atheneum Dendermonde

Naam:…… Datum: ……

Klas:…… Samengewerkt met:……

Nabootsing van de ontwikkeling van een machine learning model gebaseerd op regressie

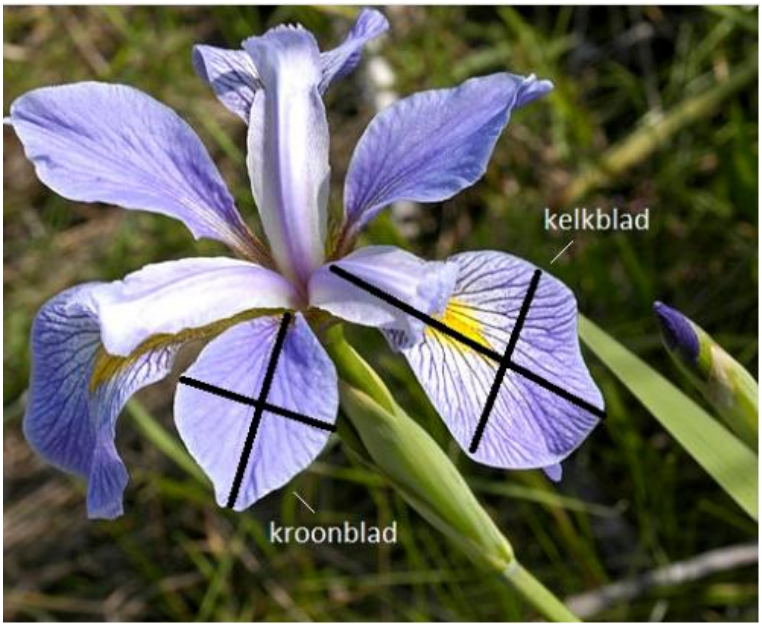
1. ***Opdracht***

Onderzoek de correlatie tussen twee datasets, bepaal een regressielijn en gebruik die om voorspellingen te doen betreffende nieuwe data.

Zo'n regressielijn zou het algoritme kunnen zijn van een eenvoudig **AI-systeem** dat in staat is om uit een bepaald gegeven, een specifieke waarde te voorspellen.

2. Context

De Iris-dataset werd in 1936 door de Brit Ronald Fischer gepubliceerd. De dataset betreft drie soorten irissen (Iris setosa, Iris virginica en Iris versicolor), 50 monsters van elke soort. Fischer kon de soorten van elkaar onderscheiden afgaande op vier kenmerken: de lengte en de breedte van de kelkbladen en de kroonbladen.



Jullie zullen focussen op de data van één van de drie soorten namelijk de Iris virginica.
De data zijn verzameld in een Exceltabel welke jullie kunnen downloaden van smartschool.
Jullie werken voor het uitvoeren van de studie verder met dit bestand in **Excel**.
Onderzoek de correlatie tussen de lengte van de kelkblaadjes en de kroonblaadjes van de Iris virginica. Bepaal een regressielijn en gebruik die om de lengte van een kroonblad te voorspellen vertrekkende van de lengte van het kelkblad van dezelfde bloem.

3. Stappenplan

I. De data weergeven in een puntenwolk

Je zal de relatie tussen de lengte van het kelkblad en de lengte van het kroonblad bestuderen.

Daarvoor zet je op een grafiek de lengte van het kroonblad uit in functie van de lengte van het kelkblad. De lengte van het kroonblad komt dus op de y-as en de lengte van het kelkblad op de x-as.

How to?

Selecteer de kolommen van de data (incl hoofding) die je in de grafiek wil voorstellen.

Kies op het tabblad 'Invoegen' voor Grafieken en selecteer de spreidingsgrafiek.

Plaats de grafiek in een apart tabblad met de naam 'spreidingsgrafiek ruwe data'. Geef de grafiek dezelfde titel als het tabblad en zorg voor een titel van de x-as en de y-as (door op het tabblad Grafiekontwerp te kiezen voor Grafiekelement invoegen).

Vragen:

a) **Vermoed je dat er een verband is tussen de beide parameters als je de puntenwolk ziet? Zo ja, welk verband?**

b) **Welke soort van regressie is in dit geval zinvol?**

II. Regressielijn berekenen

Excel berekent voor jou de regressielijn.

How to?

Selecteer een punt in de puntenwolk, klik op de rechter muisknop en kies trendlijn toevoegen. Omdat je vermoed dat er een lineair verband bestaat tussen de gegevens kies je voor een lineaire trendlijn. Vervolgens vink je aan dat de vergelijking van de trendlijn in de grafiek wordt weergegeven.

Om te benadrukken dat de regressielijn dient om voorspellingen te doen, teken je ze langer op de grafiek. In de opties voor de trendlijn kan je de voorspelling vooruit en achteruit instellen. Geef voor beide parameters waarde 1 in.

Sla jullie excelbestand op **in de map Regressie** en **benoem** als volgt:
'klasnaam'-team-'initialen van de namen van de teamleden'. Bvb: **NAWEx-team-XY-RS-UV**

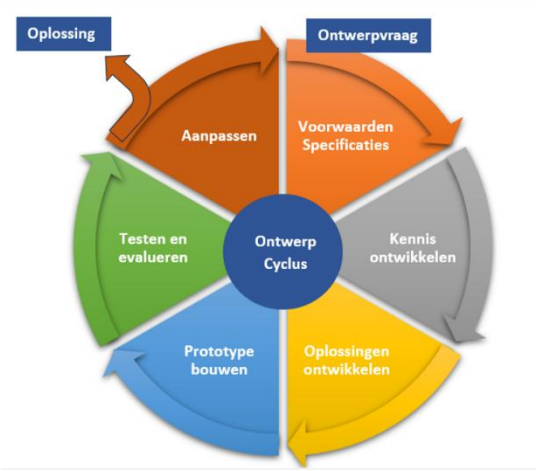
III. Voorspellingen doen met het AI-systeem

Vraag: waarom kunnen we deze regressielijn benoemen als een eenvoudig AI-systeem?

Vraag: **voorspel de lengte van het kroonblad voor een kelkblad van lengte 5.**

4. Ontwerpcyclus

Vraag: Duid aan tot welke fase van de ontwerpcyclus bovenstaande opdracht behoort?



IV. De data standaardiseren

De correlatie tussen bivariate gegevens kan men visueel inschatten door de overeenkomstige puntenwolk te bekijken. De vorm hiervan is echter maar betrouwbaar als de gegevens gestandaardiseerd zijn. Standaardiseren van variabelen houdt in dat men de variabelen op zo'n manier herschaalt dat men variabelen van bv. een verschillende grootteorde of in een verschillende eenheid met elkaar kan vergelijken of in verband brengen.

Om te standaardiseren wordt er overgegaan op de Z-scores van de kenmerken. Voor het berekenen van de Z-score van de gegevens, trek je het gemiddelde af en je deelt vervolgens door de standaardafwijking.

Na standaardisatie liggen de meeste waarden tussen 0 en 1, wat voordelig is voor het rekenen door de computer. Rekenen met vrij grote getallen leidt al snel tot nog grotere getallen en tot numerieke instabiliteit, dat is een bijkomende reden waarom de data worden gestandaardiseerd.

How to?

Werk verder in het excelbestand van jouw team **NAWEx-team-XY-RS-UV**.

Je past standaardisatie toe op de data in beide kolommen. Dit doe je door voor elke kolom afzonderlijk eerst het gemiddelde te berekenen (formule: '=Gemiddelde(C2:C51)') en vervolgens ook de standaarddeviatie (formule: '=stdev(C2:C51)'). Bereken het gemiddelde en de standaarddeviatie in de daarvoor voorziene cellen op rij 53 en rij 54

Vervolgens wordt de gestandaardiseerde waarde berekend in kolom F voor de lengte van het kelkblad en in kolom G voor de lengte van het kroonblad. De formule voor standaardisatie staat reeds klaar (normaliseren(x;gemiddelde;standaarddev))

Ter info: De functie 'normaliseren' gebruikt de volgende formule om een bepaalde gegevenswaarde te normaliseren: Genormaliseerde waarde = (x – $\bar{x}$) / s met x = gegevenswaarde, $\bar{x}$ = gemiddelde van de dataset en s = standaardafwijking van de dataset

V. Gestandaardiseerde data weergeven in een puntenwolk

Op dezelfde manier als in stap I. zet je nu op een grafiek de Z-score van de lengte van het kroonblad uit in functie van de Z-score van de lengte van het kelkblad.

How to?

Selecteer de kolommen van de data (incl hoofding) die je in de grafiek wil voorstellen. Kies op het tabblad 'Invoegen' voor Grafieken en selecteer de spreidingsgrafiek. Plaats de grafiek

in een apart tabblad met de naam ‘spreidingsgrafiek genormalis.’. Geef de grafiek dezelfde titel als het tabblad en zorg voor een titel van de x-as en de y-as (door op het tabblad Grafiekontwerp te kiezen voor Grafiekelement invoegen).

VI. Regressielijn berekenen

Excel berekent voor jou de regressielijn.

How to?

Selecteer een punt in de puntenwolk, klik op de rechter muisknop en kies trendlijn toevoegen. Omdat je vermoed dat er een lineair verband bestaat tussen de gegevens kies je voor een lineaire trendlijn.

Vervolgens vink je aan dat de vergelijking van de trendlijn in de grafiek wordt weergegeven.

Om te benadrukken dat de regressielijn dient om voorspellingen te doen, teken je ze langer op de grafiek. In de opties voor de trendlijn kan je de voorspelling vooruit en achteruit instellen. Geef voor beide parameters waarde 1 in.

Niet vergeten: sla jullie excelbestand zeker nogmaals op **in de map Regressie**

VII. Voorspellingen doen met het AI-systeem

Vraag:

> **Voorspel de gestandaardiseerde lengte van het kroonblad voor een kelkblad van gestandaardiseerde lengte 3.**

Verband met Machine Learning

Om een rechte te vinden die goed bij de gegeven data past, kan een machine learning-systeem vertrekken van een willekeurig gekozen rechte. Het systeem wordt getraind met de trainingset (de inputs en de corresponderende outputs): voor elk punt van de trainingset wordt nagegaan hoeveel de corresponderende y-waarde op de voorlopige rechte afwijkt van de gegeven y-waarde. De coëfficiënten in de vergelijking van de rechte worden aangepast, zodat de gemiddelde kwadratische afwijking voor de hele dataset minimaal is.

Ter info: AI dankzij machine learning en neurale netwerken

Wie meer wil weten over machine learning en neurale netwerken, verwijzen we naar de tekst ‘AI dankzij machine learning en neurale netwerken’ welke je terugvindt op smartschool.
Bron: https://dwengo.org/kiks/ - leerpad regressie