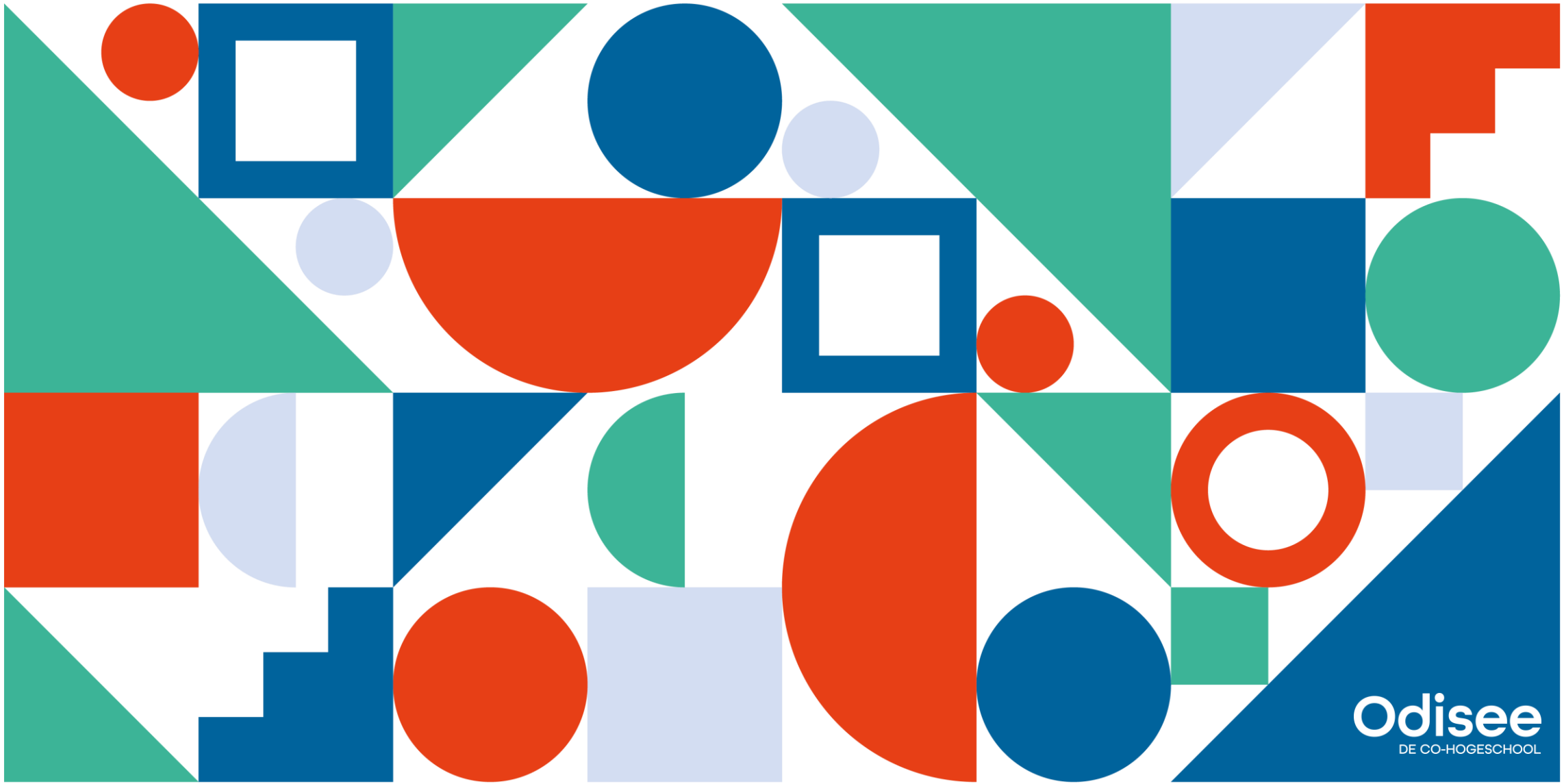

Odisee
DE CO-HOGESCHOOL

Data Science – week 6

Jens Baetens

Modelling

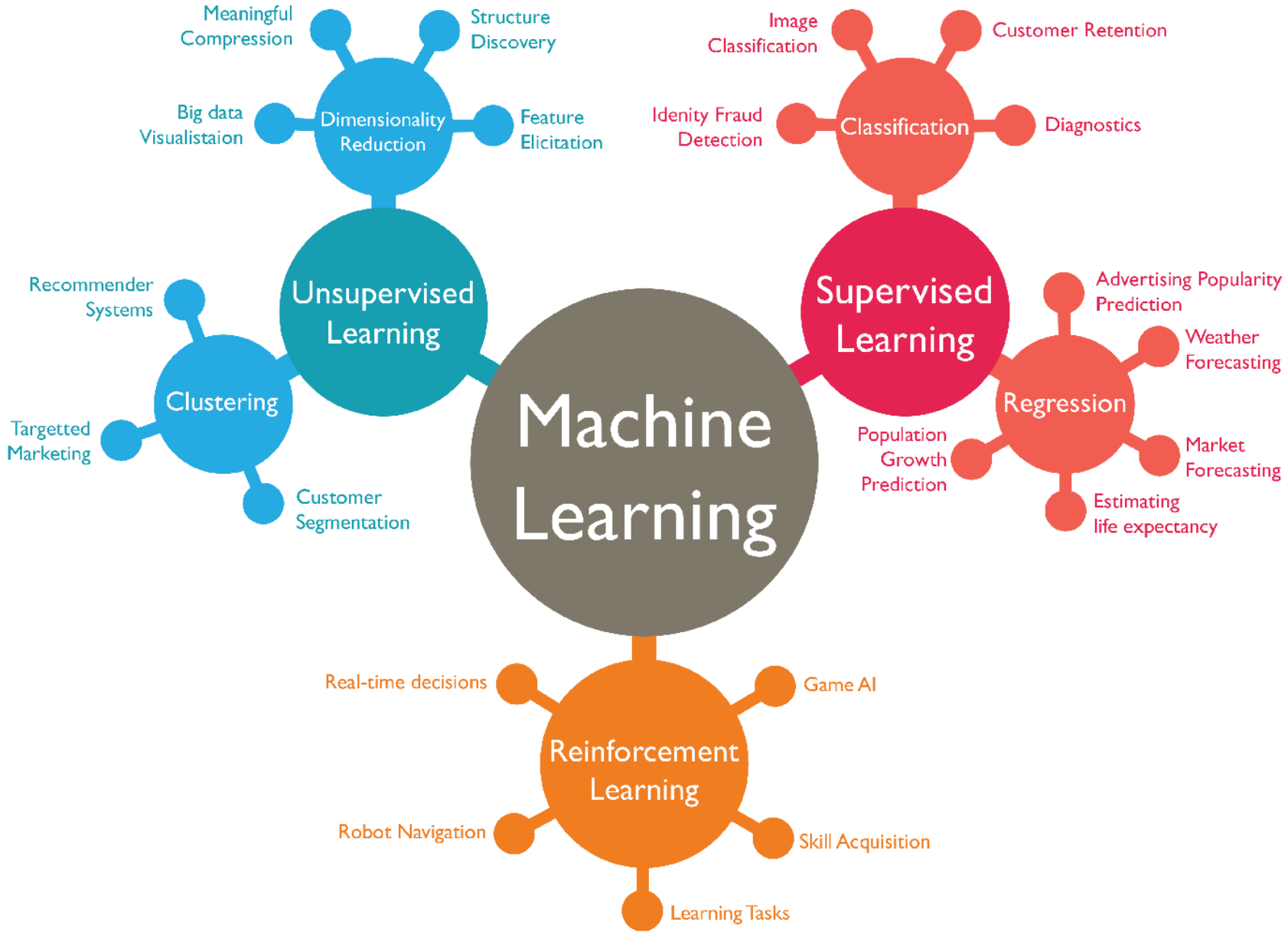
Wat is data modelling?

- Bruikbare data omzetten naar een model
  - Dit model zet inputs om naar een gewenste output
- Doel
  - Inzichten
  - Voorspellingen
  - Beslissingsvorming
  - Duidelijkere communicatie

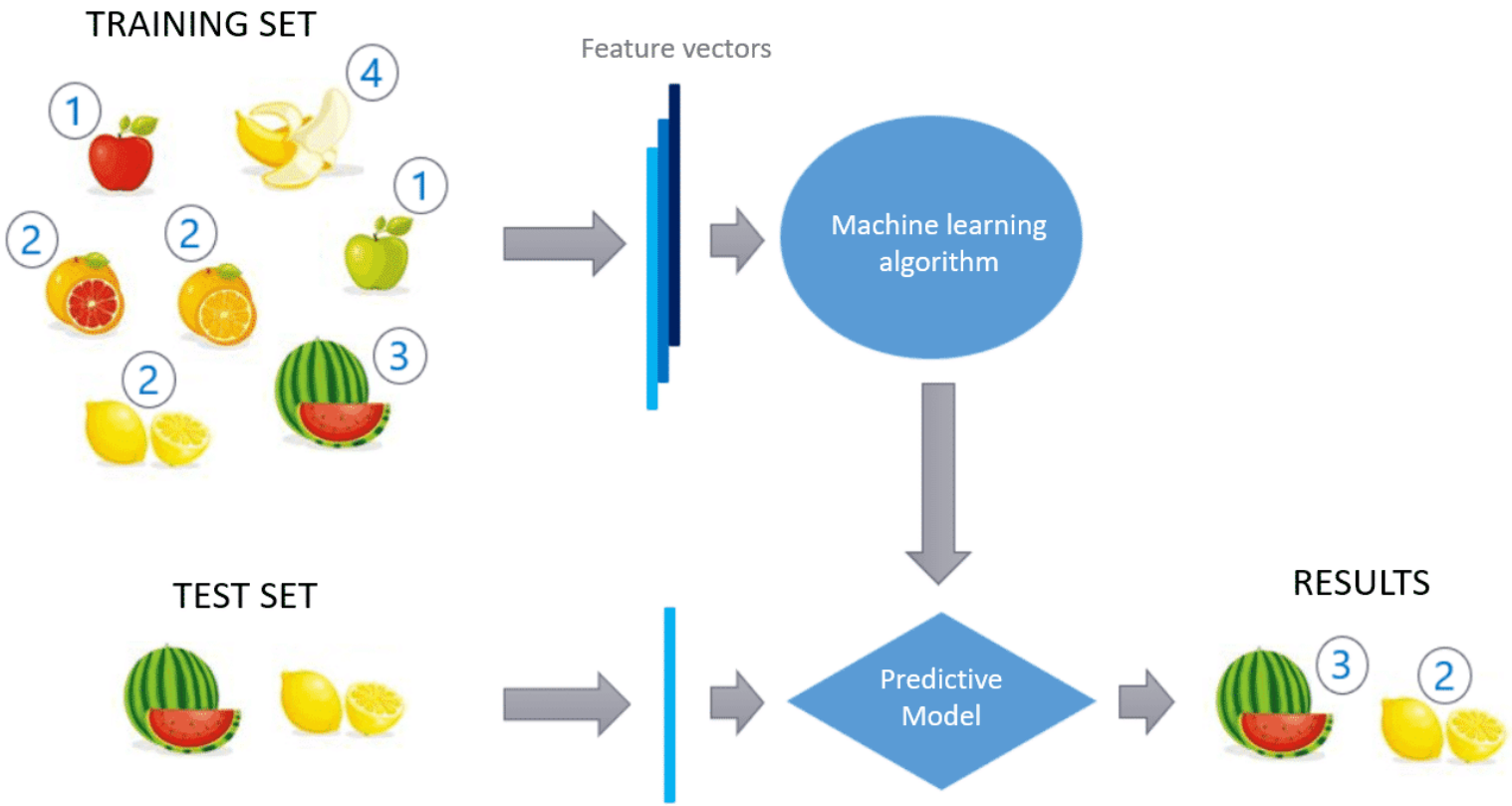
Typisch verloop van data modelling

- 1: Kies een ML-techniek om je doel te bereiken
  - Configureer de techniek met de gewenste hyperparameters
- 2: Train het model
  - Pas interne parameters/gewichten van het model aan aan de data
  - Optimaliseren van een loss-functie
- 3: Evalueer het model
  - Gebruik het model op nieuwe data en evalueer de bekomen resultaten

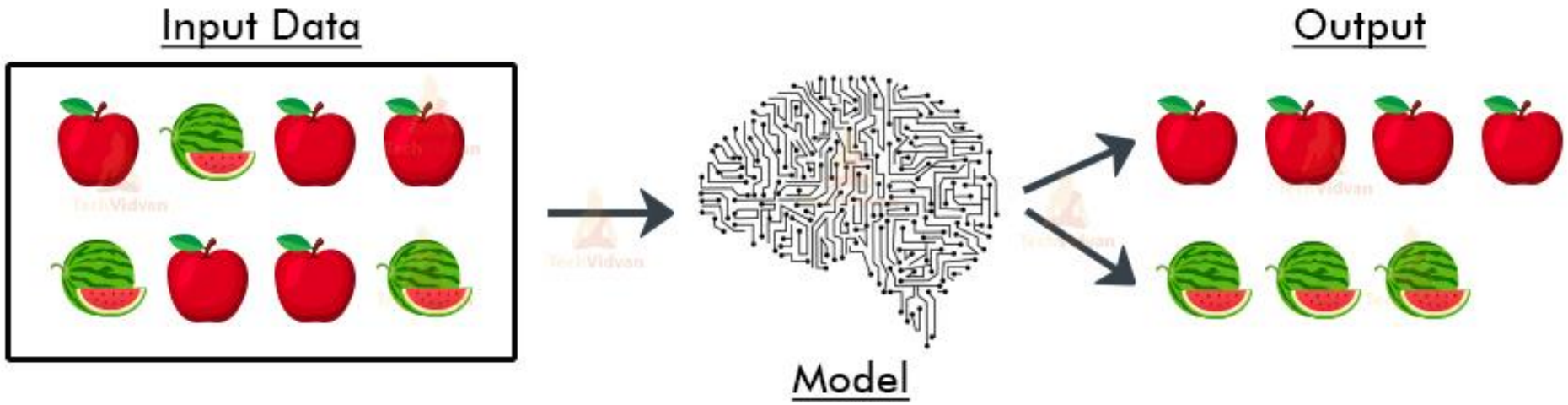
```python
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error

# Stap 1: Gegevens inladen en splitsen in features en labels
X = ...  # Features (input data)
y = ...  # Labels (target values)

# Stap 2: Data opsplitsen in trainings- en testsets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Stap 3: Model instantiëren en trainen
model = LinearRegression()
model.fit(X_train, y_train)

# Stap 4: Voorspellingen doen op de testset
y_pred = model.predict(X_test)

# Stap 5: Evalueren van het model
mse = mean_squared_error(y_test, y_pred)
print(f"Mean Squared Error: {mse}")

# Je kunt andere evaluatiemetrics toevoegen afhankelijk van je behoeften
```


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Supervised learning



Unsupervised learning



Supervised vs unsupervised

Reinforcement learning

- Modellen leren uit goed gedrag/score
  - Doel om score zo hoog mogelijk te krijgen
  - Geen data gebruikt maar wel spelregels

Reinforcement learning

ML-technieken samenvatting

Gebruik van data en labels	Supervised ML
Gebruik van data	Unsupervised ML
Gebruik van regels	Reinforcement ML

Regressie

Regressie

- Voorspel een continue output/target/label kolom op basis van inputs/features
  - 1 input-kolom -> enkelvoudige regressie
  - Meerdere kolommen -> meervoudige regressie

- Regressor

Inputs

Labels

Avg. Area I...	Avg. Area ...	Avg. Area ...	Avg. Area ...	Area Popul...	Price	Address
79545.45857431678	5.682861321615587	7.009188142792237	4.09	23086.80050268656	1059033.5578701235	208 Michael Ferry Apt. 674 Laurabury, NE 37010-5101
79248.64245482568	6.002899808275242 5	6.730821019094919	3.09	40173.07217364482	1505890.91484695	188 Johnson Views Suite 079 Lake Kathleen, CA 48958
61287.067178656784	5.865889840310001	8.512727430375099	5.13	36882.15939970458	1058987.9878760849	9127 Elizabeth Stravenue Danieltown, WI 06482-3489
63345.24004622798	7.188236094518642 5	5.586728664827653	3.26	34310.24283090706	1260616.8066294468	USS Barnett FPO AP 44820
59982.197225708034	5.040554523106283	7.839387785120487	4.23	26354.109472103148	630943.4893385402	USNS Raymond FPO AE 09386

Evaluatie van regressie

Gemiddelde kwadratische fout

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Gemiddelde absolute fout

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Determinatiecoëfficiënt

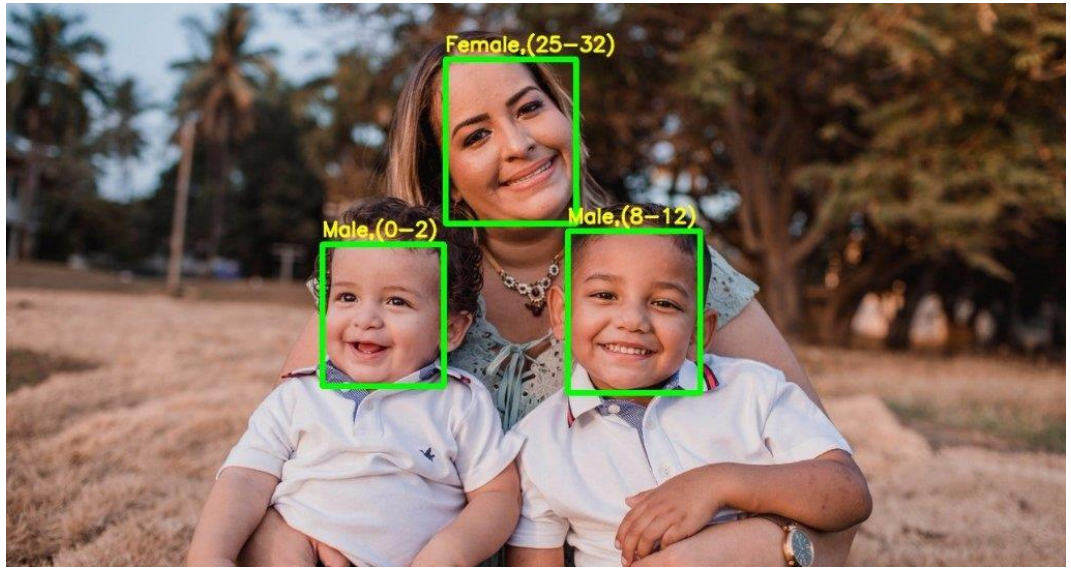
$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Classificatie

Classificatie

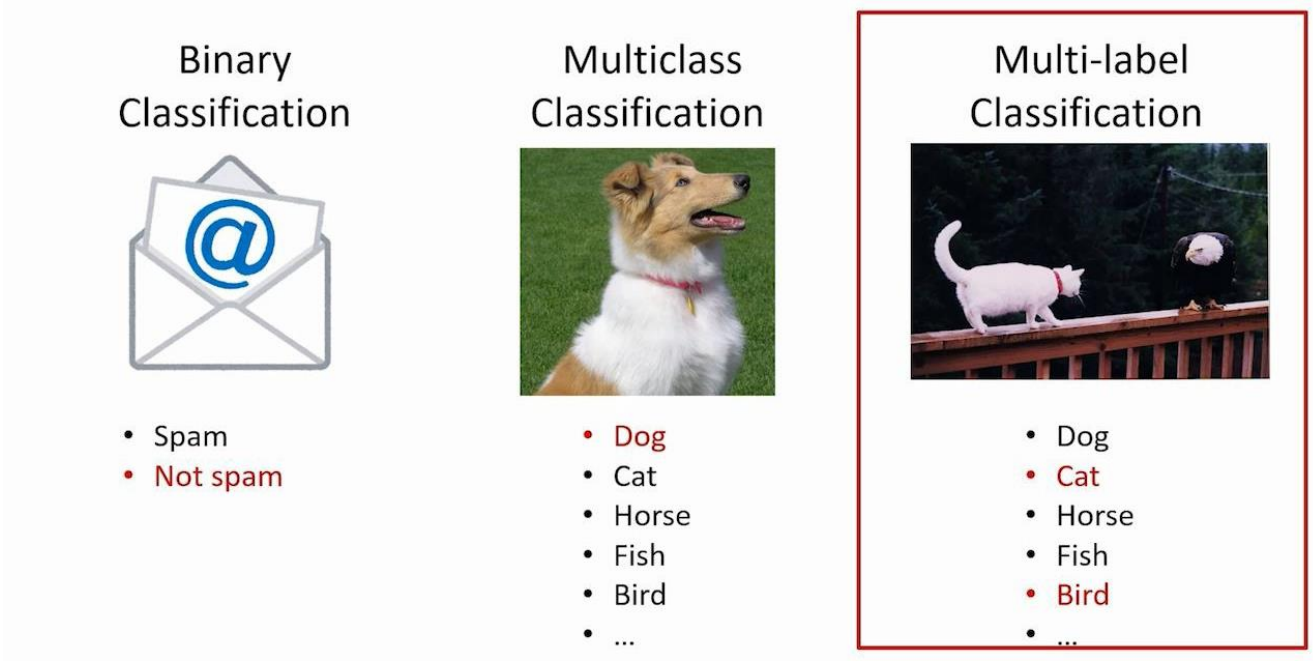
- Ken een klasse toe op basis van 1 of meerdere inputs

- Classifier
  - Spam detective
  - Tekstherkenning
  - Medische diagnoses
  - Gezichtsherkenning



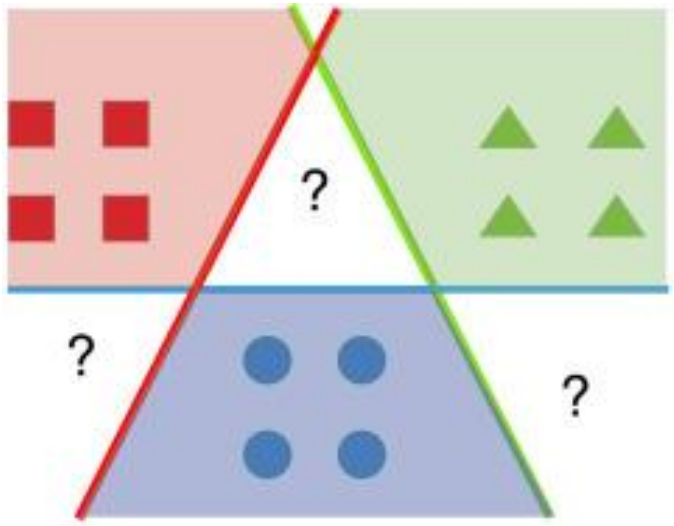
Classificatie

- Types classifier
  - Binary -> Twee klassen
  - Multi-class -> Meerdere klassen
  - Multi-label -> Meedere klassen tegelijkertijd mogelijk

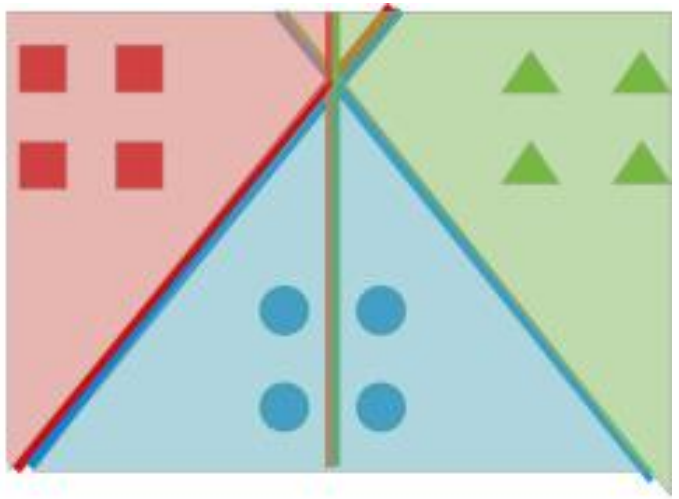


Classificatie

One-vs-All

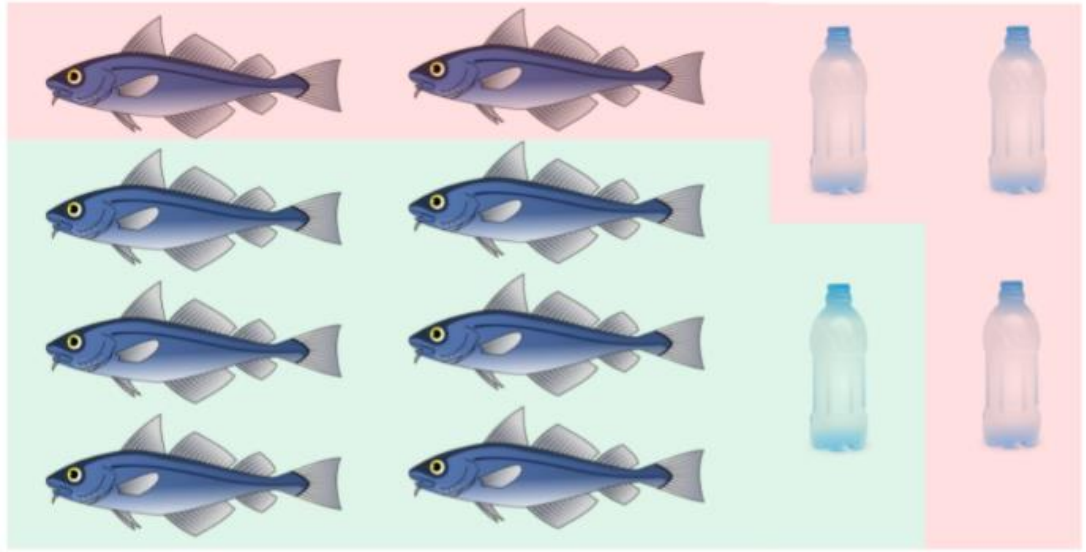


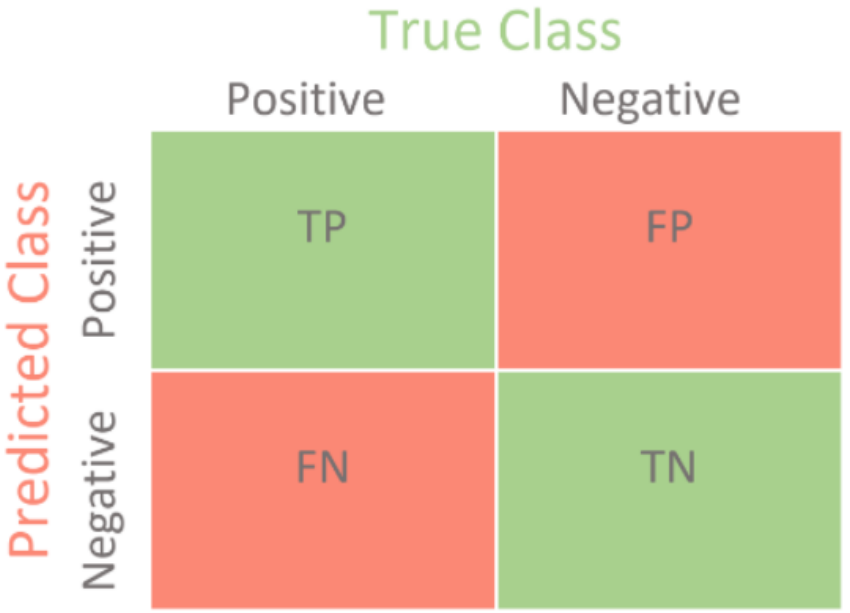
One-vs-One



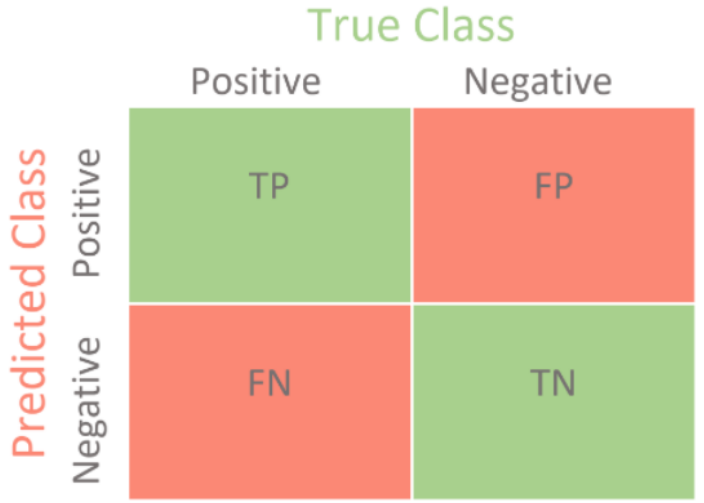
Evaluatie van classificatie

- Confusion-matrix





Evaluatie van classificatie



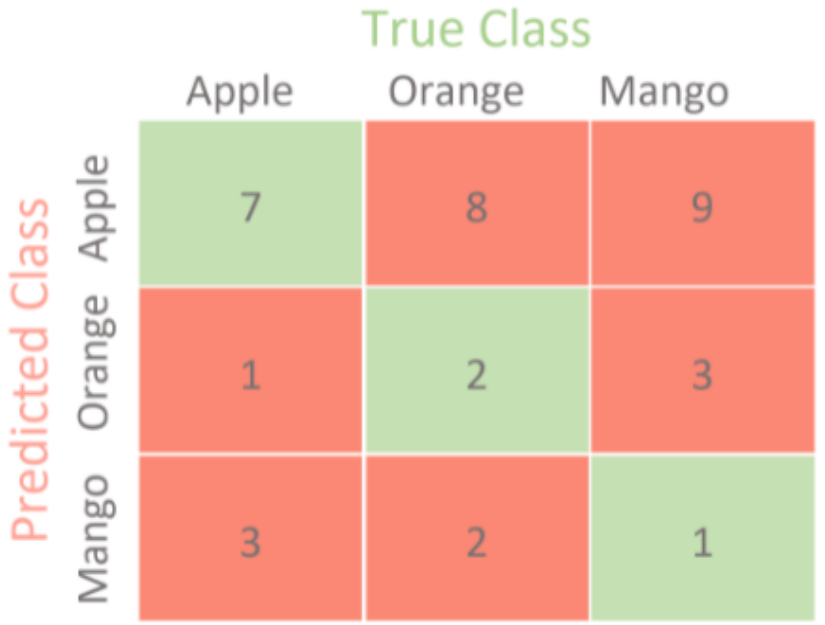
Accuraatheid	(TP + TN) / (TP + TN + FP + FN)
Precisie	TP / (TP + FP)
Specificiteit	TN / (TN + FP)
Recall	TP / (TP + FN)

F1-Score	2 * Precision * Recall / (Precision + Recall)
	2 * TP / (2 * TP + FP + FN)

Evaluatie van classificatie

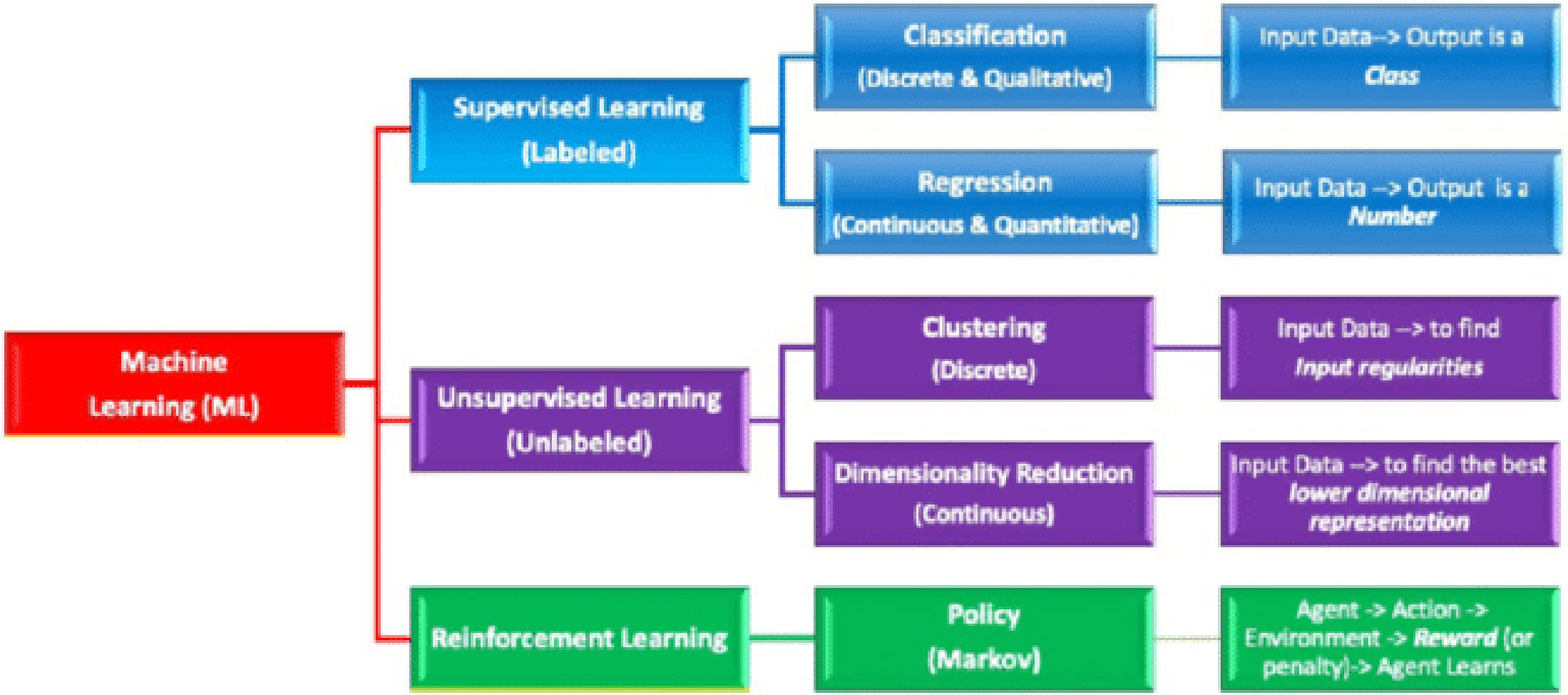
- Accuraatheid
  - Algemeen beeld maar gevoelig voor ongebalanceerdheid
- Precision
  - Kost van false-positive hoog
- Recall
  - Kost van false negative hoog
- F1-score
  - Algemeen beeld, rekening houdend met de gebalanceerdheid van de klassen

Wat in het geval met meerdere Klassen?



- Micro scores
  - Bereken confusion matrix voor correct/fout voorspeld
  - Geen verschil tussen precision, recall, accuraatheid
- Macro scores
  - TP/FP/… per klasse apart
  - Score is het (ongewogen) gemiddelde van de scores per klasse
- Weighted scores
  - Zoals macro-scores maar gewogen gemiddelde

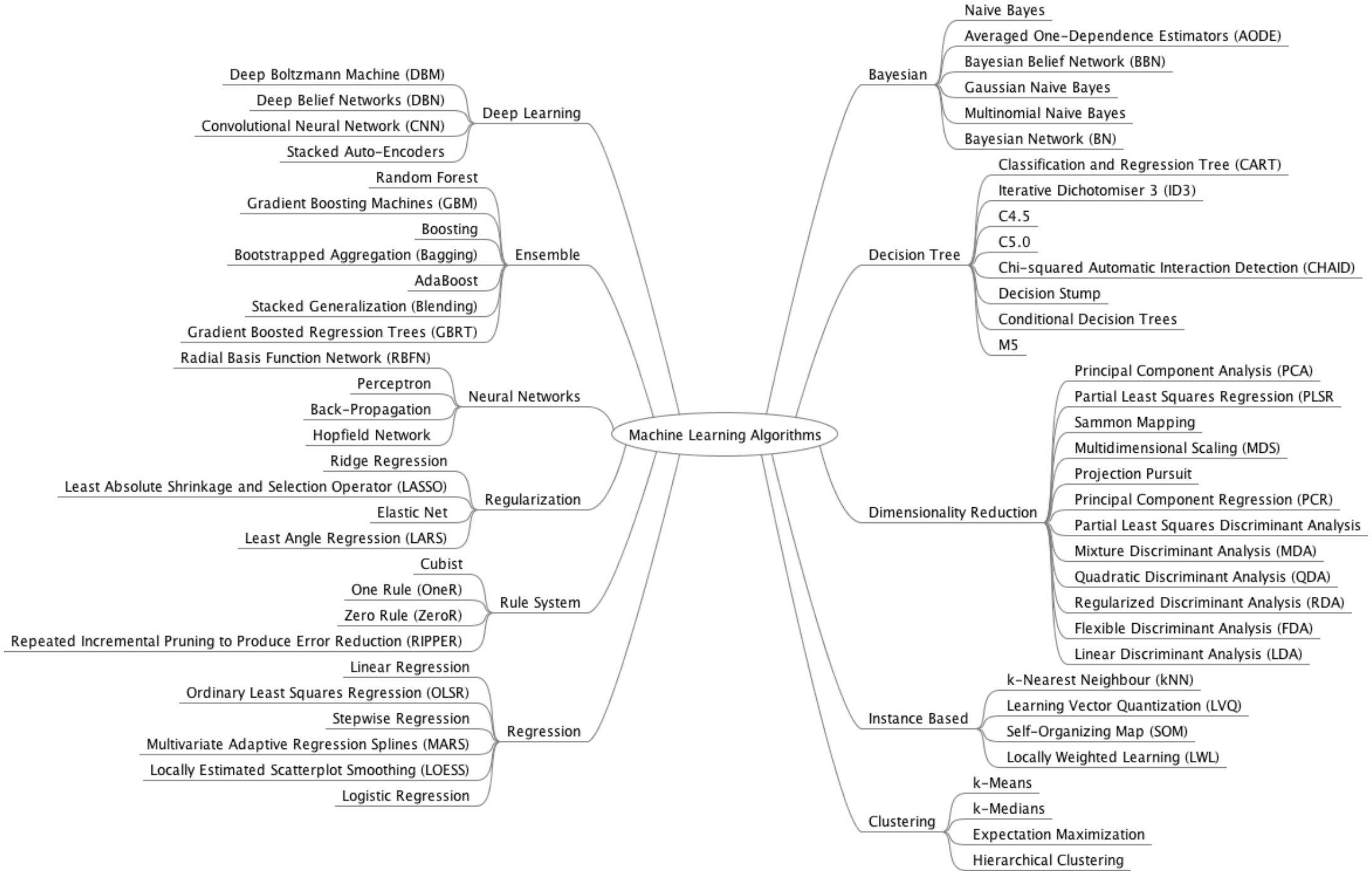
Wanneer welke techniek kiezen?



Wanneer welke techniek kiezen?

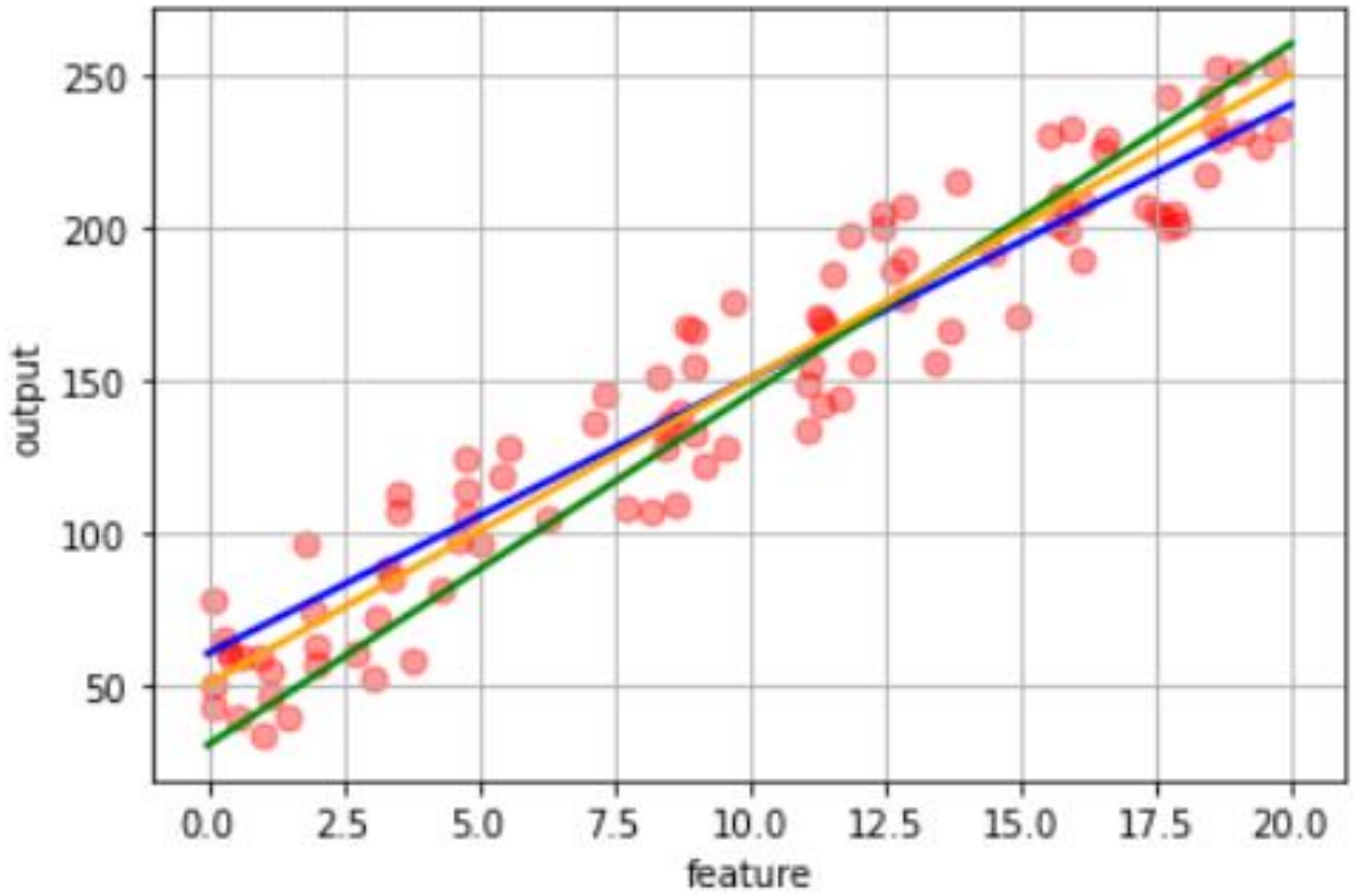
Hoeveel gaat het zijn?	→ Regression techniques
Is het type A of B?	→ Classification techniques
Zijn deze gelijk?	→ Clustering techniques
Is het vreemd?	→ Anomaly detection techniques
Wat moet ik doen?	→ Recommendation techniques
Welke zijn het belangrijkst?	→ Dimensionality reduction

ML - technieken


Machine Learning Algorithms
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression

Lineaire regressie

Wat is de beste rechte?

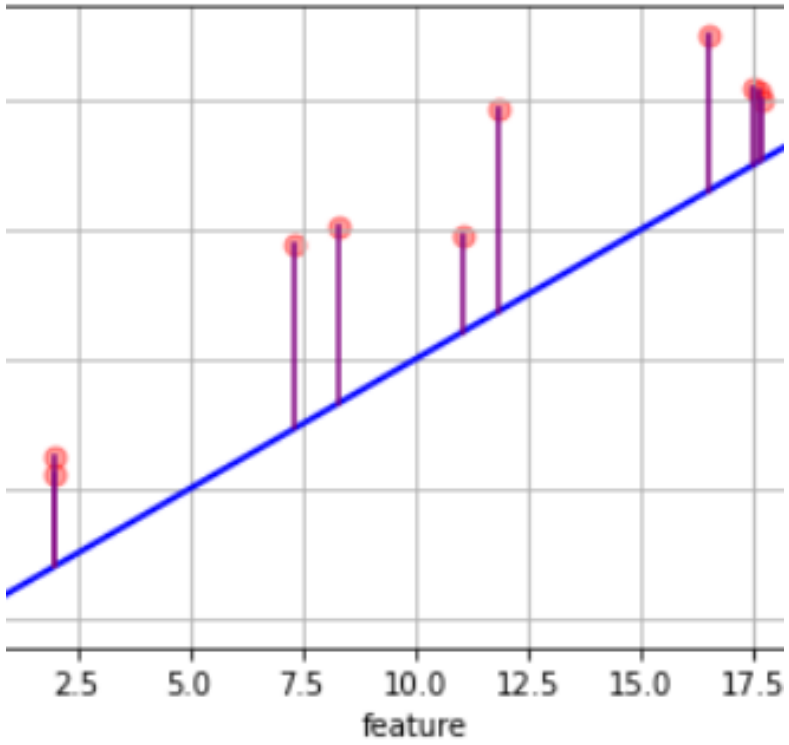


Lineaire regressie

- Beste rechte heeft de formule:
  - $f_{\mathbf{w}}(x) = w_0 + w_1 x =$ **target**
- Lineaire regressie gaat op zoek naar de beste waarden voor w_0 en w_1
  - w_0 en w_1 zijn de gewichten/parameters
  - Worden getrained/aangepast tot het ideale

Lineaire regressie

- Beste rechte heeft de kleinste fout
  - Fout bepaald door de loss-functie
  - Least Mean Squares
    - Least -> kleinste fout
    - Mean -> gemiddeld over alle datapunten
    - Squares -> kwadratische afstand
    - $L(\mathbf{w}) = \frac{1}{2N} \sum_{i=1}^{N} (f_{\mathbf{w}}(x^i) - y^i)^2$
  - Andere varianten mogelijk

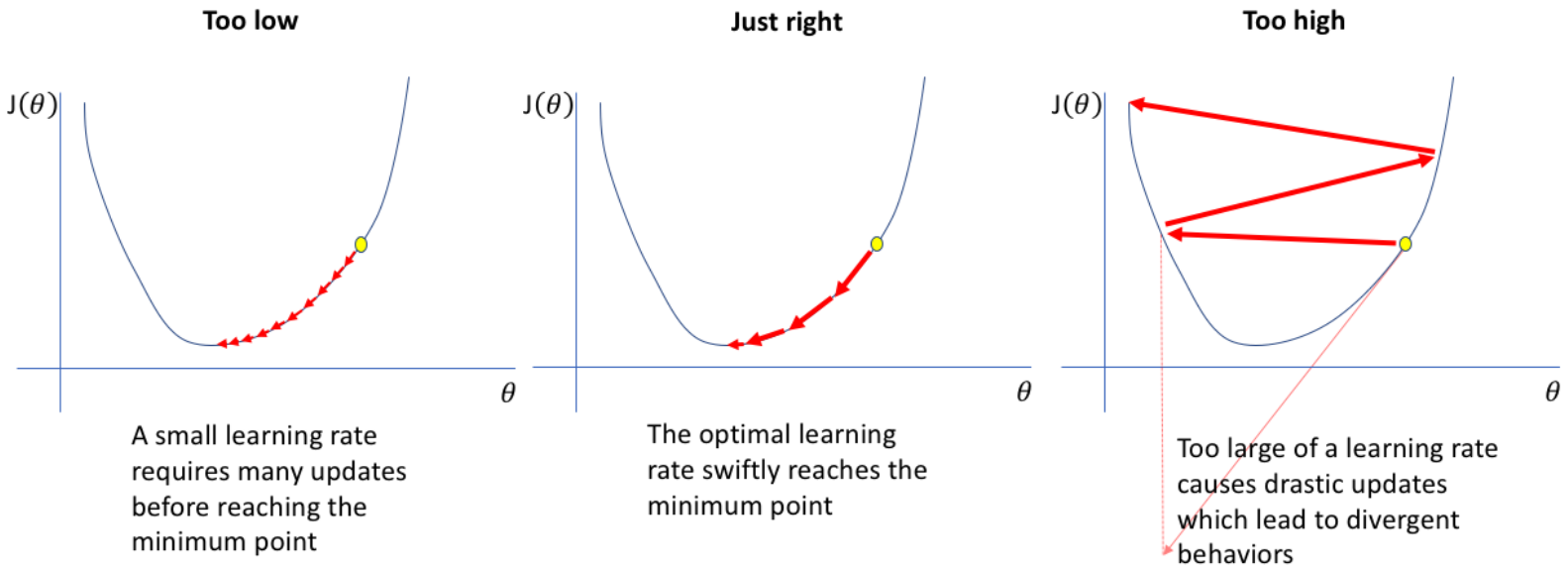


Wat zijn de beste parameters?

- Gradient descent
  - Afgeleidde van de loss-functie
  - Zoek naar het globale minimum
    - Is altijd zo bij convexe functies

Gradient Descent – Learning rate

- Learning rate bepaalt de stapgrootte
  - Heeft een grote impact op het vinden van een goed resultaat
  - Adam-optimizer past het automatisch aan



Meervoudige/multivariate regressie

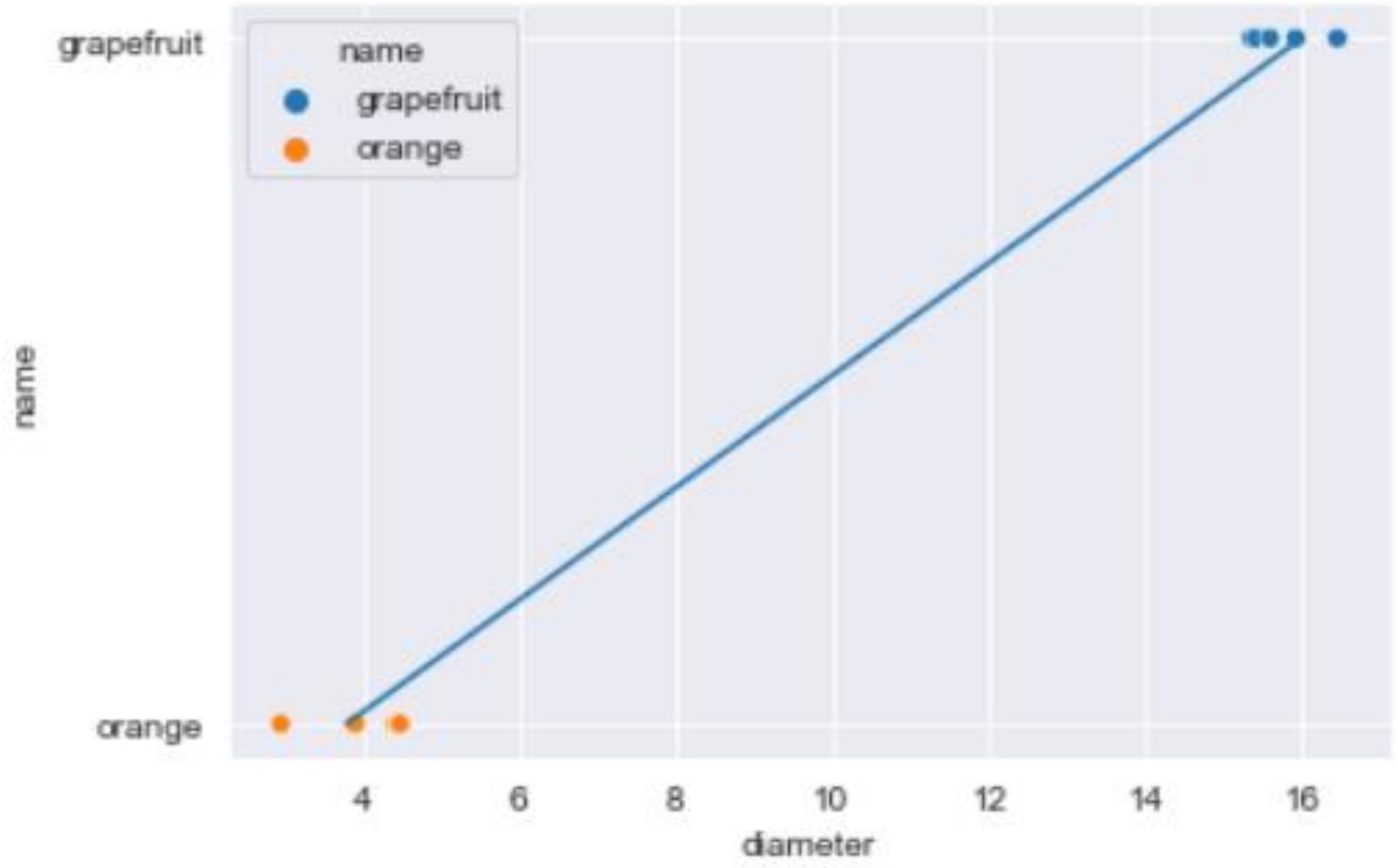
- Extra gewichten per feature
- Hoeveel hebben we er nodig?

Meervoudige/multivariate regressie

- $f_{\mathbf{w}}(x) = w_0 + w_1 x =$ **target**
- $\mathbf{w} = [w_0 , w_1]$
- $L(\mathbf{w}) = \frac{1}{2N} \sum_{i=1}^{N} (f_{\mathbf{w}}(x^i) - y^i)^2$

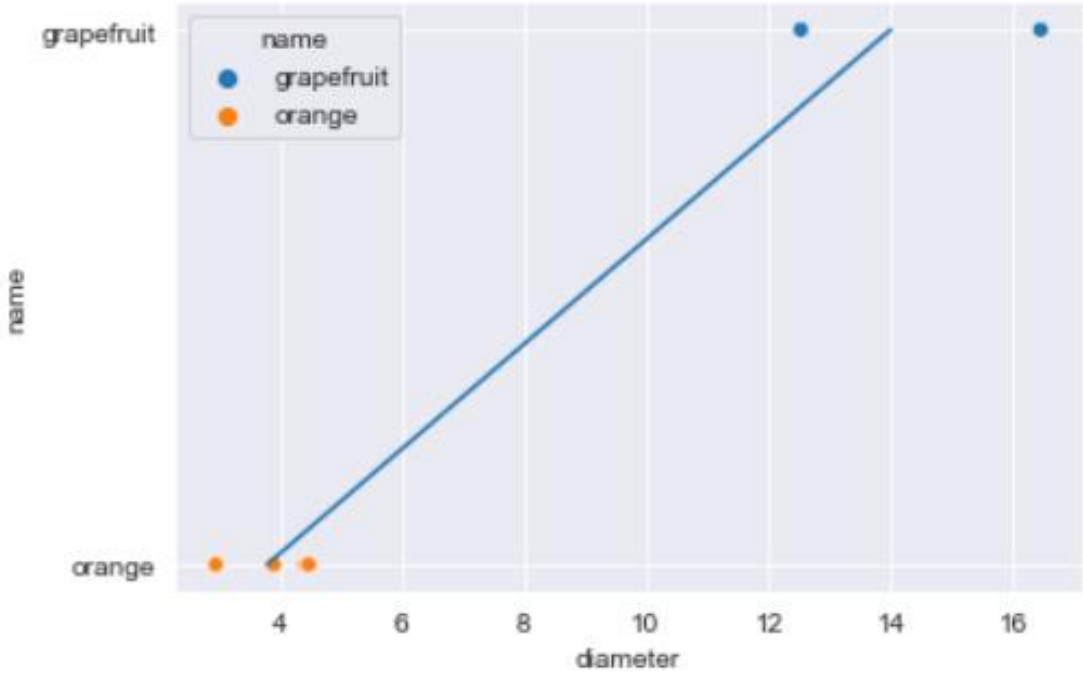
Logistische regressie

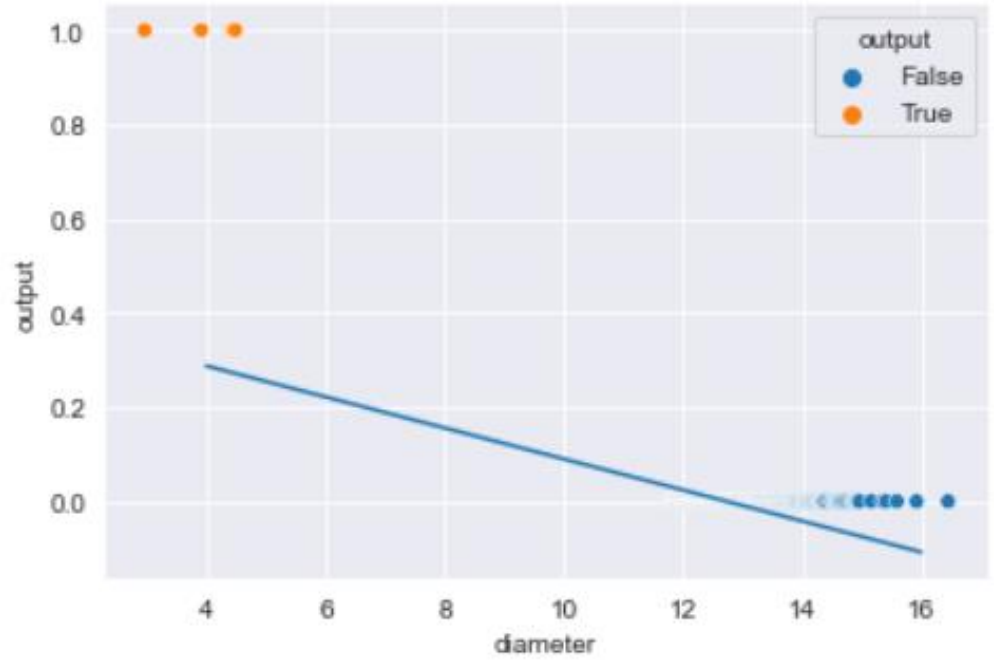
Kan classificatie met logistische regressie?



Kan classificatie met logistische regressie?

- Gevoelig voor
  - Outliers
  - Ongebalanceerde klassen
- Breed, onduidelijk midden
- Getal is geen kans



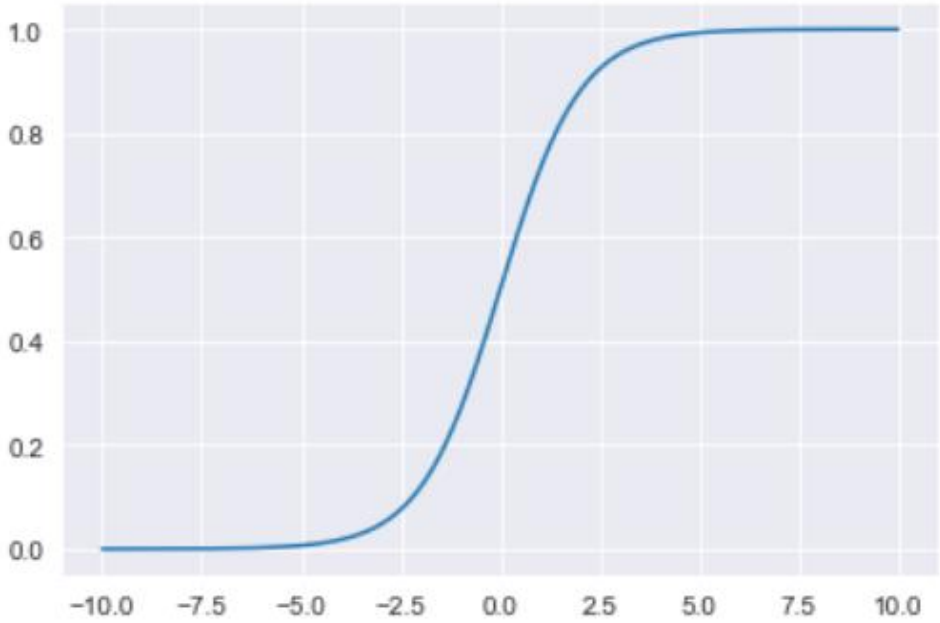


Logistische regressie

- Gebruik een sigmoid functie ipv lineaire
  - S-vorm
  - Waarde tussen 0 en 1 (kans)
  - Naam genoemd naar Belgische wiskundige

$$f_{\boldsymbol{w}}(x) = \frac{1}{1+e^{-\boldsymbol{w}^T x}}$$

$$\boldsymbol{w}^T x = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_N x_N$$



Wat met de loss-functie?

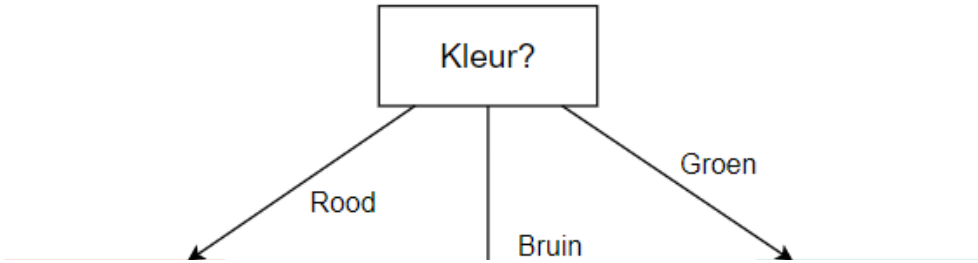
- MSE heeft geen betekenis
  - Beetje fout is even erg als heel erg fout

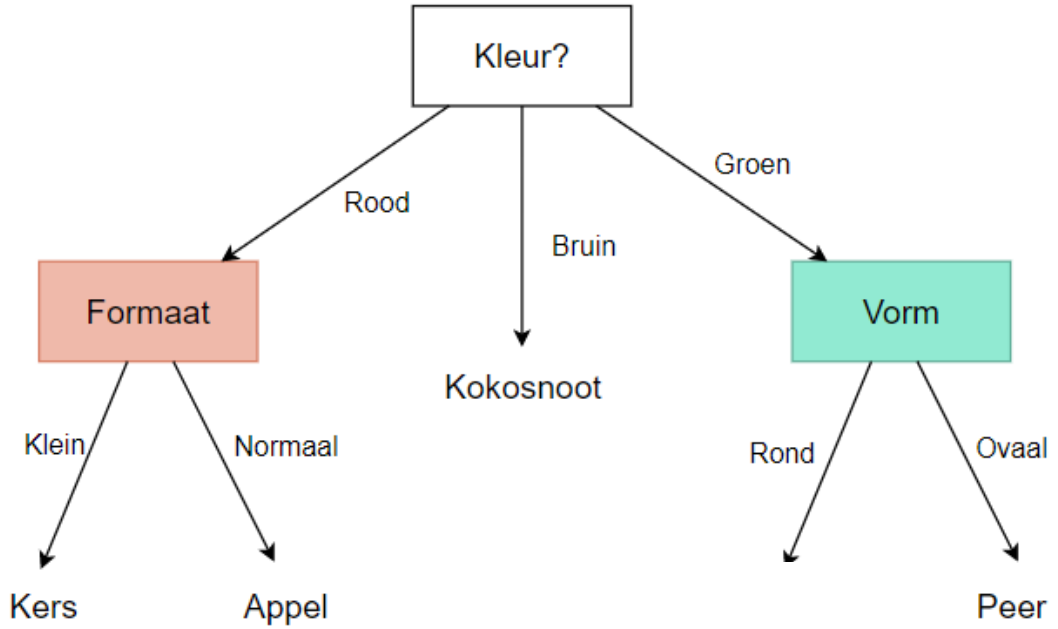
- Loss-functie op basis van correct voorspeld of niet

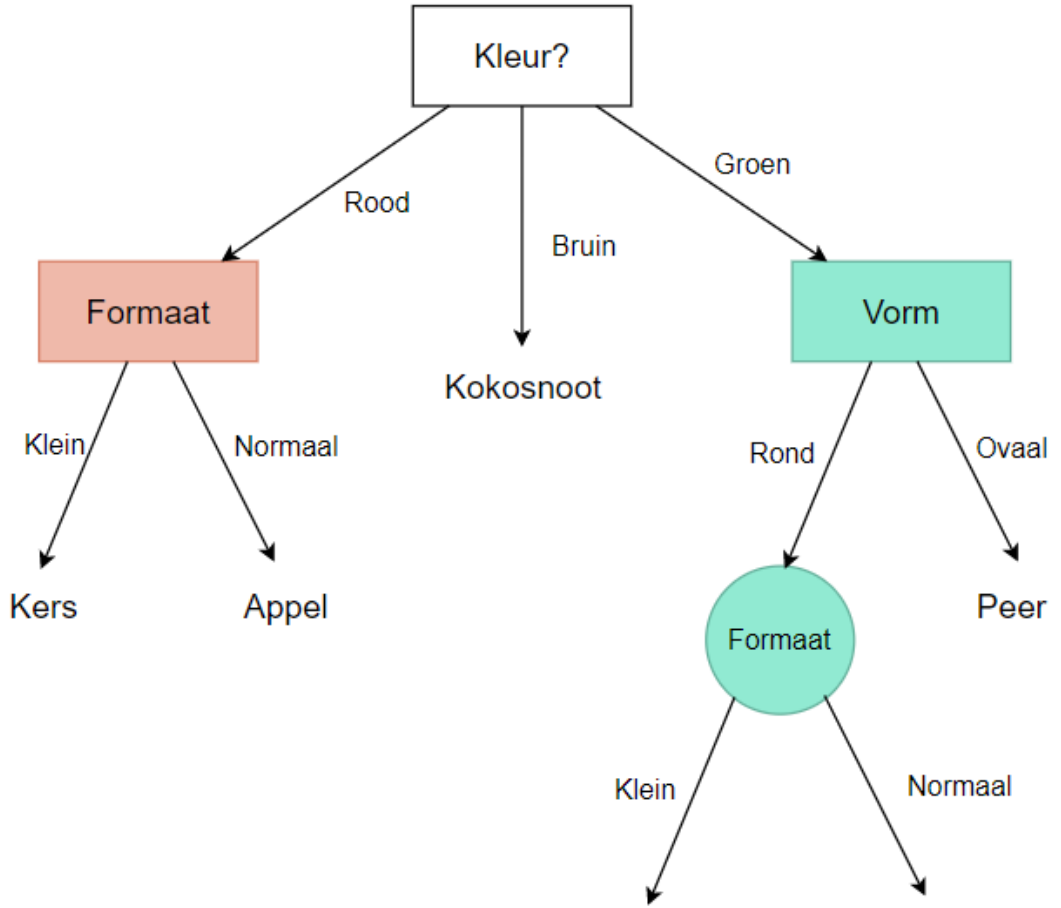
$$L(\boldsymbol{w}) = \begin{cases} -ln(f_{\boldsymbol{w}}(x)) & \text{als } y = 1 \\ -ln(1 - f_{\boldsymbol{w}}(x)) & \text{als } y = 0 \end{cases}$$

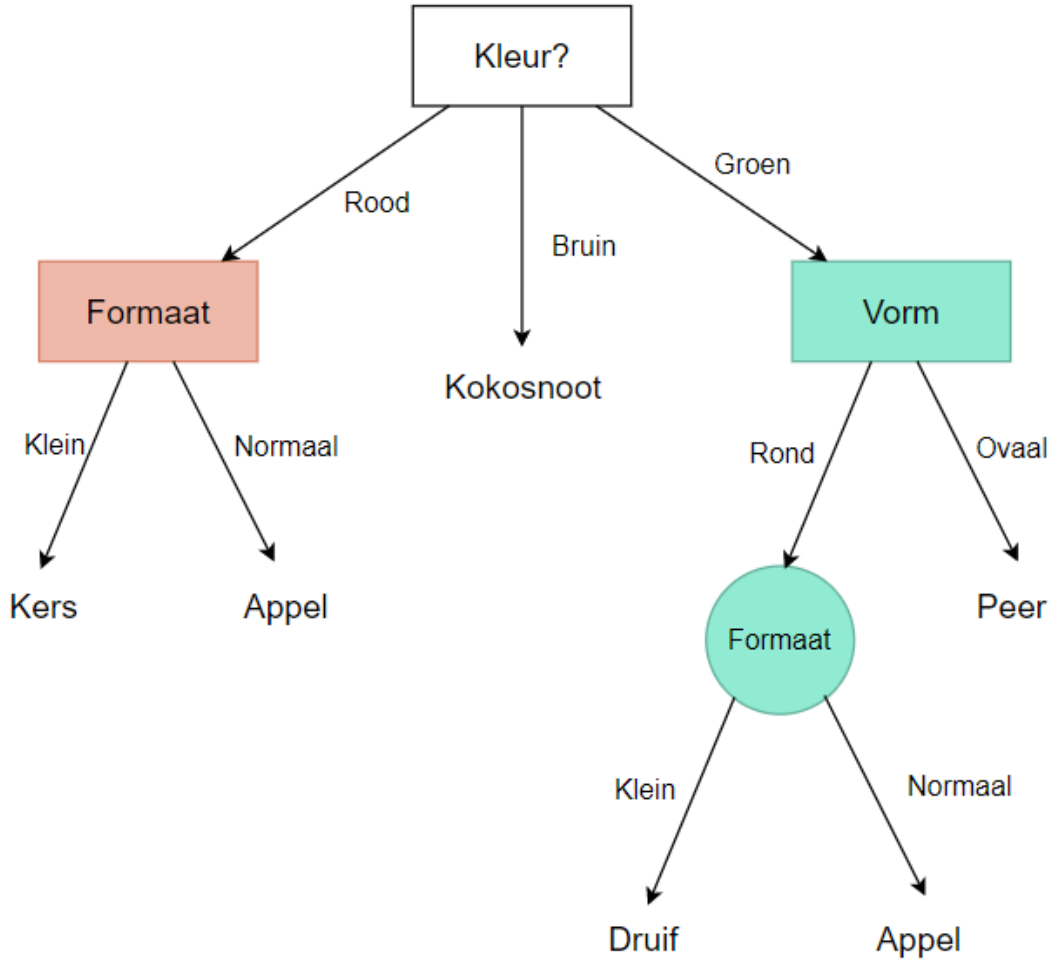
$$L(\boldsymbol{w}) = -\frac{1}{N}\left[\sum_{i=1}^{N} y_i ln(f_{\boldsymbol{w}}(x_i)) + (1 - y_i) ln(1 - f_{\boldsymbol{w}}(x_i))\right]$$

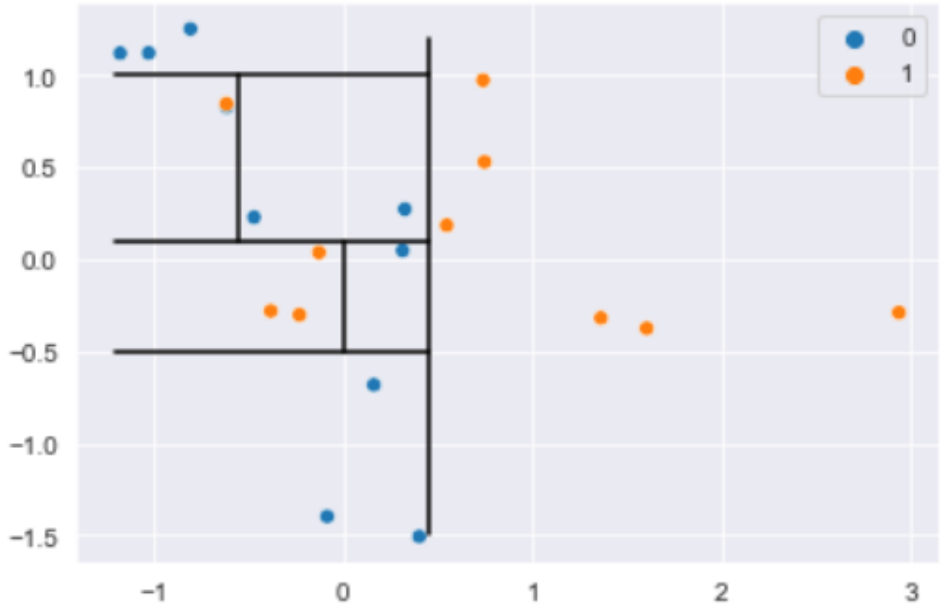
Decision Tree

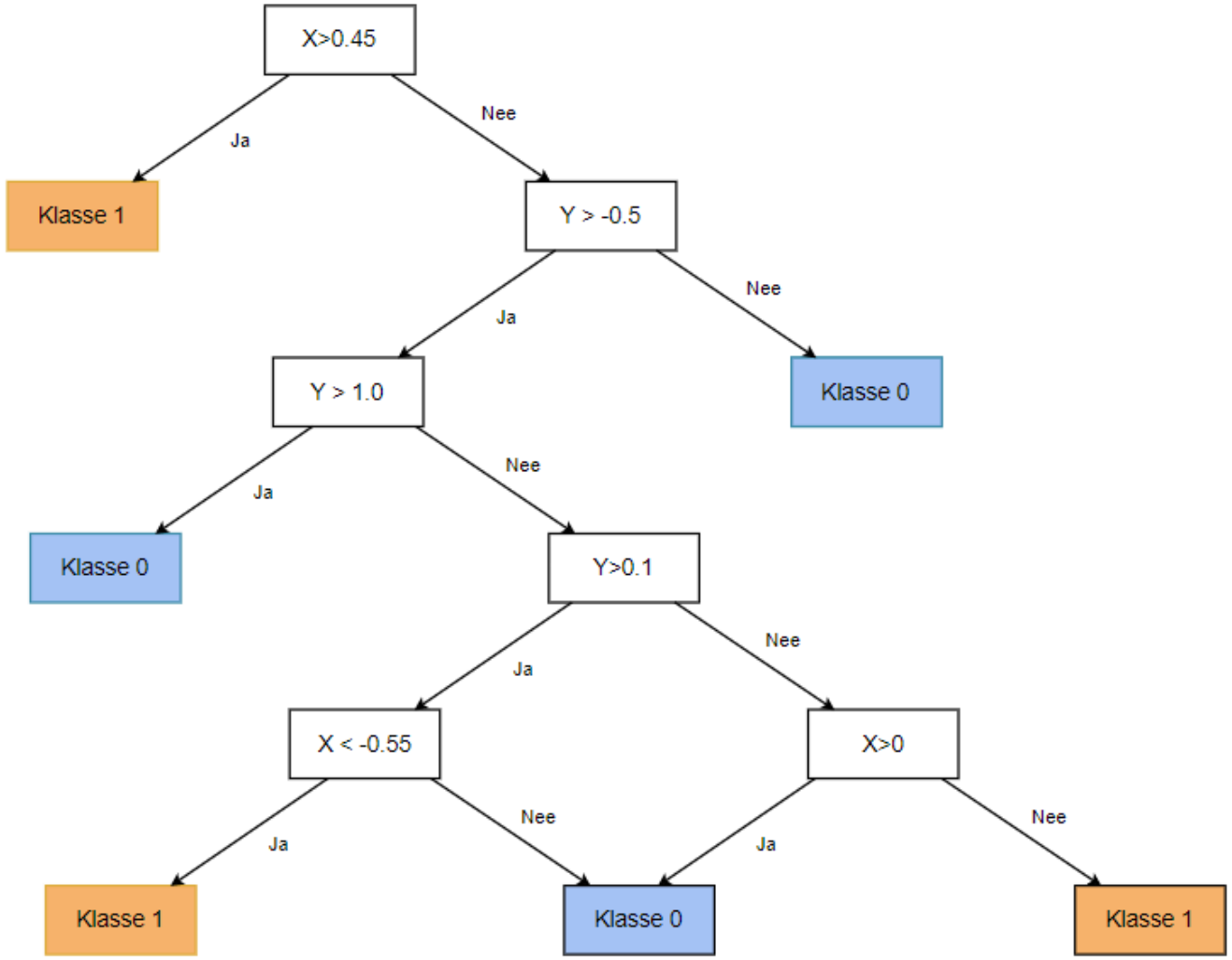

Kleur?
Rood
Bruin
Groen


Kleur?
Rood
Bruin
Groen
Formaat
Kokosnoot
Vorm
Klein
Normaal
Rond
Ovaal
Kers
Appel
Peer


Kleur?
Rood
Bruin
Groen
Formaat
Kokosnoot
Vorm
Klein
Normaal
Rond
Ovaal
Kers
Appel
Formaat
Peer
Klein
Normaal


Kleur?
Rood
Bruin
Groen
Formaat
Kokosnoot
Vorm
Klein
Normaal
Rond
Ovaal
Kers
Appel
Formaat
Peer
Klein
Normaal
Druif
Appel


0
1
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1
0
1
2
3


X>0.45
Ja
Nee
Klasse 1
Y > -0.5
Ja
Nee
Y > 1.0
Klasse 0
Ja
Nee
Klasse 0
Y>0.1
Ja
Nee
X < -0.55
X>0
Ja
Nee
Ja
Nee
Klasse 1
Klasse 0
Klasse 1

Wanneer en waar splitsen?

- Op basis van maatstaf voor de wanorde
- Heel wat hyperparameters om dit te beinvloeden
  - Hoe diep je kan gaan
  - Minimum aantal dat gesplitst kan worden
  - Minimum aantal dat samen kan zitten
  - …

Voor- en nadelen

- Eenvoudig
- Snel
- Reden voor beslissing

- Gevoelig aan ruis
- Neiging tot overfitting

Random Forest

Is Rusland groter dan Afrika?

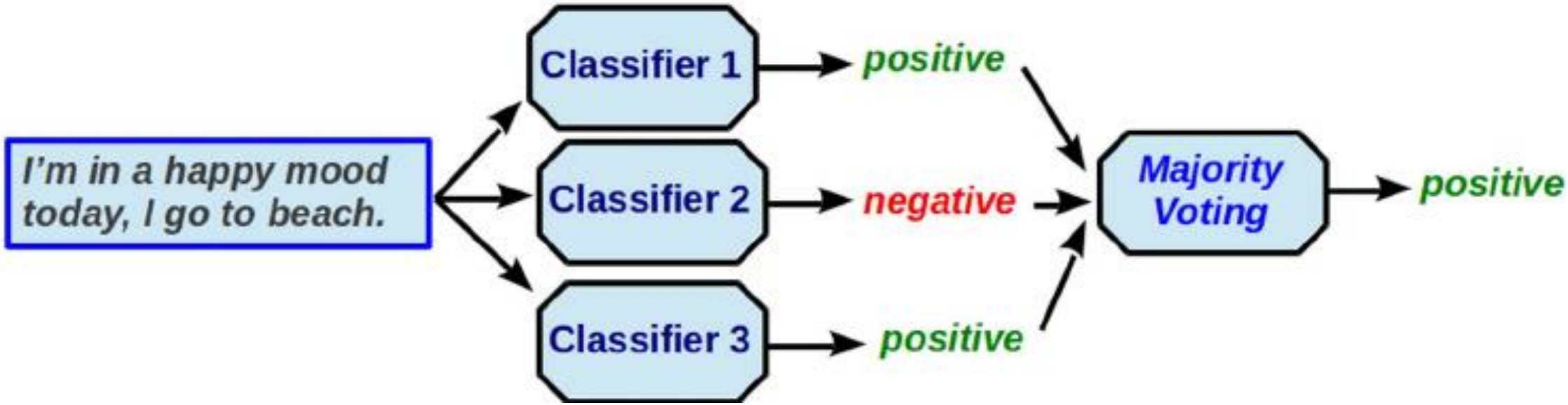
Ja: …

Nee: …

Condorcet's jury theorem

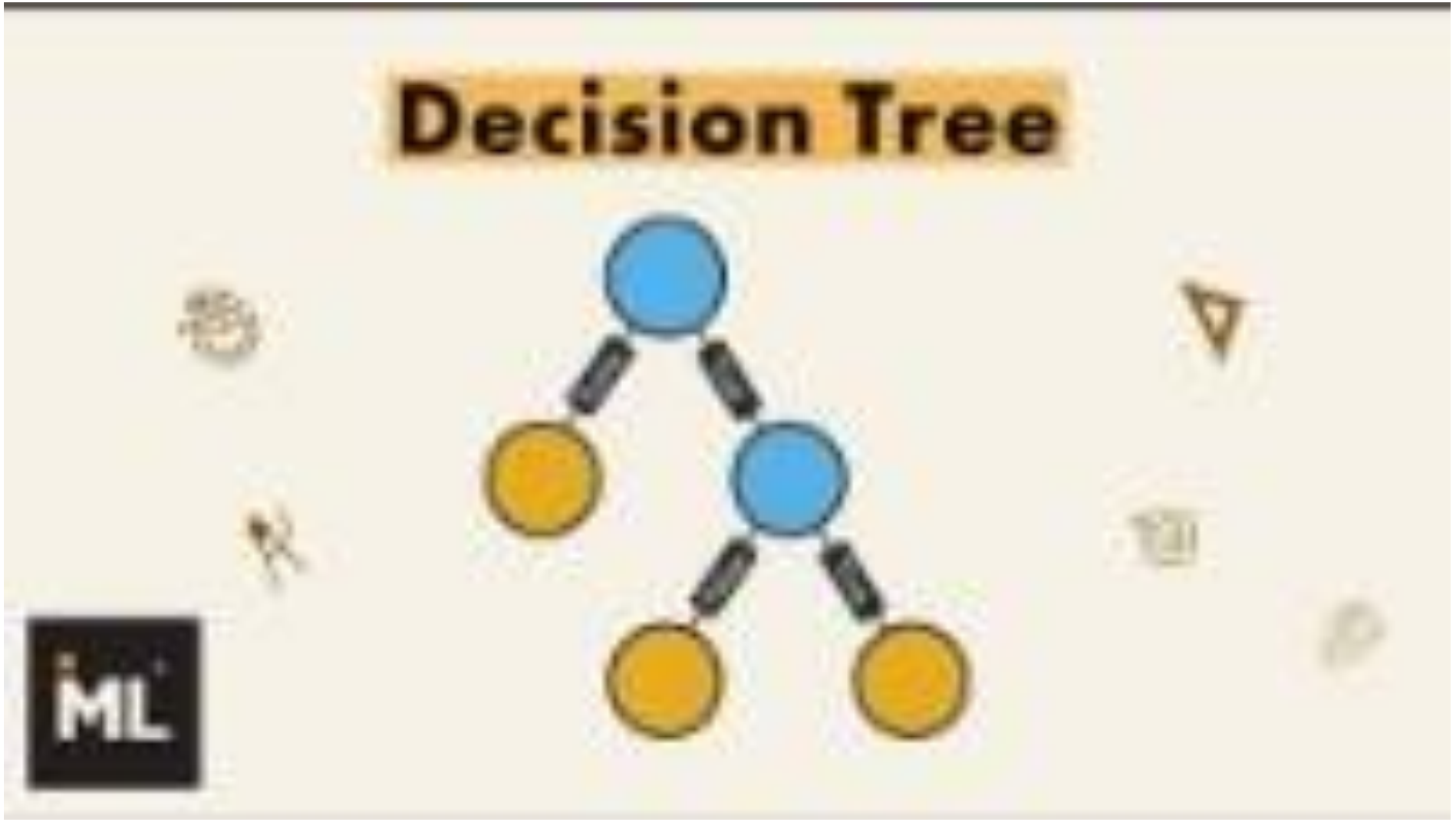
Given a jury of voters and assuming independent errors. If the probability of each single person in the jury of being correct is above 50% then the probability of the jury being correct tends to 100% as the number of persons increase.

Nicolas de Condorcet
1743 - 1794

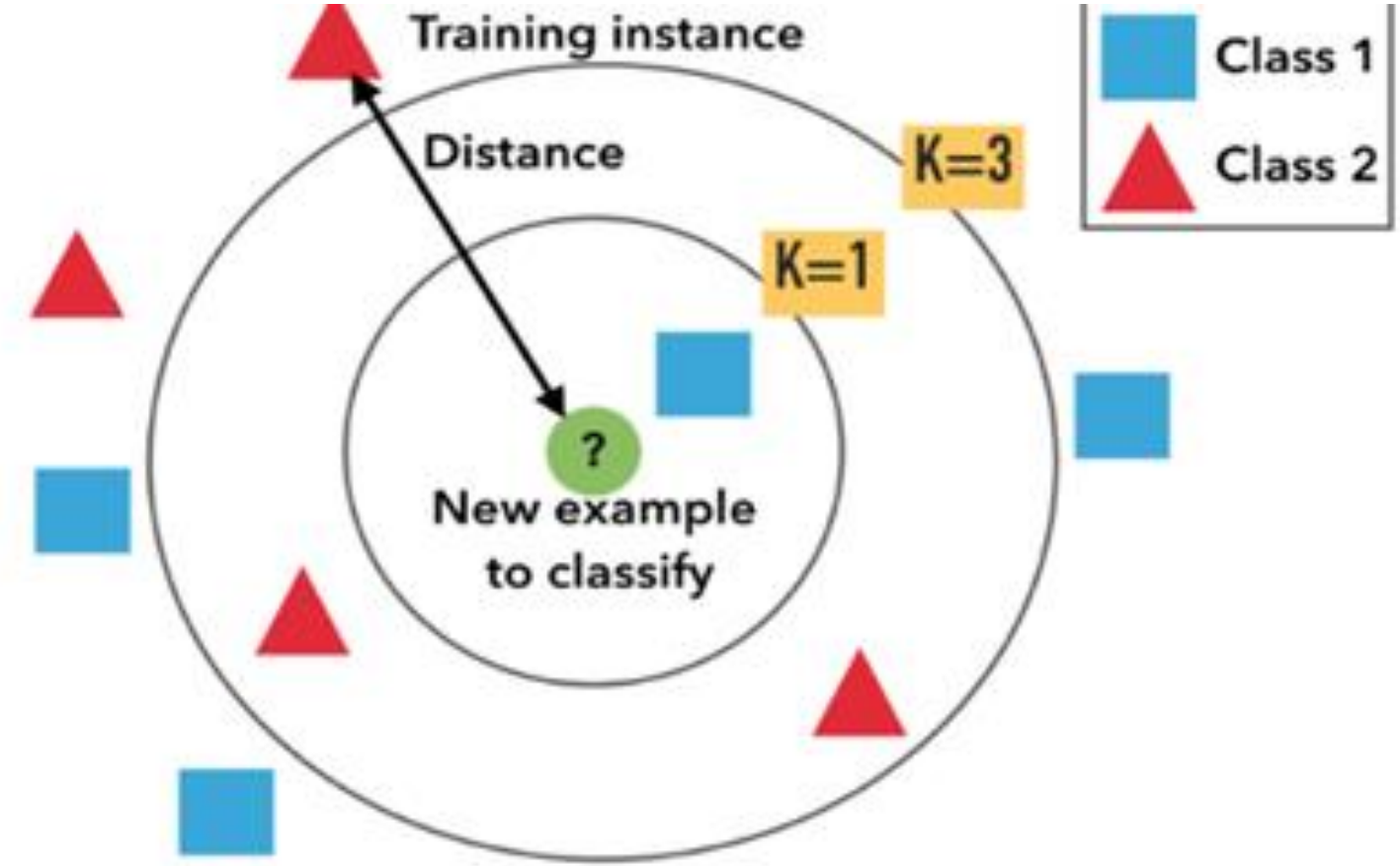

I'm in a happy mood today, I go to beach.
Classifier 1
Classifier 2
Classifier 3
positive
negative
positive
Majority Voting
positive

Beschikbare parameters

- Aantal bomen
- Aantal features per boom: int, float, sqrt, auto, log2, default
- Alle decision tree parameters
- …


Decision Tree
ML

K-Nearest Neighbours

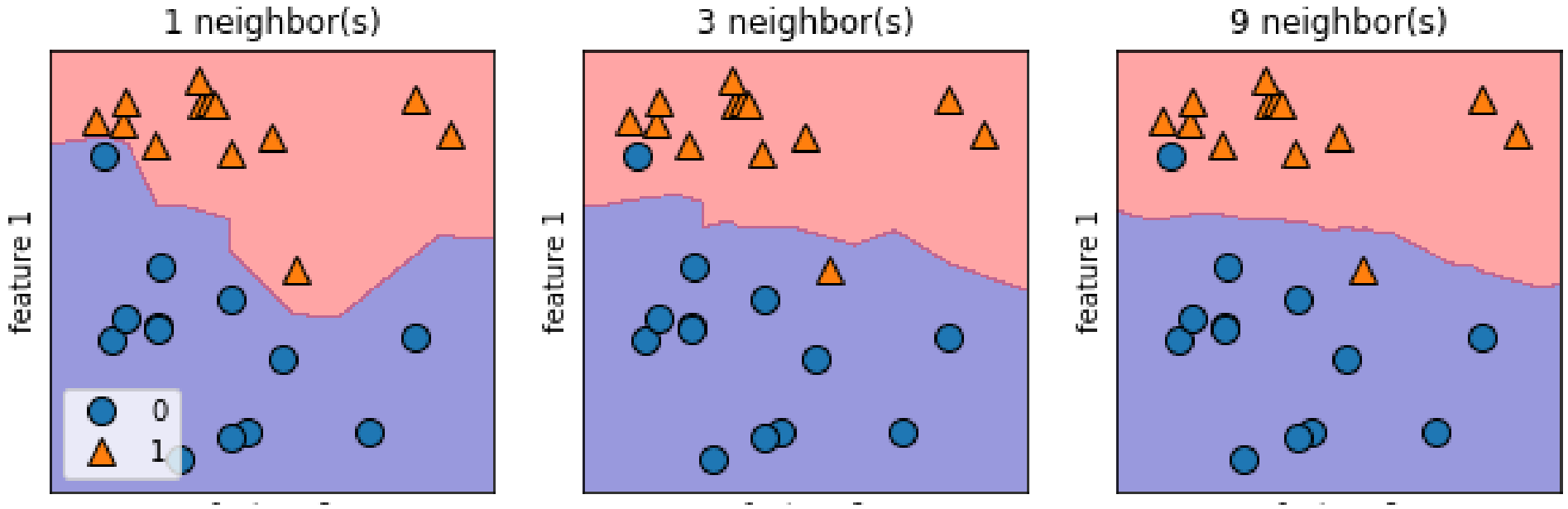

Training instance
Distance
K=3
K=1
?
New example
to classify
Class 1
Class 2

Kenmerken

- Lazy learning
  - Er is geen trainingsfase
- Eenvoudig
  - Classificatie: meest voorkomende
  - Regressie: Gemiddelde (eventueel gewogen met afstand)

Parameter K

- Lage K-waarde -> neiging tot overfitting
- Grote K-waarde -> neiging tot underfitting



Voor- en nadelen

- Eenvoudig
- Geen trainingsstap
- Goed bestand tegen ruis
- Heel accuraat met voldoende data

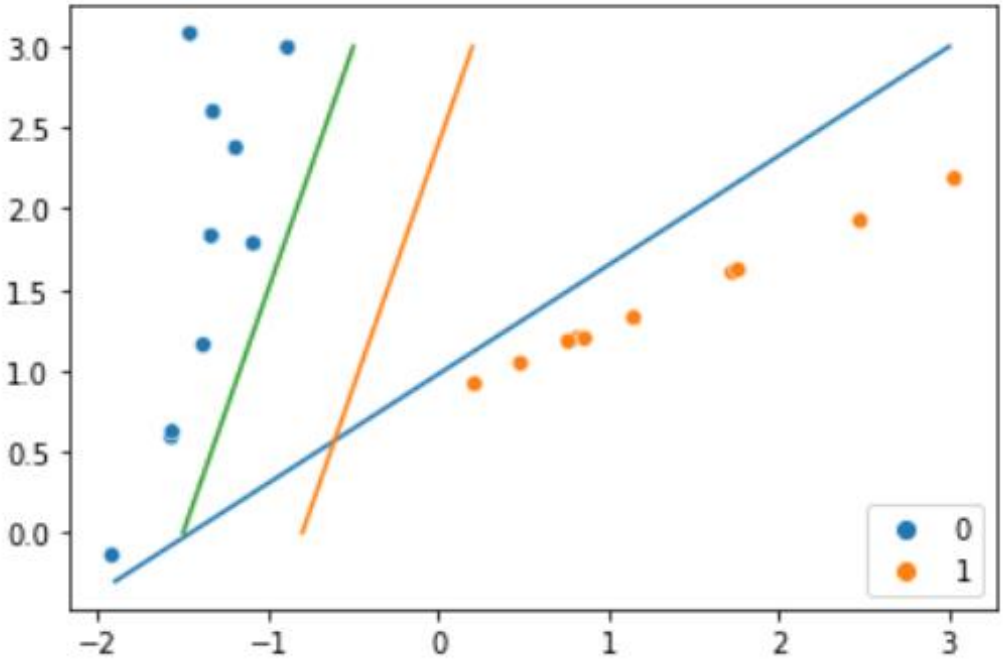
- Voorspelling is rekenintensief
- Weinig parameters
- Minder efficient bij veel features


K-NEAREST NEIGHBORS

Support Vector Machines

Support vector machine

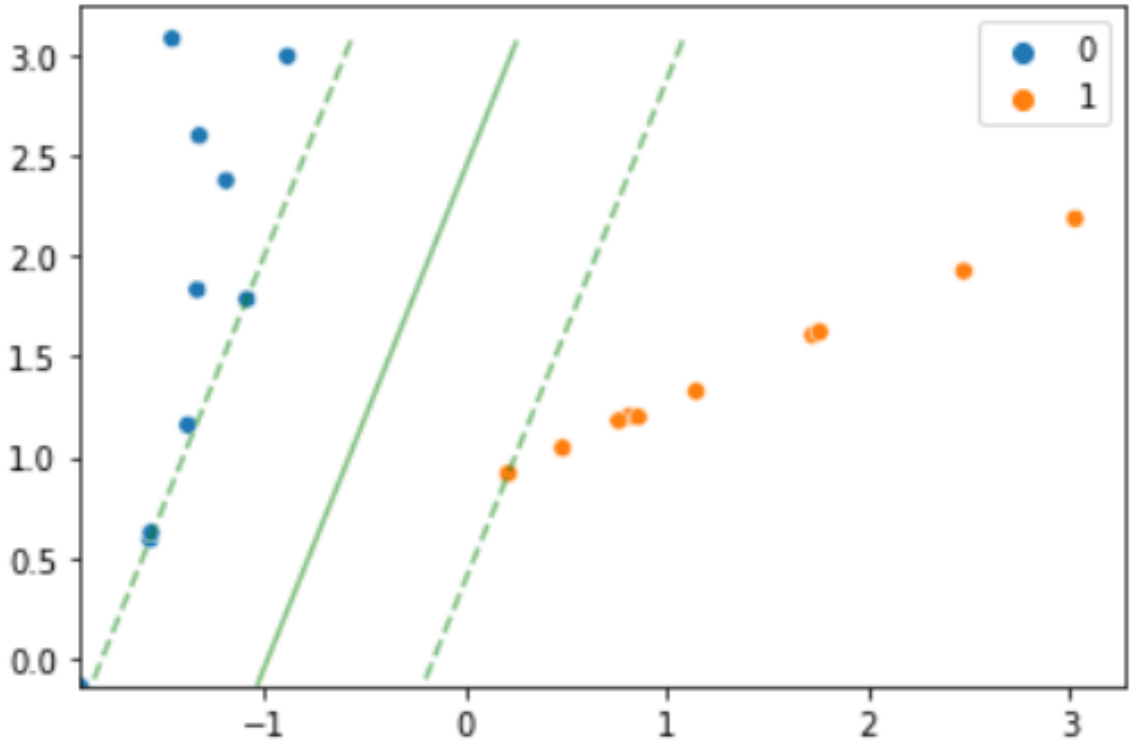
- Vaak afgekort tot SVM
- Robuster alternatief voor de kostfunctie van logistische regressie
  - Vooral voor testdata



Hoe werkt het?

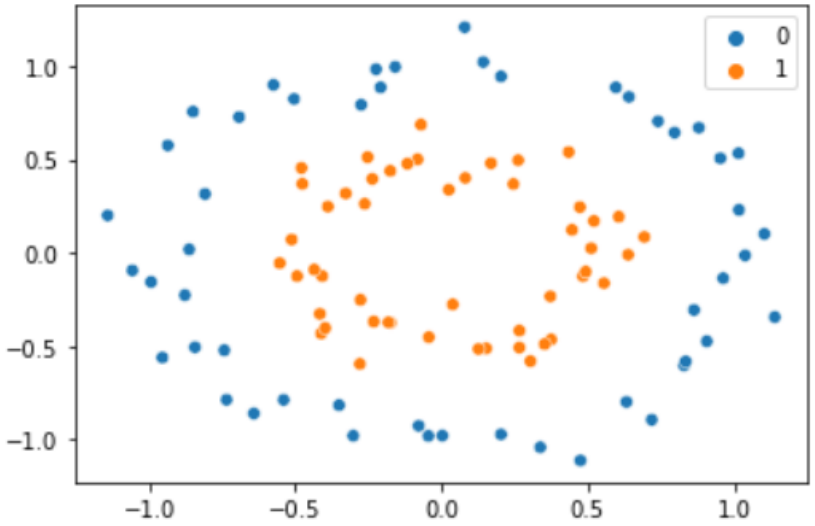
- Zoekt de scheidingslijn
- Maximaliseert de marge

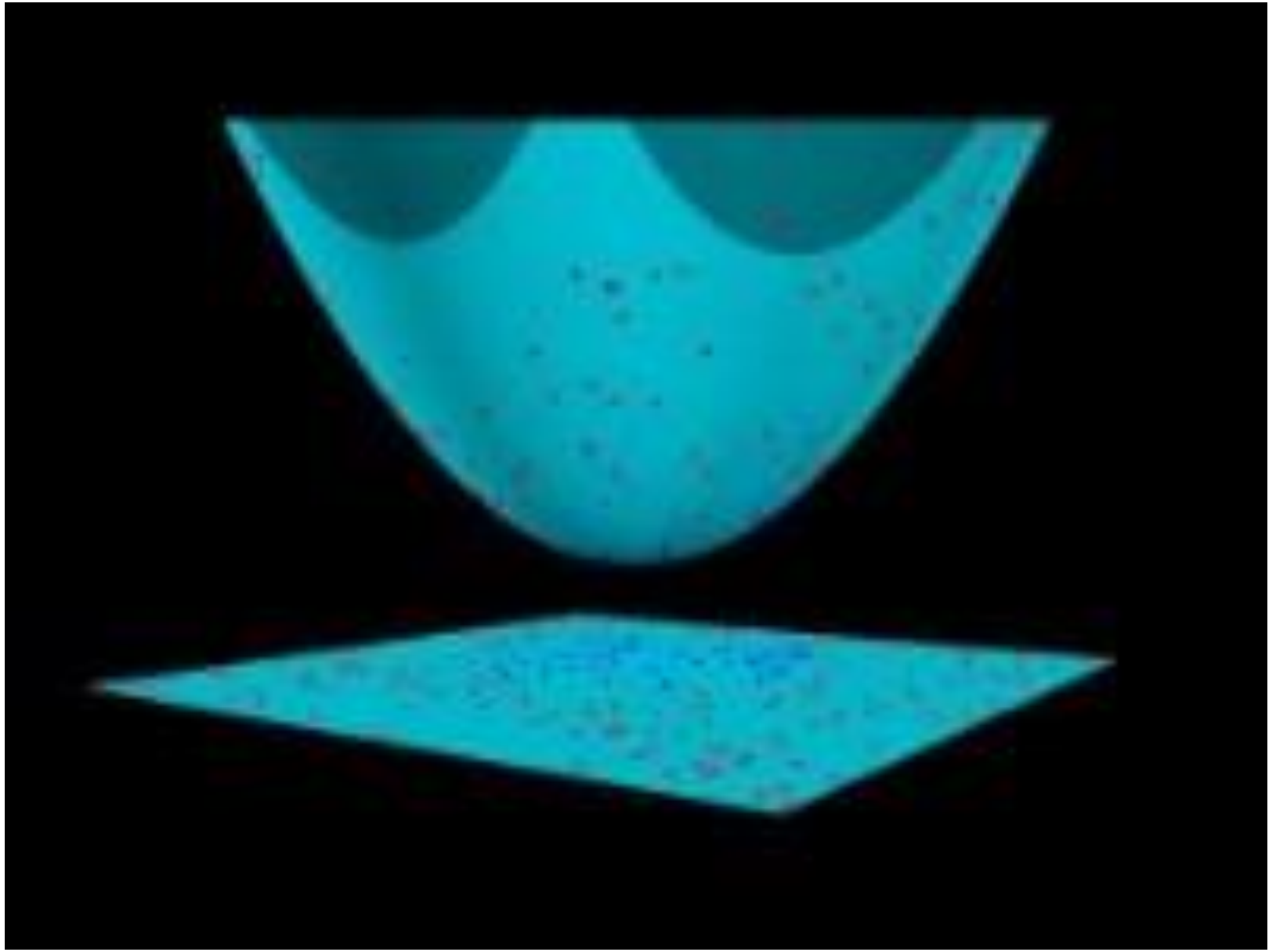
- Efficient want
  - Enkel de dichtste punten gebruikt
  - Worden de support vectors genoemd



Hoe werkt het?

- Belangrijke stap hierbij is een projectie naar een hogere dimensie
  - Gebruik hier een bepaalde functie voor
    - Kernel genoemd
  - Kan ook gebruikt worden voor niet-lineair scheidbare data



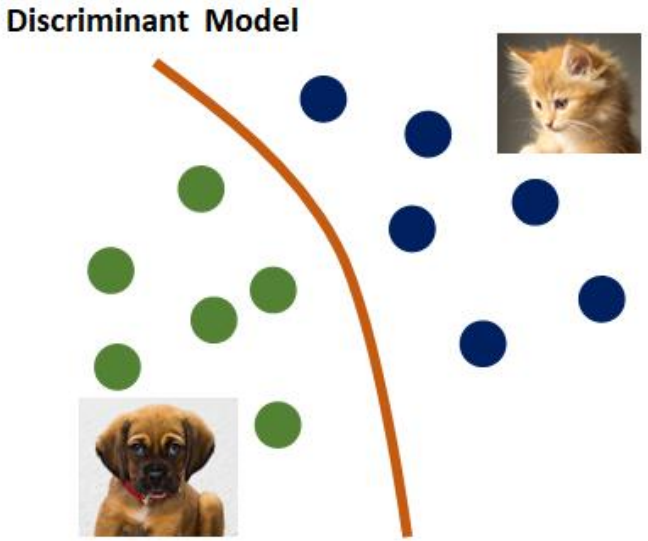


Voor- en nadelen

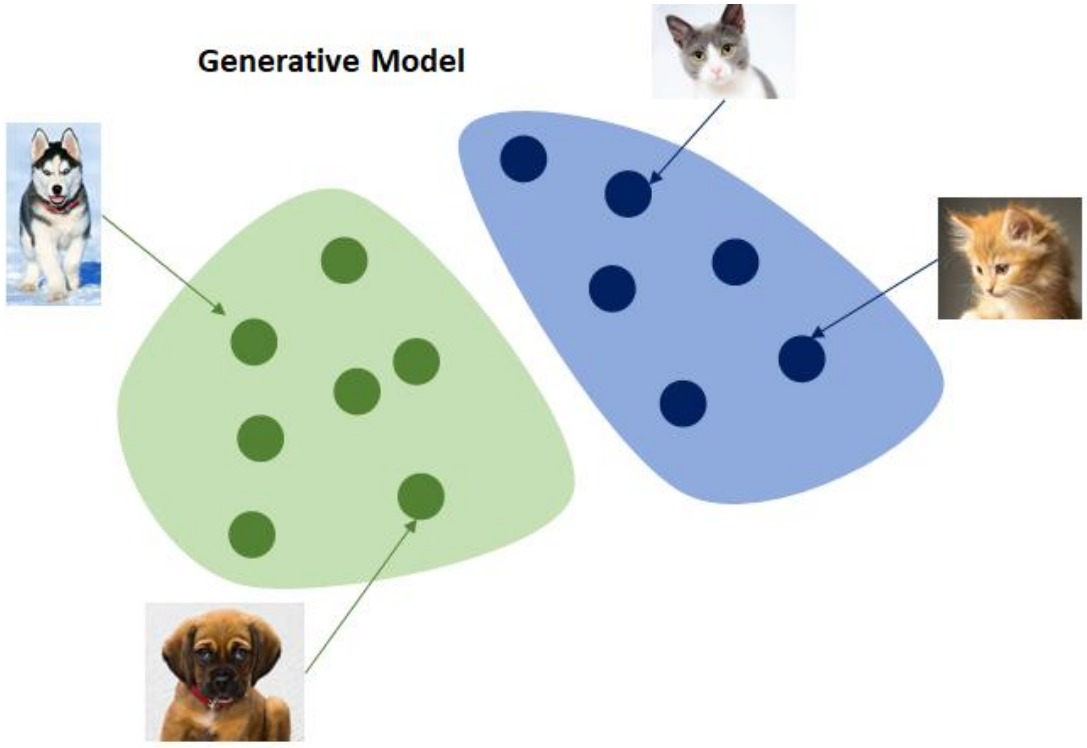
- Goede performantie op kleine data
  - Vooral als er veel features zijn
- Minder geheugen nodig
- Robuster
- Niet-gestructureerde data
- Niet gevoelig aan overfitting

- Gevoelig voor outliers
- Geen indicatie van de zekerheid
- Minder goed voor grote datasets

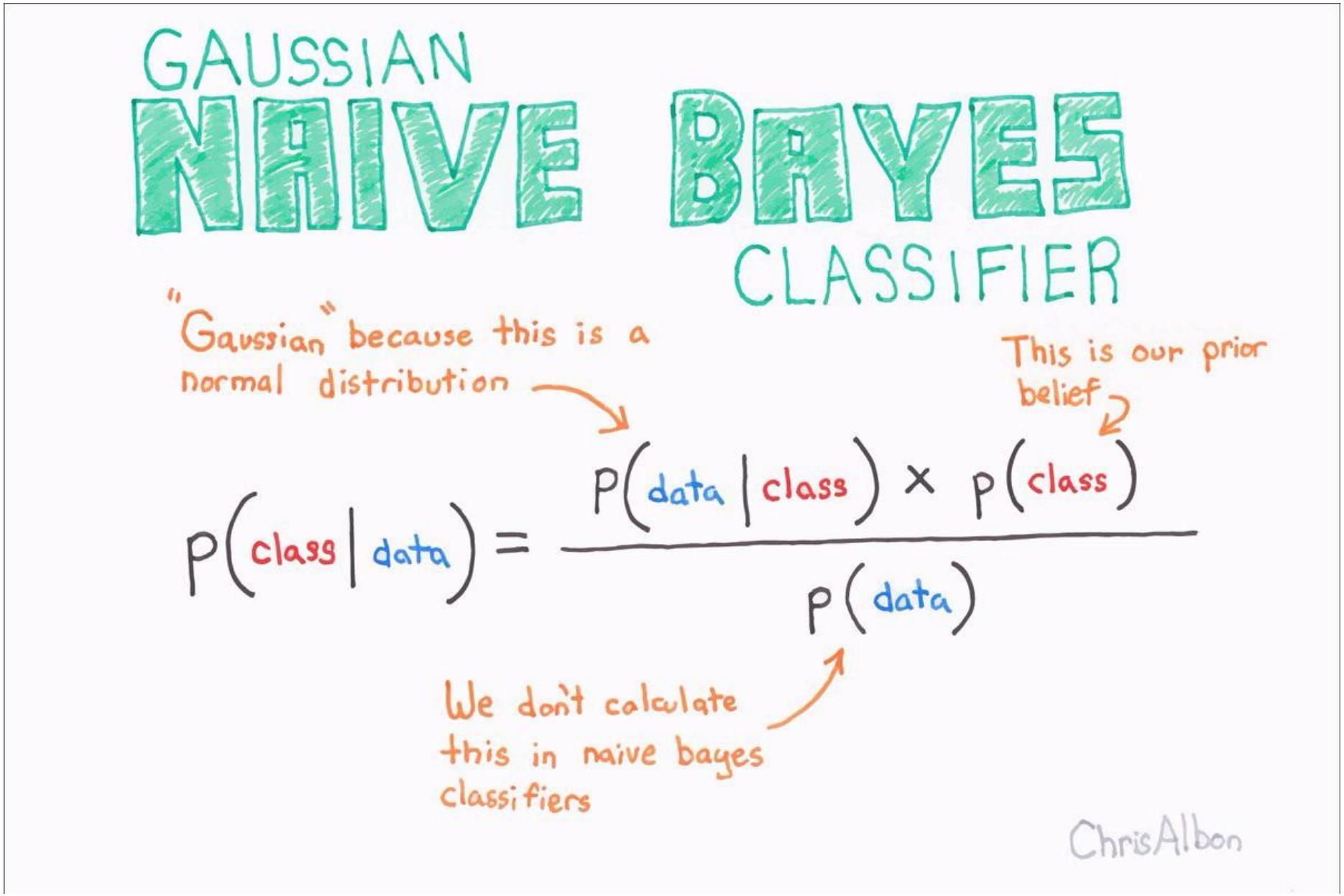
Naïve Bayes



P(y|x)
Standaardaanpak



P(x|y) en p(y)
Naïve Bayes


GAUSSIAN
NAIVE BAYES
CLASSIFIER
"Gaussian" because this is a normal distribution
This is our prior belief
P(class | data) = P(data | class) × P(class) / P(data)
We don't calculate this in naive bayes classifiers
ChrisAlbon

Huiswerk

Belangrijke termen

- Data Modelling
- Supervised learning
- Unsupervised learning
- Reinforcement learning
- Classification
- Regression
- Clustering
- Anomaly detection
- Training
- Evaluation
- Parameter
- Hyperparameter
- Classifier
- Regressor

Belangrijke termen

- Gewichten
- Loss-functie
- Least Mean Squares
- Gradient Descent
- Learning rate
- F1-score
- Confusion matrix
- Multiclass
- Multilabel
- One-vs-One
- One-vs-All
- Precision
- Recall

Belangrijke termen

- Lineaire regressie
- Logistische regressie
- Decision Tree
- Random Forest
- Naïve Bayes
- K-Nearest Neighbours
- Support Vector Machines

Data modelling tutorial

- Ga naar:
  - https://www.kaggle.com/learn/intro-to-machine-learning
  - Volg de tutorial volledig
  - Als er vragen zijn, stel ze zeker want volgende week is hier een oefening op

  - De informatie in de tutorials is te kennen leerstof en helpt bij het maken van de oefeningen

Data modelling: oefening in notebooks uit leerstof

- Bekijk de notebooks voor deze les

- Bevat extra informatie over hoe elke techniek werkt, de beschikbare parameters en voorbeeldcode