# Wat is machine learning?

Machine learning is het proces waarbij computers zich kunnen aanpassen zonder het volgen van vooraf geschreven regels, dit wordt gedaan door het analyseren van data.

## Wat kan machine learning wel?

Machine learning kan ons helpen in het maken vaan beslissingen en het vinden van onregelmatigheden die anders niet gevonden zouden zijn.

## Wat kan machine learning niet?

## Wat is het verschil tussen supervised en unsupervised learning?

Bij supervised learning heeft elke datapunt in de train (en test) data een label, als je een model wilt trainen om underscheid te maken tussen een hond en een kat, zal elke foto een label hebben met “hond” of “kat”. In unsupervised learning wordt er geen label mee gegeven, in plaats daarvan wordt er een bepaalde “structuur” in de data gevonden die gebruikt wordt om je model te trainen.

# Hoe werkt k-means?

K-means clustering is een “unsupervised learning” clustering algorithm, dit betekend dat het gebruikt kan worden om een set van datapoints in groepen de klusteren, zonder labels aan te geven.

Zo werkt het:

De input is een set van punten (x1…. xN), De K in K-means zijn de aantal groepen (clusters) voor punten. Daarna plaatsen we K centroids in de ruimte van de punten. Vervolgens herhalen we deze stappen:

Voor elk punt vind je de dichtstbijzijnde centroid, en geef dat punt aan de gevonden centroid, zo maak je een cluster, vervolgens voeren we de volgende stappen uit op elke cluster:

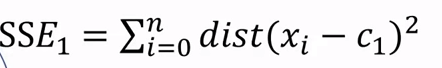
Voor elke cluster gaan we de positie van de centroid opnieuw berekenen door alle punten te nemen en hun gemiddelde positie uit te rekenen.

Voer dit uit tot dat er geen nieuwe punten aan een centroid worden gegeven.

## Hoe vind je de beste K?’

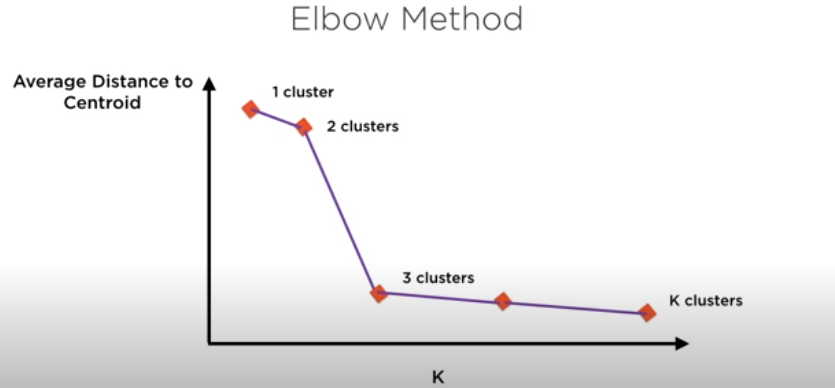
Omdat je de data meestal zelf kan inzien wordt er meestal een gepaste K gekozen

Er zijn ook andere manieren om dit te doen, je kan de “elbow method” gebruiken. Hierbij gebruiken we een gekozen kost functie met als input K, deze functie kan bijvoorbeeld SSE zijn:



*Xi is hier de ith datapunt, en c1 is de centroid behorend tot dat datapunt, je kan euclidean distance gebruiken voor dist()*

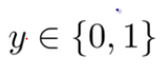
als je de kostfunctie op een grafiek zet kan het er zo uit zien:

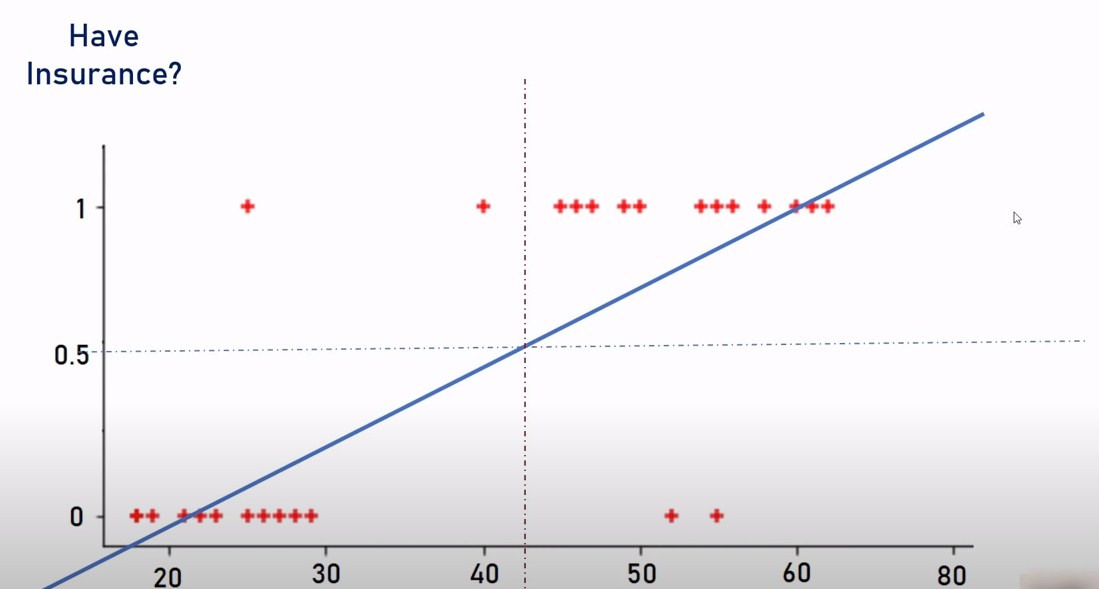


We zeggen dan dat de plek van de ‘eelboog’ de beste K is (in dit geval 3)

# Hoe werkt logistic regression?

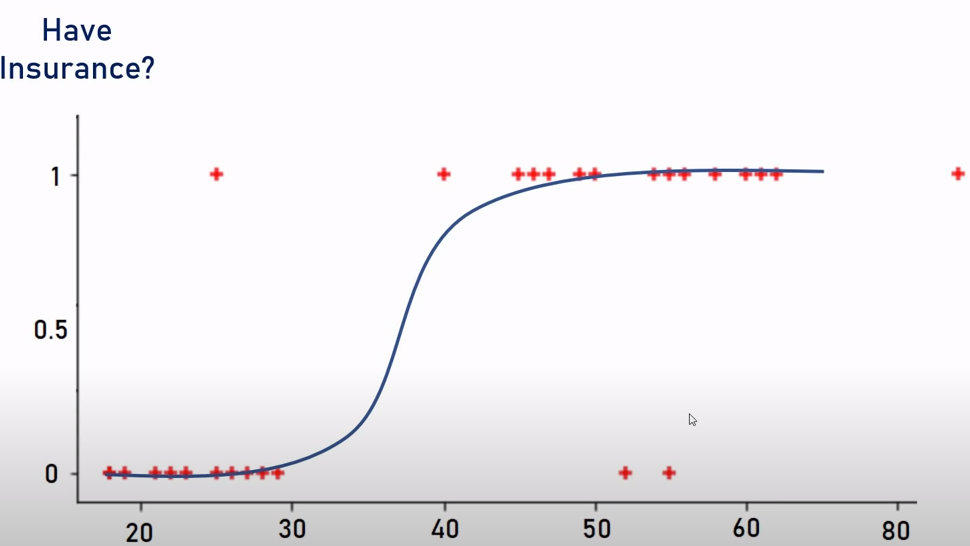
Logistic regression is een classificatie techniek, we hebben meerdere classificatie types: binary en multi classificatie. Binary classificatie is van toepassing wanneer je een dataset in “Ja” en “nee” moet groeperen, multi classificatie is van toepassing wanneer je een dataset moet classificeren in groepen zoals voor welke partij iemand zal stemmen.

Voor binary classificatie kan je dit zo zien, voor een willekeurige x wil je een y zodat

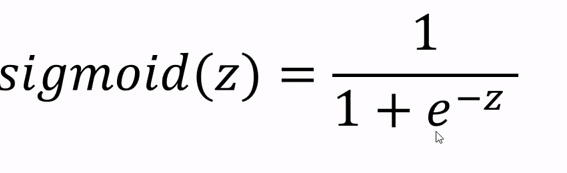
Waar y “ja” of “nee” representeert. Als je data hebt dat “linear” omhoog gaat zou je linear regression kunnen gebruiken, dit ziet er zo uit voor data tussen y = [0,1] 

Je zou dan kunnen zeggen dat alle waarden boven 0.5 waar zijn, en onder onwaar. Linear regression veranderd snel als er wat “outliers” in je data set zit, dit is dus niet de optimale manier.

Met logistic regression kunnen wij een lijn krijgen die er zo uitziet:



Dit werkt beter.



We gebruiken de sigmoid functie voor deze lijn, waar z = de lijn is van linear regression.

# Hoe werken decision trees