PORTFOLIO-OPDRACHT

**Introductie**

Het mobiliteitsgedrag van burgers speelt een sleutelrol bij de uitstoot van broeikasgassen en het behalen van de nationale klimaatdoelstellingen. Om gericht beleid te kunnen ontwikkelen, is inzicht nodig in hoe persoonlijke kenmerken, woonlocatie en mobiliteitskeuzes samenhangen met de individuele CO₂-uitstoot. Deze portfolio-opdracht biedt de mogelijkheid om dit complexe vraagstuk vanuit een datawetenschappelijk perspectief te verkennen.

**Vraagstellingen**

Welke persoonlijke en mobiliteitsgerelateerde factoren verklaren verschillen in de jaarlijkse CO₂-uitstoot tussen burgers?

In welke mate kunnen we individuele uitstoot nauwkeurig voorspellen op basis van een combinatie van sociaaldemografische kenmerken en mobiliteitsgedrag?

**Beschrijving van de dataset**

De dataset bevat informatie van 1500 burgers. Voor elke persoon zijn de volgende gegevens beschikbaar:

* Sociaaldemografische kenmerken: geslacht, leeftijd, gewicht, burgerlijke staat, jaarsalaris, grootte van het huishouden, opleidingsniveau.
* Woonlocatie: stad en provincie.
* Mobiliteitsmiddelen: bezit van auto, elektrische auto en fiets.
* Mobiliteitsgedrag: jaarkilometers per vervoerswijze (auto, fiets, openbaar vervoer, vliegtuig), aantal thuiswerkdagen per week, aantal dagen per jaar in het buitenland, gebruik van carsharing.
* Jaarlijkse totale CO₂-uitstoot

**Het is belangrijk voor de opdrachten dat je bij elk antwoord aangeeft waarom je een keuze maakt of een toelichting geeft op wat je doet. Enkel eindantwoorden of stukken code zijn 0 punten waard**. Hieronder zie je het beoordelingsformulier dat voor elke vraag gebruikt wordt:

|  |  |
| --- | --- |
| Beoordeling: | Criteria: |
| 0% | Er is **geen** toelichting gegeven of de argumentatie, het antwoord of de gebruikte methode bevat grote fouten. |
| 50% | Er is toelichting gegeven. De argumentatie, het antwoord of de gebruikte methode bevat kleine fouten. |
| 100% | Er is toelichting gegeven. De argumentatie, het antwoord of de gebruikte methode bevat (bijna) geen fouten. |

**Bronvermelding:**

Bij gebruik van (GenAI) bronnen moet hier conform de richtlijnen naar worden verwezen. Deze richtlijnen staan beschreven op de BrightSpace pagina van DataLab ll.

Deel A: Data exploratie en voorbereiding

1. Geef voor elke variabele in de dataset weer tot welk meetniveau deze behoort.
2. Waarom is het nodig om vooraf te bekijken tot welk meetniveau elke variabele behoort?
3. Ga na of er missende/niet kloppende/extreme waarden zijn in de dataset en ga hier op een adequate manier mee om.
4. Kies 2 onafhankelijke variabelen waarvan je denkt dat het invloed kan hebben op de jaarlijkse CO2 uitstoot. Visualiseer de verdeling van deze twee variabelen en de verdeling van de afhankelijke variabele.
5. Geef een beschrijving van de 3 visualisaties met behulp van statistische termen.
6. Toets, met behulp van statistiek, of de verdelingen (vraag 4) normaal verdeeld zijn.
7. Om ervoor te zorgen dat de onafhankelijke variabelen gebruikt kunnen worden voor Machine Learning moeten ze getransformeerd worden. Geef per meetniveau aan wat er moet gebeuren en waarom.

Voer de transformatie uit. **Let op** dat je de getransformeerde variabelen in een nieuw dataframe opslaat, zodat je de oorspronkelijke waarden behoudt. De getransformeerde variabelen gebruik je alleen bij deel D.

1. Zorg dat je inzicht krijgt in de relatie tussen de variabelen. Wat zijn je bevindingen?
2. Welke methode(n) heb je gebruikt en waarom?
3. Is er sprake van multicollineariteit? Waarom zou dit een probleem kunnen zijn en wat zou jouw oplossing zijn?
4. Bekijk de relatie tussen uitstoot CO2 en de andere onafhankelijke variabelen per geslacht. Kom je hier tot een andere conclusie dan bij vraag 8?

Deel B: Kansrekenen

1. Wat is de kans dat een willekeurig persoon 5000 KG CO2 uitstoot? Gebruik de standaard normaal verdeling in de berekening.
2. Op welke manier wordt, bij deze berekening, een van de technieken van calculus toegepast?
3. Is deze kans groter, gelijk of kleiner bij mannen t.o.v. vrouwen?

Deel C: Toetsende statistiek

Kies een variabele uit de dataset waarmee de dataset gegroepeerd kan worden (bijv. Mannen vs. Vrouwen). Je mag **niet** geslacht kiezen. Je gaat onderzoek doen of er een verschil zit in de uitstoot van CO2 tussen de groepen die je hebt gekozen.

1. Stel de hypothesen op
2. Welke type fouten kunnen er gemaakt worden?
3. Welke statistische toets moet je uitvoeren om te onderzoeken of er een significant verschil is?
4. Voer de toets uit.
5. Geef in je eigen woorden een toelichting op de betekenis van de p-waarde.
6. Welke conclusie kan je trekken?

Deel D: Modelleren

Split de data nu op in een train- en test set. Kies zelf voor een geschikte split.

1. Op hoeveel verschillende mogelijkheden kan deze split gemaakt worden?

Train een model om de jaarlijkse CO2 uitstoot te voorspellen

1. Welke metric zou je gebruiken om te toetsen hoe goed je model is en waarom?
2. Geef de wiskunde functie van het model en pas deze toe op jouw eigen situatie (kies 1 groepsgenoot)

Kies de vijf onafhankelijke variabelen die de meeste invloed hebben op de voorspelling. Train en test nu twee modellen (1: met alle variabelen en 2: met de vijf meest invloedrijke variabelen). Kies 1 metric om de nauwkeurigheden van de modellen te kwantificeren. Gebruik hiervoor 20-fold crossvalidation.

1. Gebruik toetsende statistiek en de resultaten van de 20-fold cross validation om te onderzoeken welk model significant beter is.

Splits de target variabele op in 2 klassen, kies hiervoor een geschikte punt. Codeer de eerste groep als 0 en de tweede groep als 1.

1. Welk model zou je nu kunnen gebruiken en waarom?

Train een nieuw model op de nieuwe target, gebruik hierbij alleen de vijf meest invloedrijke onafhankelijke variabelen die je eerder hebt gekozen.

1. Geef de wiskunde functie van dit model en pas deze toe op jouw eigen situatie (kies 1 groepsgenoot)
2. Schrijf een conclusie op basis van de twee hoofdvragen van het portfolio (zie kopje ‘Vraagstellingen’)