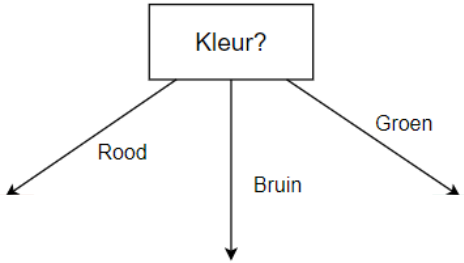
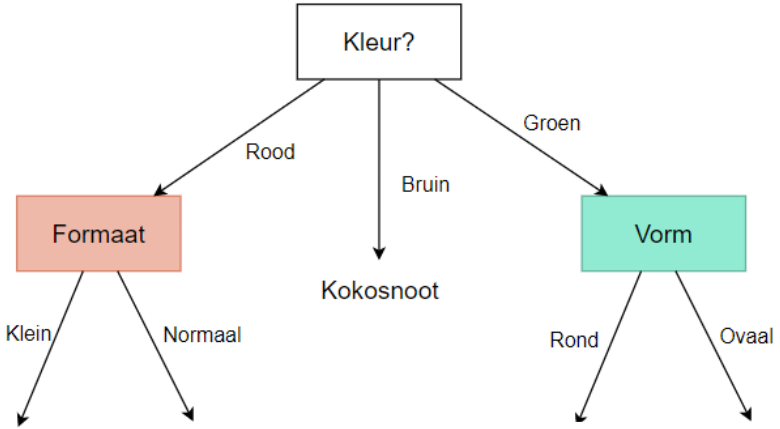
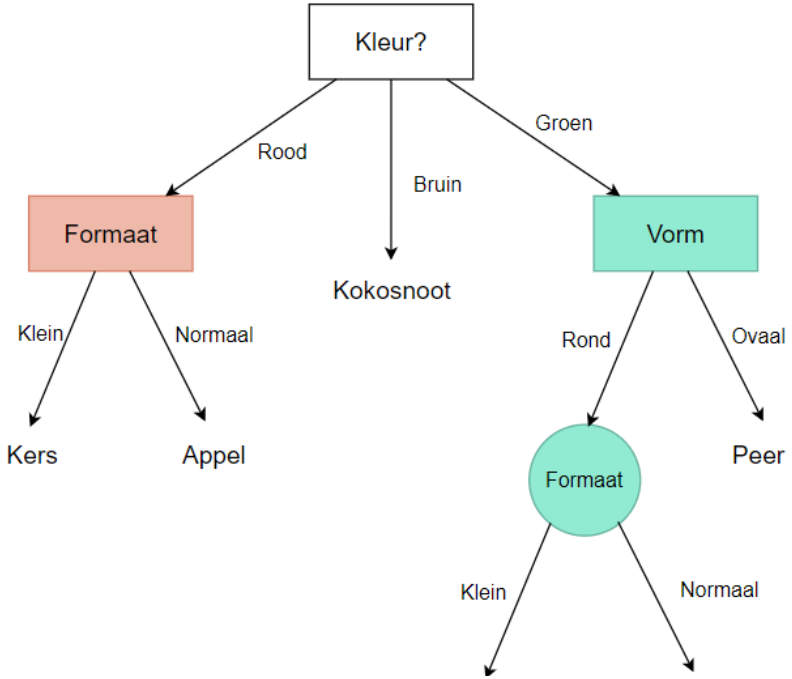
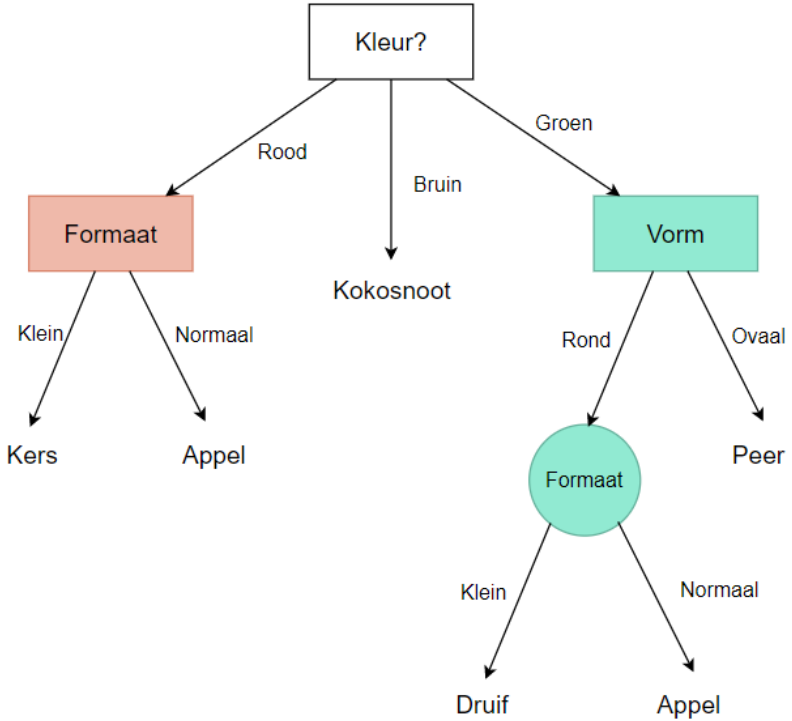
DECISION TREES

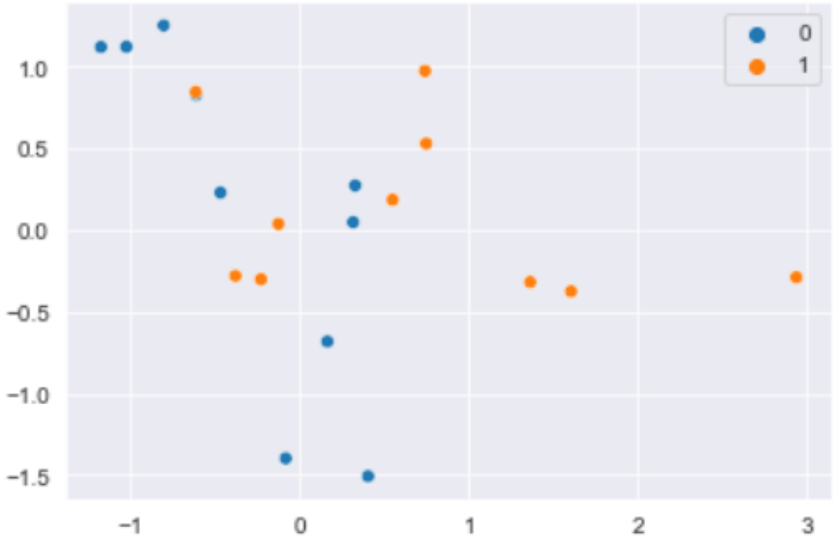
JENS BAETENS

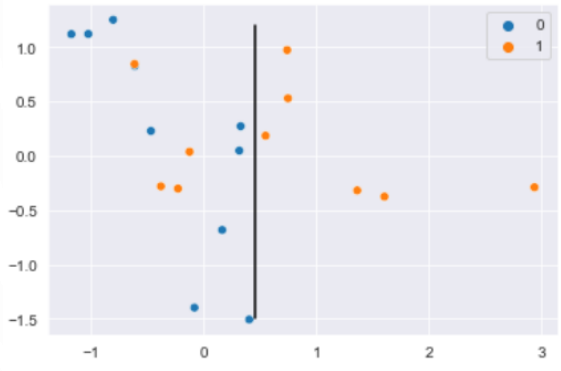

Kleur?
Rood
Bruin
Groen

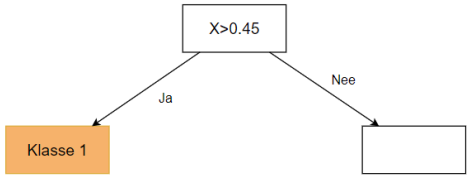

Kleur?
Rood
Bruin
Groen
Formaat
Kokosnoot
Vorm
Klein
Normaal
Rond
Ovaal

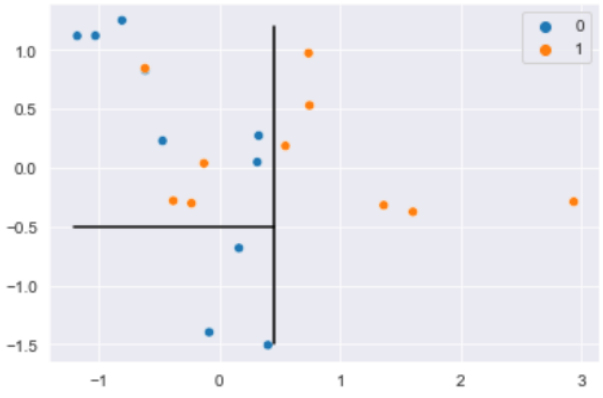

Kleur?
Rood
Bruin
Groen
Formaat
Kokosnoot
Vorm
Klein
Normaal
Rond
Ovaal
Kers
Appel
Formaat
Peer
Klein
Normaal

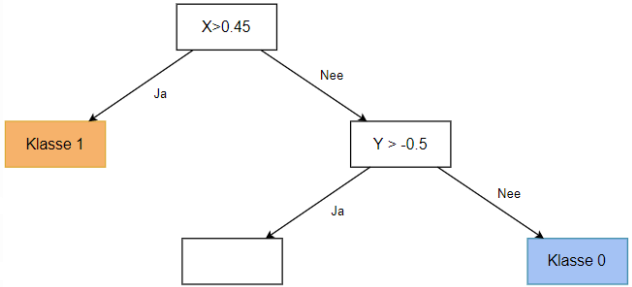

Kleur?
Rood
Bruin
Groen
Formaat
Kokosnoot
Vorm
Klein
Normaal
Rond
Ovaal
Kers
Appel
Formaat
Peer
Klein
Normaal
Druif
Appel

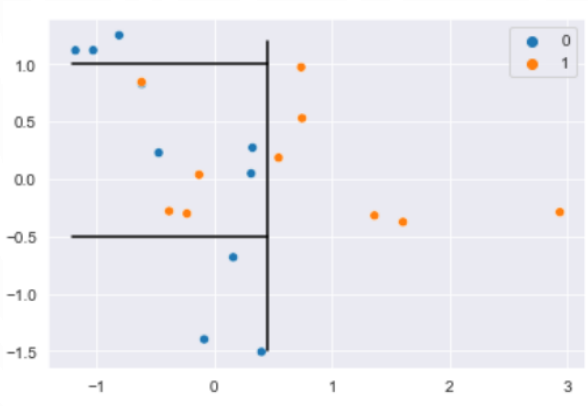

0
1
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1
0
1
2
3

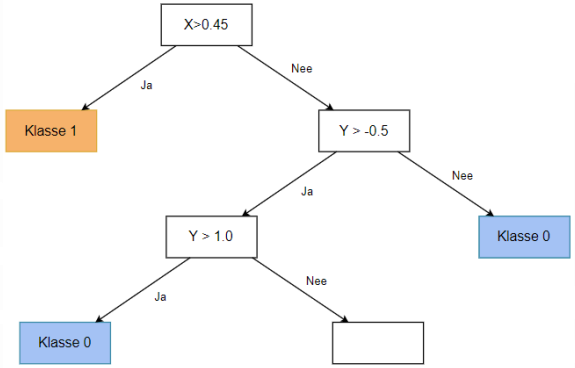

0
1
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1
0
1
2
3

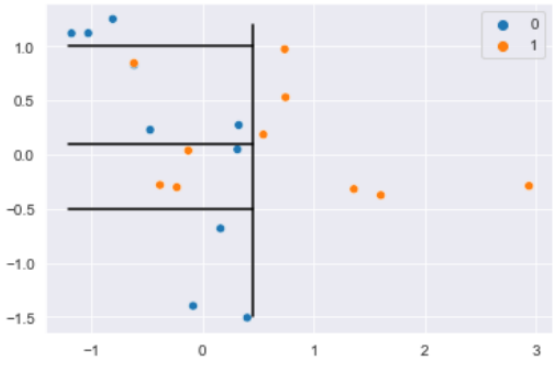

X>0.45
Ja
Nee
Klasse 1

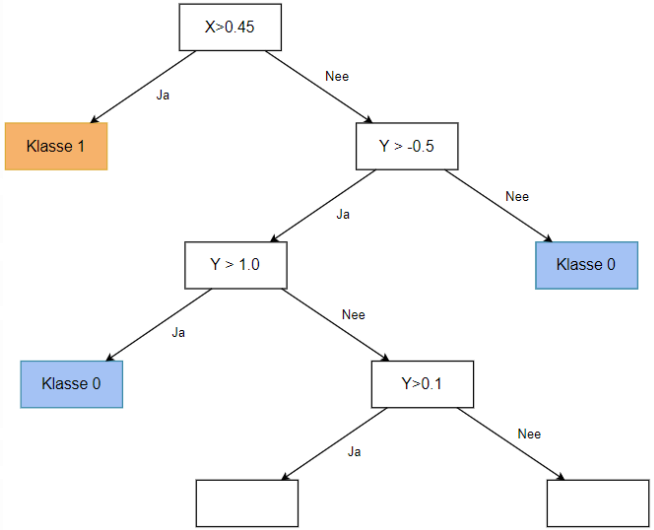

0
1
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1
0
1
2
3

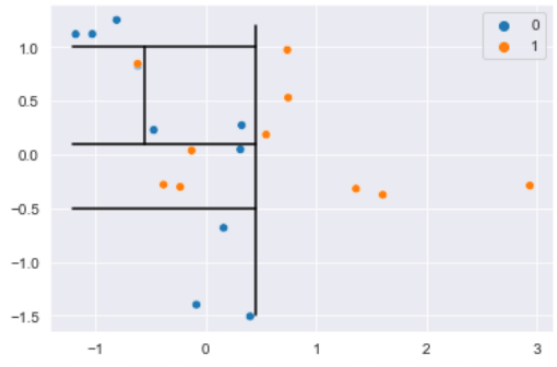

X>0.45
Ja
Nee
Klasse 1
Y > -0.5
Ja
Nee
Klasse 0

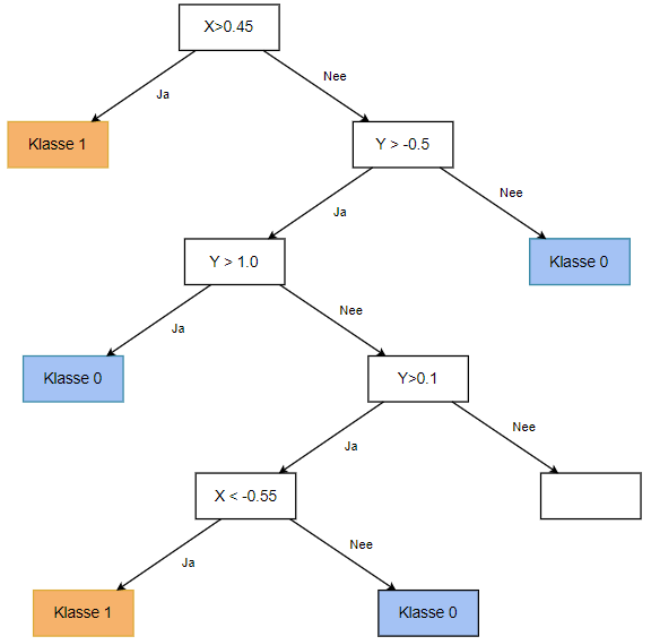

0
1
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1
0
1
2
3

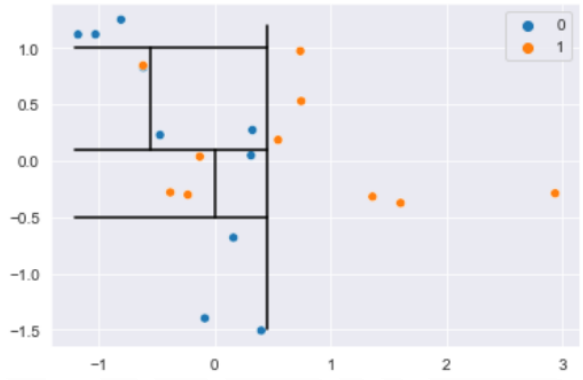

X>0.45
Ja
Nee
Klasse 1
Y > -0.5
Ja
Nee
Y > 1.0
Klasse 0
Ja
Nee
Klasse 0

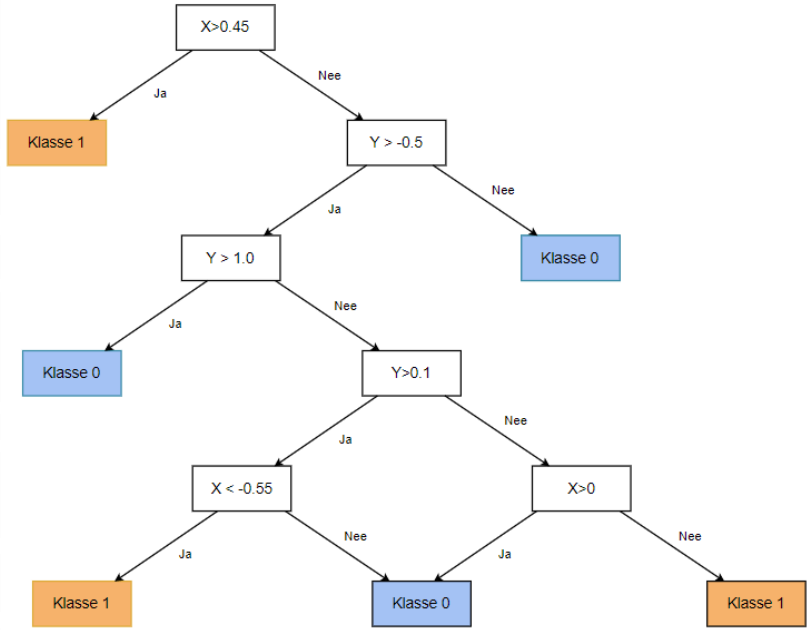

0
1
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1
0
1
2
3


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


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
Ja
Nee
Klasse 1
Klasse 0


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

HOE BEPALEN VAN DE SCHEIDINGSLIJN?

Bereken entropie (maatstaf voor de wanorde) voor elke mogelijke verdeling

$$H = \sum_{i=1}^{N} p_i \log_2\left(\frac{1}{p_i}\right) = -\sum_{i=1}^{N} p_i \log_2(p_i)$$

P_i is het percentage van elke klasse in een gebied

Bereken het verschil met de entropie voor en na de verdeling

=> Information gained => Hoe groter hoe beter

ALTERNATIEF VOOR ENTROPIE

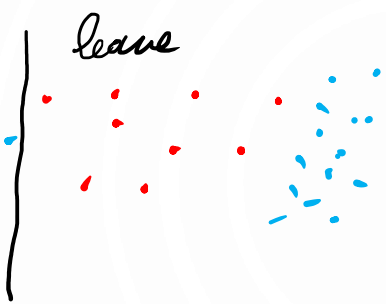
Logaritmes voor berekenen entropie zijn rekenintensief

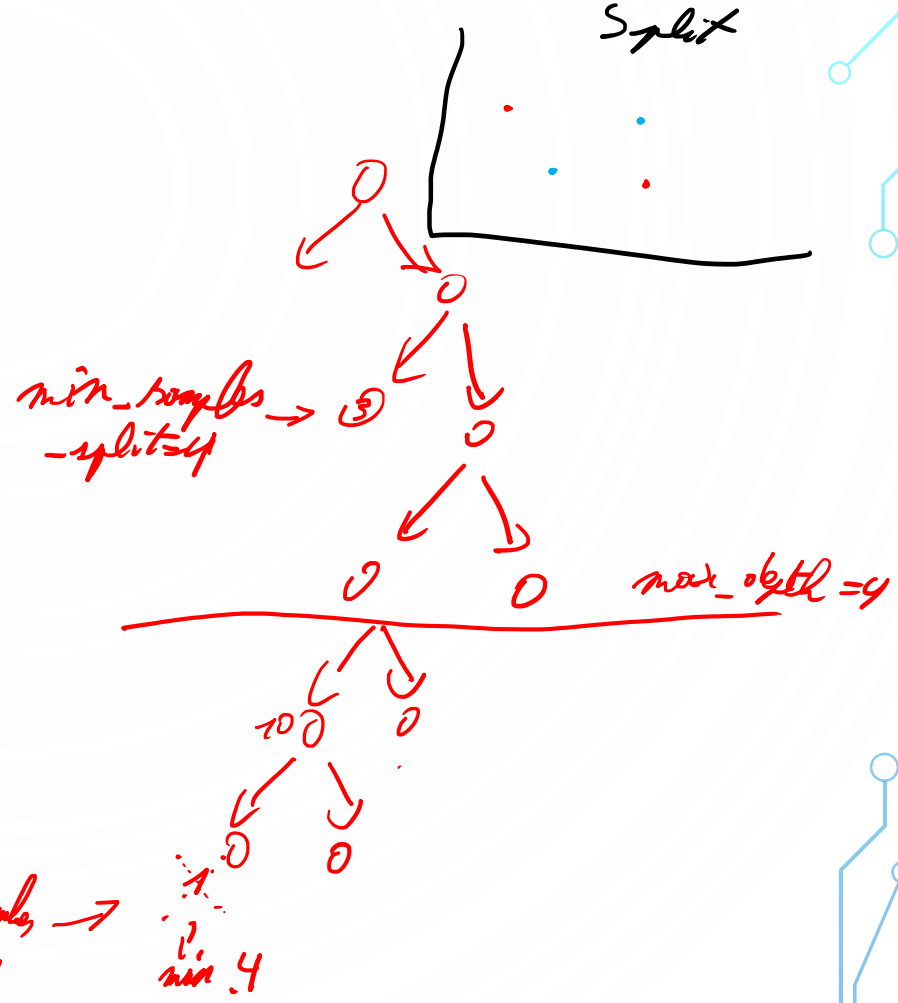
Alternatieve manier: Gini impurity

$$G = 1 - \sum_{i=1}^{N} p_i^2$$

Beste scheidingslijn heeft de kleinste Gini Impurity

NADELEN





Gevoelig aan ruis

Sterke neiging tot overfitting

Hyperparameters voor regularisatie:

- max_depth
- min_samples_split
- min_samples_leave

VOORDELEN

Eenvoudig algoritme en daardoor ook snel

Men kan redeneren op het resultaat

- Niet mogelijk bij SVM, LR, Naïve Bayes, …

https://www.youtube.com/watch?v=JcI5E2Ng6r4

RANDOM FORESTS

RANDOM FORESTS

Is Rusland groter dan Afrika?

Ja: …

Nee: …

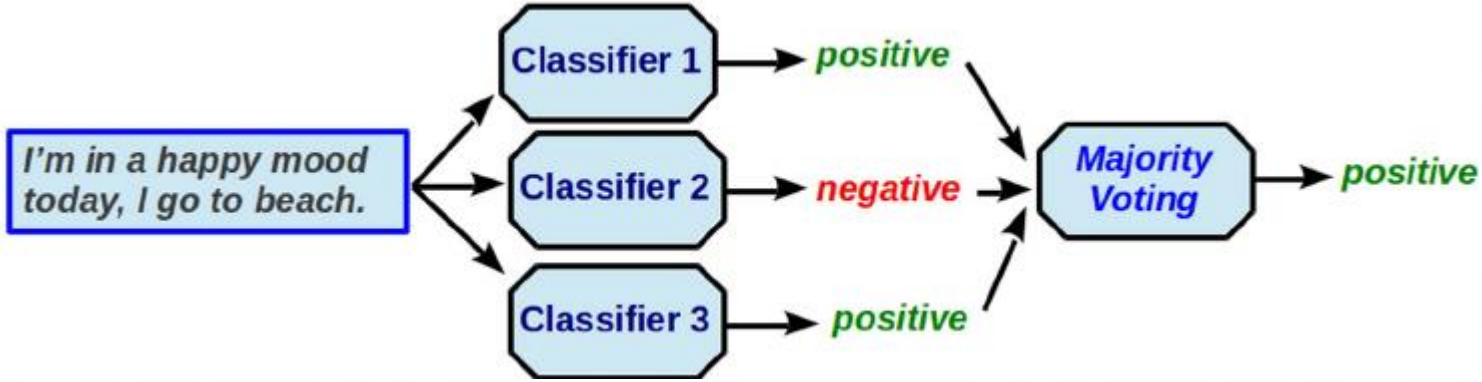
RANDOM FORESTS

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)

RANDOM FORESTS



RANDOM FORESTS - PARAMETERS

Aantal bomen

Aantal features per boom: int, float, sqrt / auto, log2, default → splitst op kolom

Bagging: Gebruik slechts een deel van de data om elke boom te trainen → splitst op rijen (toevalsgetal)

Oob_score: Aangezien niet alle data gebruikt wordt om elke boom te trainen. Gebruik deze data om de bomen te valideren. Vooral handig als er niet genoeg data is om een aparte validatieset te behouden.