Gevorderde Verkennende Analyse

# Inleiding

Eerder heb je al een basis verkennende analyse gedaan op de Student Performance dataset. Nu doen we een gevorderde analyse. Maar deze keer kijken we naar de volledige dataset. Je krijgt de dataset in de les. Dat is een aangepaste variant op de dataset(s) die je hier vindt: <https://archive.ics.uci.edu/ml/datasets/student+performance>

Het belangrijkste verschil met de eerdere opdracht is dat nu alle 34 variabelen in de dataset zitten. Daarom werkt de systematische aanpak van de basis analyse niet meer zo goed. Er zijn bijvoorbeeld veel te veel mogelijke tweetallen om alle relaties tussen twee variabelen afzonderlijk te bekijken en interpreteren.

In deze opdracht verken je systematisch de hele dataset. Maar je werkt wel met een doel in je achterhoofd. Namelijk dat je aansluitend een machine learning oplossing wilt maken om de afhankelijke variabele G3 te voorspellen

# Data, metadata, variabelen indelen

1. Lees de data in, toon de data, en bekijk de datatypes die pandas toekent.
2. Interpreteer de output van bovenstaande acties. Je kunt eventueel verwijzen naar <https://archive.ics.uci.edu/ml/datasets/student+performance>.
3. Deel de variabelen in. Hint: Het is voor de Student Performance dataset het meest werkbaar om een onderscheid te maken tussen de volgende soorten: afhankelijke variabelen, (onafhankelijke) categorische variabelen en (onafhankelijke) numerieke variabelen.
4. Zorg ervoor dat je notebook een verhaal vertelt.

# Individuele variabelen

1. Onderzoek de afhankelijke variabelen. De afhankelijke variabelen zijn G1, G2 en G3. Bekijk ze alle drie afzonderlijk: maak een frequentieverdeling en interpreteer het resultaat.
2. Onderzoek de categorische variabelen:
   1. Plot in een keer alle frequentieverdelingen. Je kunt gebruiken maken van de code die in de les is gedemonstreerd. Bonus: Pas deze functie aan naar je eigen wensen. Of maak zelf een soortgelijke oplossing.
   2. Interpreteer de figuur. Benoem een aantal dingen die je opvallen.
3. Onderzoek de numerieke variabelen:
   1. Plot wederom alle frequentieverdelingen.
   2. Interpreteer de figuur. Benoem een aantal dingen die je opvallen. Ga met name in op hoe variabelen verdeeld zijn.
4. De verkennende analyse doen we met in ons achterhoofd een machine learning probleem, namelijk het voorspellen van een cijfer met een regressiemodel. Relateer je interpretatie waar mogelijk aan dit machine learning probleem.
5. Zorg ervoor dat je notebook een verhaal vertelt.

# Missing values

1. Analyseer de missing values in de dataset. Maak daarbij met name een onderscheid tussen nullable en not nullable variabelen.
2. Formuleer een strategie om de missing values te vullen. Houdt daarbij ook rekening met het volgende: een machine learning model is uiteindelijk bedoeld om voorspellingen te doen op nieuwe data. In deze nieuwe data kunnen andere missing values zitten dan in de dataset waar je nu mee werkt. Je moet er dus vanuit gaan dat elke variabele een missing value kan hebben.
3. Maak een Estimator-Transformer die de missing values vult. Hint: let goed op welke stuk in de functie fit() moet, en welk stuk in de functie transform().
4. Zorg ervoor dat je notebook een verhaal vertelt.

# Bonus: nog een Estimator-Transformer

Bouw naar eigen inzicht nog een Estimator-Transformer. Misschien wil je bijvoorbeeld een log transformatie doen, of misschien heb je zelf nog een ander idee.

# Combinaties van variabelen

## De afhankelijke variabelen onderling

Deze dataset heeft drie afhankelijke variabelen: G1, G2 en G3.

1. Onderzoek de relaties tussen deze variabelen. Je kunt bijvoorbeeld een pairplot maken of een correlatiematrix. Of misschien wil je wel een tweetal apart nemen.
2. Interpreteer de resultaten.
3. Breng je verhaal ook in verband met het machine learning probleem: namelijk het voorspellen van G3 op basis van de onafhankelijke variabelen (dus zonder G1 en G2).
4. Zorg ervoor dat je notebook een verhaal vertelt.

## De numerieke onafhankelijke variabelen

1. Maak een correlatiematrix en/of een pairplot van alle numerieke onafhankelijke variabelen. Zoek naar een representatie van de relaties die werkbaar is. Je moet lees- en interpreteerbare output produceren.
2. Interpreteer je resultaten. Zoom waar nodig in op tweetallen van variabelen.
3. Relateer deze analyse aan het machine learning probleem.
4. Zorg ervoor dat je notebook een verhaal vertelt.

## Bonus: de categorische onafhankelijke variabelen

Het analyseren van relaties met categorische variabelen is lastig als het er veel zijn. Maar:

1. Bedenk een creatieve oplossing om de relaties tussen de categorische variabelen onderling te onderzoeken.
2. Bedenk een creatieve oplossing om de relaties tussen tweetallen van numeriek en categorisch te onderzoeken.
3. Interpreteer je resultaten. Zoem waar nodig in op tweetallen van variabelen.
4. Relateer deze analyse aan het machine learning probleem.
5. Zorg ervoor dat je notebook een verhaal vertelt.

## De afhankelijke variabele G3 en de onafhankelijke variabelen

De dataset heeft drie afhankelijke variabelen: G1, G2 en G3. Wij bekijken alleen G3 omdat dat ook de variabele is die je later met een machine learning algoritme wilt voorspellen. Je kunt in deze stap gebruikt maken van de functie die in de les is gedemonstreerd waarmee je één (afhankelijke) variabele vergelijkt met alle onafhankelijke variabelen.

1. Vergelijk de afhankelijke variabele G3 in één figuur met alle onafhankelijke numerieke variabelen.
2. Bonus: Bereken de correlatie coëfficiënten tussen G3 enerzijds en de onafhankelijke numerieke variabelen anderzijds.
3. Vergelijk de afhankelijke variabele G3 in één figuur met alle onafhankelijke categorische variabelen.
4. Bonus: Gebruik een Kruskal-Wallis of ANOVA test om te toetsen hoe afhankelijk G3 is van de verschillende categorische variabelen.
5. Interpreteer alle output. Benoem de variabelen waarvan je ziet dat ze het meest bepalend zijn voor G3. Ga ook (kort) in op de vraag waarom zo’n variabele invloed zou kunnen hebben. Relateer deze analyse aan het machine learning probleem.
6. Zorg ervoor dat je notebook een verhaal vertelt.

# Afronding

1. Geef een opsomming van de observaties die je hebt gedaan die betrekking hebben op het voorspellen van de afhankelijke variabele G3 met behulp van machine learning. Zorg ervoor dat de conclusie van deze verkennende analyse een bruikbaar startpunt is voor het daadwerkelijk maken van zo’n machine learning oplossing.
2. Tenslotte: zorg ervoor dat je notebook als geheel een verhaal vertelt. Denk ook aan een titel, een inleiding, en een passende afbeelding.