# Workshop AI in smart industry - Iris dataset

Onno Huijgen

28 maart 2025

## Introductie

In dit eerste voorbeeld gaan we werken met de iris dataset. Deze bestaat uit afmetingen van 200 bloemen van drie verschillende types: **iris setosa**, **iris versicolor** en **iris virginica**. In dit eerste voorbeeld gaan we de data inladen, visualiseren en een model trainen om het type bloem te kunnen voorspellen.

## 1 Data inladen

De iris dataset kun je direct in Orange inladen

1.1 Open de Orange tool en begin een nieuw project.

1.2 Voeg een **Datasets** module toe aan je canvas. Deze kun je vinden in de toolbar aan de linkerkant onder het kopje **Data**. De module toevoegen kan door erop te dubbelklikken of door hem naar het canvas te slepen.

1.3 Dubbelklik op de **Datasets** module. Je krijgt een overzicht met alle datasets die al automatisch beschikbaar zijn.

1.4 Laad de iris dataset in door in de zoekbalk **iris** te zoeken en selecteer de juiste dataset.

1.5 De data is nu ingeladen. Je kunt de module **Datasets** nu sluiten.

De modules in Orange zijn losse gereedschappen. Modules kunnen data als input krijgen, doen er dan iets mee en kunnen de output doorsturen aan andere modules. Zo ontstaat er een pipeline van bewerkingen.

1.6 Op dit moment hebben we de data wel ingeladen, maar is het nog niet inzichtelijk voor ons als gebruiker hoe de data eruit ziet. Om de data inzichtelijk te maken, kun je de module **Data Table** toevoegen. Deze vind je ook in de linkerbalk onder het kopje **Data**.

1.7 Sleep een verbinding van de rechterkant (output) van de dataset naar de linkerkant (input) van de data table.

1.8 Dubbelklik de data table. Je kunt nu zien hoe de data eruit ziet. De verschillende kolommen markeren de verschillende eigenschappen van de bloemen die gemeten zijn, ook wel de *features*. De rijen markeren alle bloemen die gemeten zijn in deze dataset.

## 2 Data visualiseren

De modules zijn in de toolbar aan de linkerkant ingedeeld in verschillende categorieën. Nu we de data hebben ingeladen kunnen we visualisaties maken.

2.1 Open de tab **Visualize** in de toolbar aan de linkerkant en selecteer voeg de module **Distributions** toe aan het canvas.

2.2 Sleep een verbinding direct van de dataset naar de distributions.

2.3 Dubbelklik de distributions.

Je krijgt meteen een visualisatie te zien van de verdeling van de data. Aan de linkerkant kun je de instellingen van de visualisatie aanpassen. Je kunt selecteren welke feature je wilt weergeven en hoe fijn die worden opgedeeld in verschillende ranges.

2.4 Speel met de instellingen van de visualisatie. Kun je een feature vinden waarmee je één soort iris direct kunt herkennen?

2.5 Voeg nu ook de module **Scatter Plot** toe. Sleep een verbinding van de **Dataset** direct naar **Scatterplot** en open de interface van de **Scatterplot**.

2.6 Speel op dezelfde manier met de instellingen van **Scatterplot**. Kun je een features vinden waarvoor een gedeelte van de data *linear separeerbaar* is? Dat wil zeggen dat je een van de bloemsoorten kunt afscheiden van de rest door een rechte scheidingslijn door de data heen te trekken. *Hint:* Orange kan je helpen met het zoeken naar inzichtelijke combinaties van features.

2.7 Sleep nu ook een lijn van **Data Table** naar **Scatter Plot**. **Scatter plot** heeft nu twee inputs: de **Dataset** module en de **Data Table** module.

2.8 Selecteer alleen alle *Iris setosa* punten in de **Data Table**. Wat gebeurt er met de scatterplot?

2.9 Selecteer weel alle data in **Data Table**. Zoek in de **Scatter plot** naar de meest inzichtelijke twee features en selecteer deze.

# 3 Model trainen

Nu het wat inzichtelijker is met wat voor data we te maken hebben kunnen we een model trainen. Je kunt in de toolbar links verschillende Machine Learning modellen vinden onder het kopje **Model**.

3.1 Selecteer het model **Tree** en voeg het toe aan je canvas.

3.2 Voeg ook de module **Test and Score** onder het kopje **Evaluate** toe aan je canvas.

3.3 Sleep een lijn van je **Dataset** naar **Test and Score**

3.4 Sleep een lijn van **Tree** naar **Test and score**

3.5 Dubbelklik op **Test and score**. Deze module laat zien hoe goed het machine learning model scoort op de dataset door middel van een paar verschillende metrieken: Area Under the Receiver Operator Curve (AUC of Area under ROC), Class Accuracy (CA), F1 score, Precision, Recall en Matthew's Correlation Coefficient (MCC). Zoek van de metrieken die je niet kent op wat ze betekenen.

3.6 Speel als je tijd over hebt met wat andere modellen. Probeer van modellen die je niet kent op te zoeken hoe ze ongeveer werken.