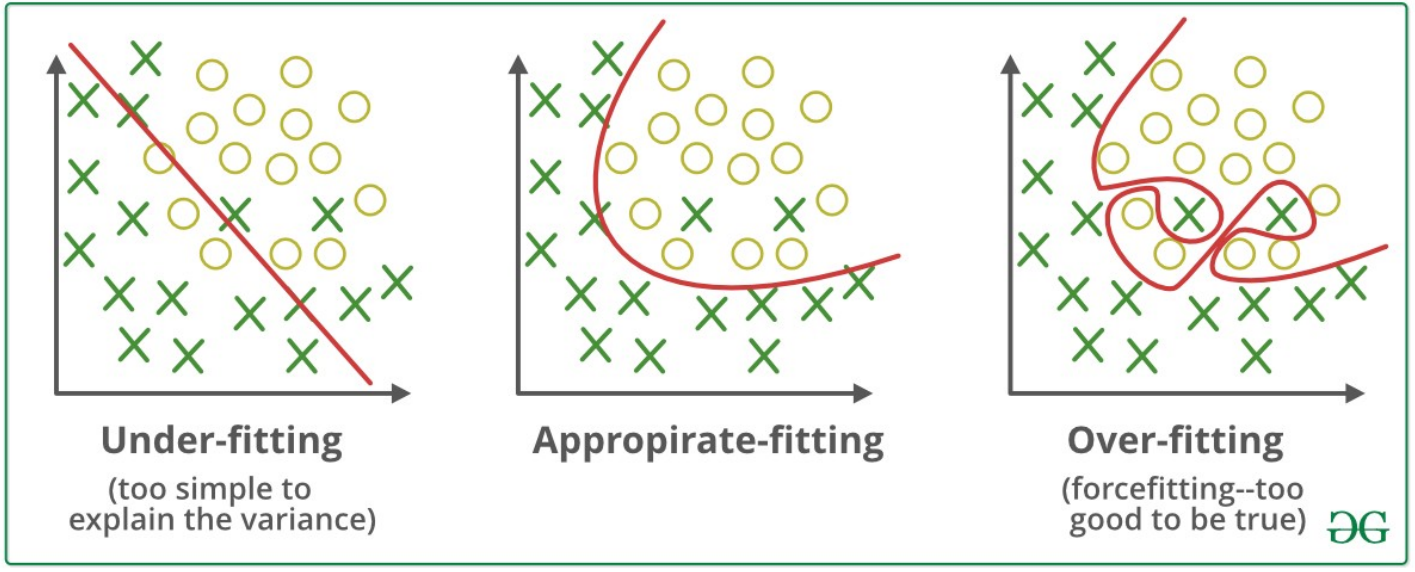
Lagere bias → geen underfitting

Het model kan complexere relaties in de data modelleren, wat leidt tot een nauwkeurigere voorspelling.

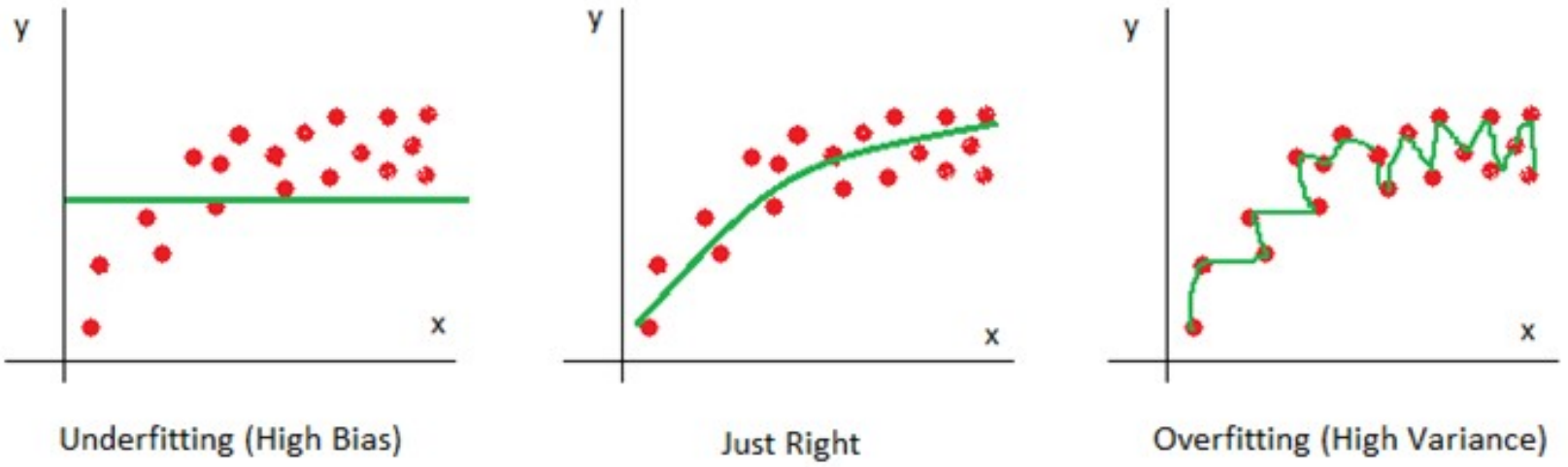
Balanceren


Optimal Model Complexity
Total Error
Variance
Bias²
Error
Model Complexity

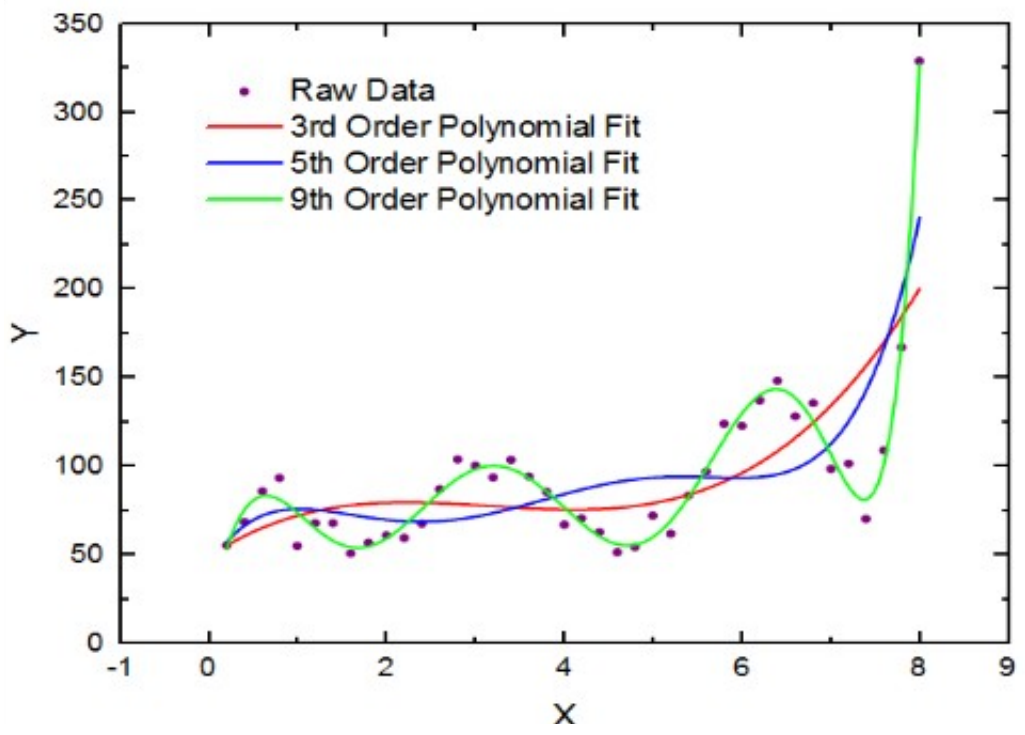
Overfitting bij classificatie



Overfitting bij regressie



Overfitting bij regressie



Training en test dataset

2

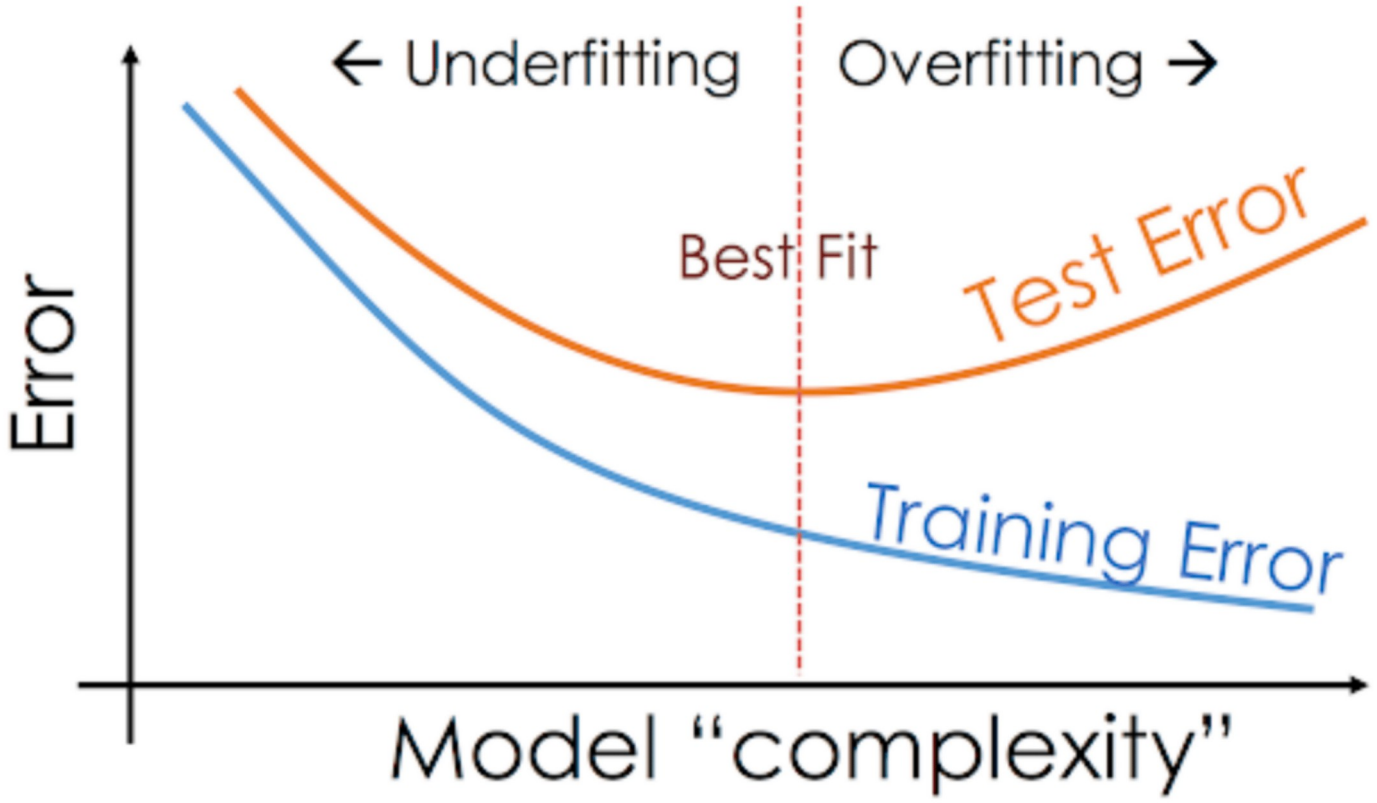
Oplossing voor overfitting?

- vergelijk: studeren voor een vak: wat als je alles vanbuiten leert?
- gaan we op het examen letterlijke oefeningen vragen of verzinnen we er nieuwe?

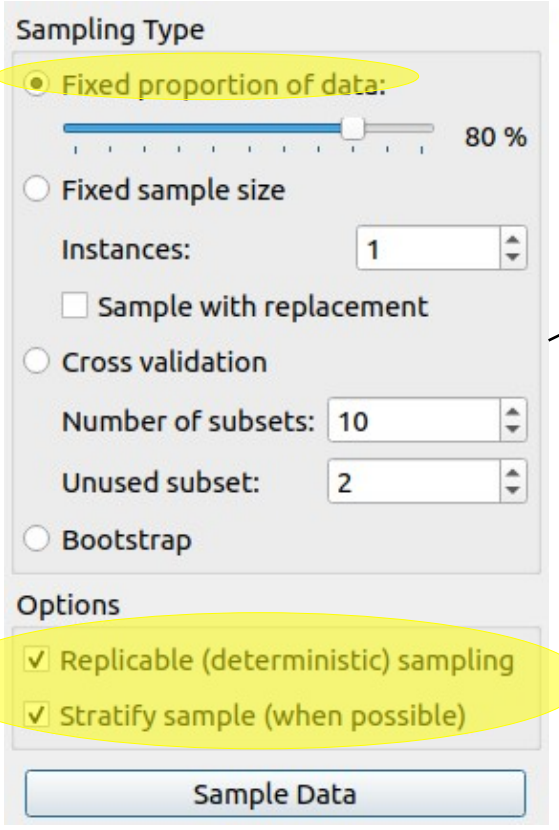
Oplossing

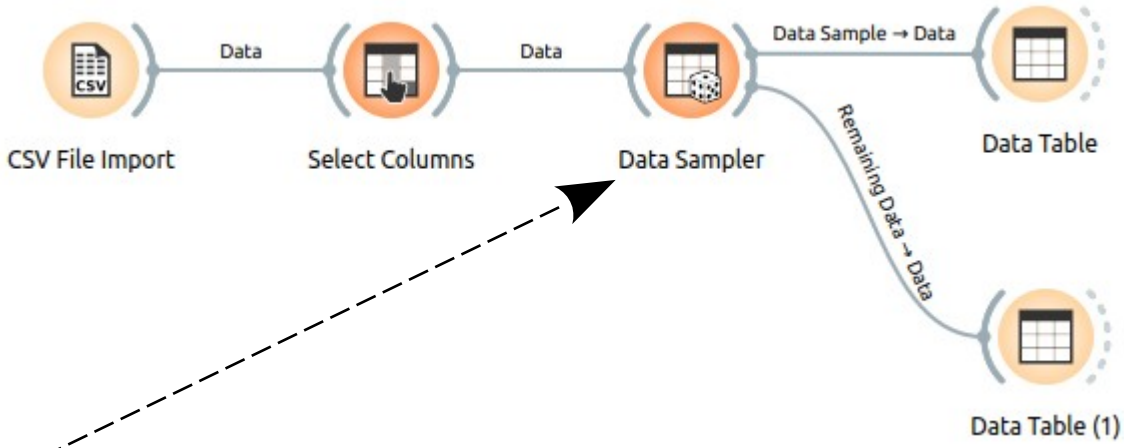
- we hebben nood aan een test dataset!
- splits data op in 2 delen:
  - training dataset (70%-80%)
  - test dataset (20%-30%)
- we trainen het model met de training dataset en gaan meten hoe goed het model werkt op de test dataset
- men noemt dit de “holdout” methode
- we hebben “metrieken” nodig om de prestatie van een model te meten (zie volgende slidesets)

Oplossing

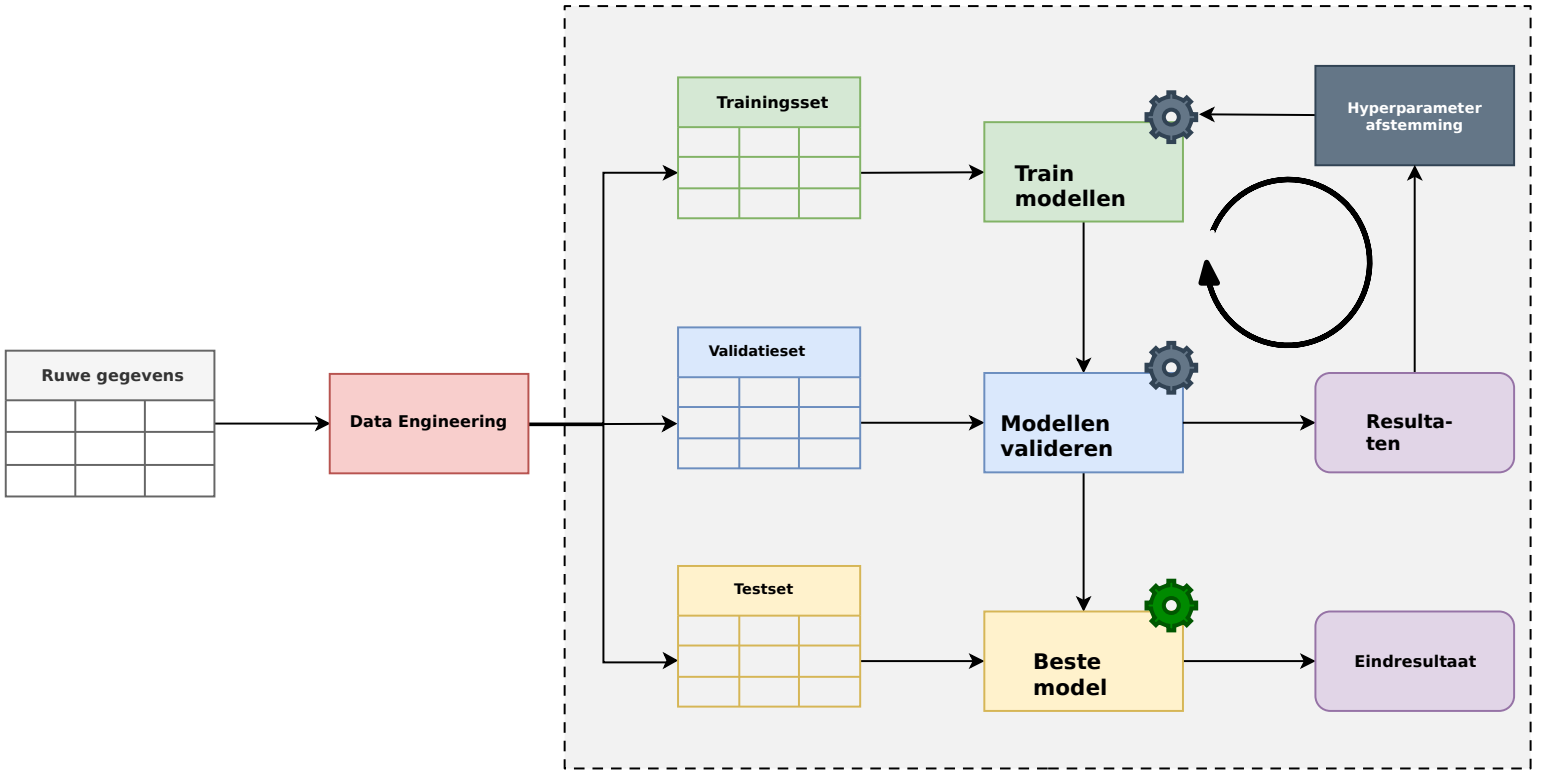

← Underfitting
Overfitting →
Best Fit
Test Error
Training Error
Error
Model "complexity"

Orange


Sampling Type
Fixed proportion of data:
80 %
Fixed sample size
Instances:
1
Sample with replacement
Cross validation
Number of subsets:
10
Unused subset:
2
Bootstrap
Options
Replicable (deterministic) sampling
Stratify sample (when possible)
Sample Data


Data
Data
Data Sample → Data
Remaining Data → Data
CSV File Import
Select Columns
Data Sampler
Data Table
Data Table (1)

Later: “modelvalidatie”

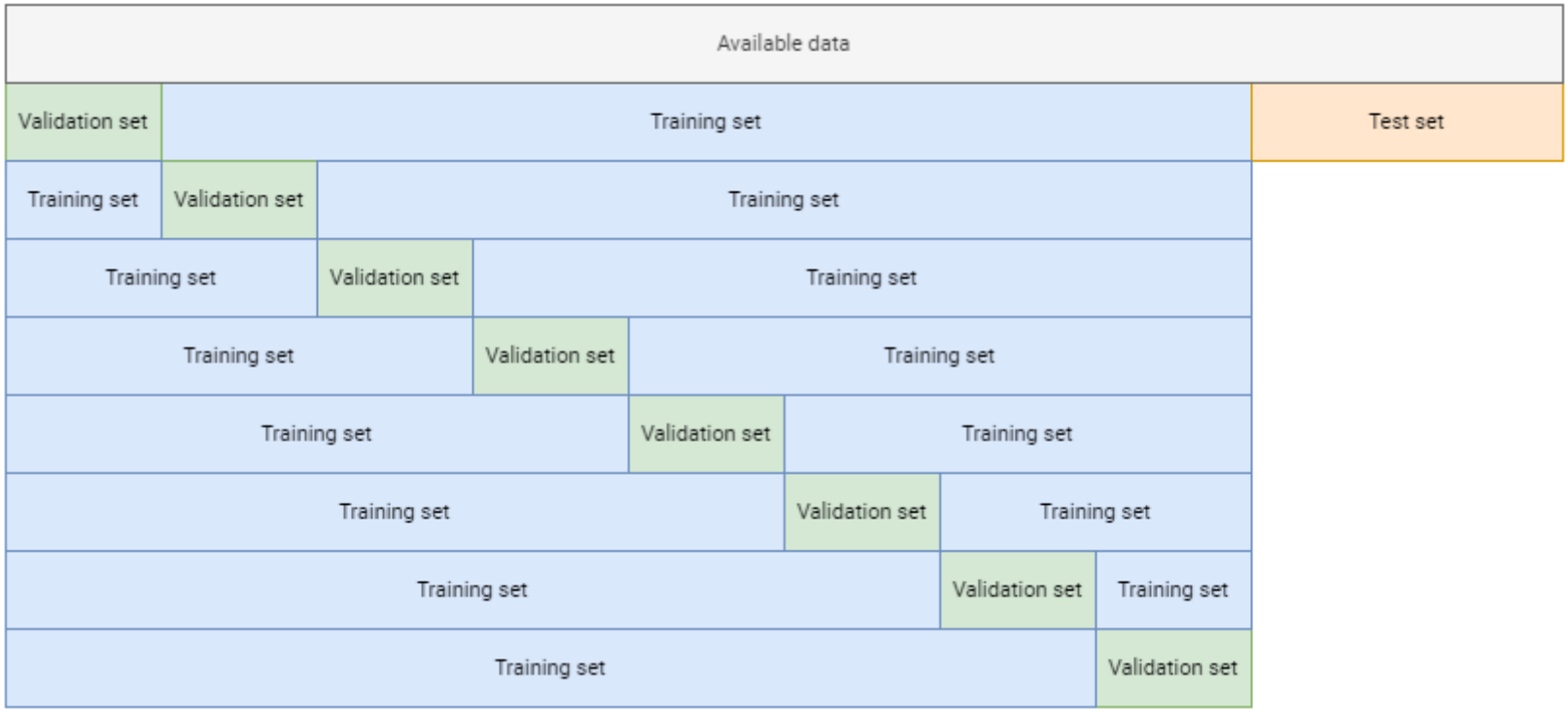


Modelvalidatie

Later: cross validation

- partitioneer de data in blokken
- doorloop de blokken:
  - gebruik dit blok om te testen en de andere om te trainen
  - valideer het model
- neem het gemiddelde van alle resultaten

Later: cross validation



Besluit

- data ALTIJD opsplitsen in training data en test data
- model genereren op training data
- model testen met test data
  - maar hoe testen we???
- twee verschillende technieken:
  - regressie
  - classificatie