

Zaahir Joeman(23021098)  
Chris Aartman(23081872)  
Jayden Jagroep(23052279)

Inhoudsopgave

[Onderzoek de dataset (Iris Flowerset) 3](#_Toc198639407)

[Wat betekent elke kolom? 3](#_Toc198639408)

[Zijn er ontbrekende waarden? 3](#_Toc198639409)

[Welk type probleem is het? (Classificatie of regressie?) 4](#_Toc198639410)

[Onderzoek de dataset(World Happiness) 4](#_Toc198639411)

[Wat betekent elke kolom? 4](#_Toc198639412)

[Zijn er ontbrekende waarden? 4](#_Toc198639413)

[Welk type probleem is het? (Classificatie of regressie?) 5](#_Toc198639414)

[Vergelijk en reflecteer 5](#_Toc198639415)

[Vergelijking van de prestaties van de modellen 5](#_Toc198639416)

[Welk model werkte beter, en waarom? 5](#_Toc198639417)

[Uitdagingen bij voorbereiding en training 5](#_Toc198639418)

# Onderzoek de dataset (Iris Flowerset)

## Wat betekent elke kolom?

Afbeelding met tekst, schermopname

Door AI gegenereerde inhoud is mogelijk onjuist.

Zoals we hier zien, zijn er vier onafhankelijke variabelen: ***Sepal\_Length*, *Sepal\_Width*, *Petal\_Length* en *Petal\_Width*.** De afhankelijke variabele bij de Decision Tree is *Species*, omdat we proberen te voorspellen tot welke soort een bloem behoort.

Bij het lineaire regressiemodel gebruiken we dezelfde onafhankelijke variabelen, maar voorspellen we de *Petal\_Length*. In dat geval is *Petal\_Length* de afhankelijke variabele.

## Zijn er ontbrekende waarden?

Bij de iris dataset zijn er geen ontbrekende waardes.

## Welk type probleem is het? (Classificatie of regressie?)

Het type probleem hangt af van wat we willen voorspellen. In het geval van de Iris-dataset spreken we van een classificatieprobleem wanneer we bepalen tot welke soort bloem een bepaalde meting behoort.

Wanneer we daarentegen een numerieke waarde willen voorspellen, zoals de lengte van het bloemblad (*Petal\_Length*), dan hebben we te maken met een **regressieprobleem**. In dat geval is de uitkomst een continue waarde.

# Onderzoek de dataset(World Happiness)

## Wat betekent elke kolom?

Afbeelding met tekst, schermopname, zwart-wit, zwart

Door AI gegenereerde inhoud is mogelijk onjuist.

In de World Happiness Report dataset onderzoeken we welke factoren de **Happiness Score** van inwoners in verschillende landen beïnvloeden. Voor het lineaire regressiemodel gebruiken we de Happiness Score als afhankelijke variabele, omdat we een continue numerieke waarde willen voorspellen.

Voor het decision tree-model proberen we te voorspellen tot welke region behoort. Hierbij is de regio de afhankelijke variabele.

De afhankelijke variabele is:

Happiness Score(Lineare Regressie) en Region(Decision Tree)

De onafhankelijke variabele is:

Economy (GDP per Capita), Family, Health (Life Expectancy, Freedom, Generosity, Trust (Government Corruption) en Dystopia Residual

## Zijn er ontbrekende waarden?

Er zijn geen ontbrekende waarden.

## Welk type probleem is het? (Classificatie of regressie?)

Hier geldt weer hetzelfde. Voor de Region is het een classificatieprobleem. Voor de World Happiness Score, dat is natuurlijk een numerieke waarde. Dat is een regressieprobleem.

# Vergelijk en reflecteer

## Vergelijking van de prestaties van de modellen

Voor deze opdracht heb ik twee datasets gebruikt: de *World Happiness Report* en de *Iris*-dataset. Op beide heb ik verschillende modellen toegepast om zowel regressie- als classificatieproblemen te onderzoeken.

Voor de World Happiness Report heb ik een *lineair regressiemodel* gebruikt om de *Happiness Score* te voorspellen. Dit model gaf goede resultaten met een lage MSE-score, wat erop wijst dat de voorspellingen vrij nauwkeurig waren. Ook heb ik een *decision tree classifier* getraind om de *regio* van landen te voorspellen. Deze had een accuracy score van 0.75

Voor de Iris-dataset gebruikte ik ook een *decision tree classifier* om de soort bloem (*Species*) te voorspellen op basis van metingen als *Sepal Length*, *Sepal Width*, *Petal Length* en *Petal Width*. Dit model presteerde uitstekend, met een nauwkeurigheid van 1.0. Dit komt omdat het model klein en gescheiden is. We hadden ook een linear regressie model gemaakt voor de petal\_length. Deze had ook een hele lage MSE score.

## Welk model werkte beter, en waarom?

Van alle modellen werkte de *decision tree op de Iris-dataset* het best. De klassen in deze dataset zijn overzichtelijk en goed gescheiden, wat ideaal is voor een decision tree. De *lineaire regressie* op de World Happiness dataset werkte ook goed, omdat het een geschikte methode is voor het voorspellen van een continue waarde zoals de Happiness Score.

## Uitdagingen bij voorbereiding en training

We vonden de iris te makkelijk en heel erg goed werkend. Hierdoor hebben we nog een ander dataset gebruikt om het ons wat moeilijker te maken. Hierdoor kunnen we ook wat beter vergelijken hoe de modellen werken op verschillende datasets.