Basis verkennende data analyse

# Inleiding

Een onderdeel van data analyse is het doen van een verkennende analyse op je dataset. Bij de verkennende analyse kijk je welke variabelen in de dataset zitten. Je analyseert bovendien de waarden van die variabelen. Tenslotte onderzoek je de relatie tussen (relevante) combinaties van variabelen.

In deze opdracht ga je een verkennende data analyse doen op de Student Performance dataset. Deze dataset gaat over de prestaties van studenten op een Portugese school. Het bevat cijfers en gegevens over de studenten. Zie <https://archive.ics.uci.edu/ml/datasets/student+performance>. Wij gebruiken een aangepaste variant van deze dataset. Daarin zitten minder variabelen en ook een aantal andere aanpassingen. Je krijgt de dataset in de les.

In de volgende paragrafen krijg je opdrachten om deze verkennende analyse stap voor stap op de te bouwen.

# Data en data dictionary

We gaan ervanuit dat je een ontwikkelomgeving hebt waarin je Jupyter Notebooks kunt maken.

In deze stap ga je de dataset inlezen en een data dictionary maken. In je data dictionary geeft je uitleg over de variabelen in je dataset.

1. Lees de dataset in. Je kunt het bestand vinden in de digitale leeromgeving.
2. Toon het dataframe.
3. Toon ook de lijst met variabelen en de datatypes die pandas eraan heeft toegekend.
4. Maak een data dictionary. Je kunt gebruik maken van het bestand student.txt op <https://archive.ics.uci.edu/ml/datasets/student+performance>. Gebruik de mogelijkheid om in markdown een tabel te maken. Geef per variabele:
   1. Een korte omschrijving van de variabele;
   2. Het datatype van de variabele;
   3. Of de variabele numeriek of categorisch is;
   4. Het soort variabele: nomimaal, ordinaal, interval, ratio;
   5. De mogelijke waarden die de variabele kan aannemem.
5. Voeg markdown cellen toe om ervoor te zorgen dat je notebook een verhaal vertelt. Gebruik indien nodig headers, lijsten, tabellen, plaatjes.

# Afzonderlijke variabelen

In deze stap ga je de afzonderlijke variabelen bekijken.

Het is daarbij nodig om een onderscheid te maken tussen categorische variabelen en numerieke variabelen. In je data dictionary kun je teruglezen of een variabele numeriek of categorisch is.

1. Onderzoek de categorische variabelen. Per categorische variabele:
   1. Toon welke waarden de variabele aan kan nemen en hoe vaak die waarde voorkomt.
   2. Maak een grafiek met een frequentieverdeling.
   3. Schrijf tekst waar je deze waarden en grafieken interpreteert.
2. Onderzoek de numerieke variabelen. Per numerieke variabele:
3. Toon maten van spreiding en ligging.
4. Maak een histogram.
5. Kijk of de variabele normaal verdeeld is. Zo niet, kijk of je de variabele zodanig kunt transformeren dat ie dat wel is.
6. Maak een boxplot om te kijken of er outliers zijn.
7. Schrijf tekst waar je deze waarden en grafieken interpreteert.
8. Zorg ervoor dat je notebook een verhaal vertelt.

# Relaties tussen variabelen

In deze stap gaan we kijken naar de relaties tussen variabelen. We kijken iedere keer naar tweetallen. Het is nodig onderscheid te maken tussen de verschillende mogelijke combinaties van numerieke en categorische variabelen.

1. Onderzoek relaties tussen categorische variabelen. Per tweetal:
   1. Maak een gelabeld staafdiagram. Op de x-as de ene variabele, op de y-as het aantal, als label de andere variabele.
   2. Bonus: Doe een Chi kwadraat test.
   3. Interpreteer je resultaten.
2. Verken de relaties tussen tweetallen numerieke variabelen.
   1. Maak een correlatiematrix.
   2. Maak eventueel een correlatieheatmap.
   3. Maak een pairplot met scatterplots.
   4. Interpreteer je resultaten.
   5. Kies twee numerieke variabelen waarvan je vermoedt dat ze een sterk lineair verband hebben. Je kunt gebruik maken van het pairplot met scatterplots die je in de zojuist hebt gemaakt. Je kunt ook gebruik maken van de heatmap van correlatiecoëfficiënten. Of je kunt kijken naar de namen van de variabelen en op grond van de betekenis een verband vermoeden.
   6. Maak voor de twee variabelen die je hebt gekozen een scatterplot met daarin een regressielijn. Bonus: bereken de R2 en geef een interpretatie van de waarde.
   7. Kies nog een tweetal, nu eentje waarvan je vermoedt dat ze nauwelijks een lineair verband hebben. Herhaal stap f voor dit tweetal.
   8. Kies nog een tweetal, nu eentje waarvan niet zo duidelijk of er een sterk lineair verband is, of nauwelijks een verband. Iets tussen je twee vorige tweetallen in dus. Herhaal opnieuw stap f voor dit tweetal.
3. Onderzoek relaties tussen alle mogelijke tweetallen van numerieke en categorische variabelen. Per tweetal:
   1. Maak een boxplot. Bonus: Experimenteer met varianten op de boxplots.
   2. Vergelijk de gemiddelden van de numerieke variabele per waarde van de categorische variabele.
   3. Bonus: doe een ANOVA test of een Kruskal-Wallis test.
   4. Interpreteer je resultaten.
4. Bonus: onderzoek ook de relatie tussen drietallen variabelen.
5. Zorg ervoor dat je notebook een verhaal vertelt.

# Afronding

1. Wat zijn op dit moment je belangrijkste conclusies over deze dataset? Schrijf deze op in je notebook.
2. Welke vervolgvragen heb je over deze dataset? Schrijf deze op in je notebook.
3. Tenslotte: zorg ervoor dat je notebook als geheel een verhaal vertelt. Denk ook aan een titel, een inleiding, en een passende afbeelding.