# Opdracht 2: Voorspel verhuuraantallen

Inhoud

[Opdracht 2: Voorspel verhuuraantallen 1](#_Toc147494077)

[1. Inleiding 1](#_Toc147494078)

[2. Data 1](#_Toc147494079)

[2. Opdrachtbeschrijving 2](#_Toc147494080)

[Vereisten notebook 2](#_Toc147494081)

[Opdracht 1: Exploratieve Data Analyse (10 punten) 2](#_Toc147494082)

[Opdracht 2: Time series feature engineering (35 punten) 2](#_Toc147494083)

[Opdracht 3: Modelleren (45 punten) 3](#_Toc147494084)

[Opdracht 3: Bevindingen (5 punten) 3](#_Toc147494085)

[Opdracht 4: Conclusie en aanbevelingen (5 punten) 3](#_Toc147494086)

[Bijlage 1: Bronvermelding en regels gebruik Generative AI (GAI) 4](#_Toc147494087)

## 1. Inleiding

In deze opdracht voorspel je het aantal stuks dat per uur wordt verhuurd van een product. Na afronding van deze opdracht vertellen we om welk product het gaat.

## 2. Data

**train.csv:** Hierin vind je in de targetkolom ‘cnt’ informatie over het aantal stuks dat per uur is verhuurd. Daarnaast vind je de volgende kolommen:

* date\_hour: Je hebt informatie over de periode van 1-1-2011 t/m 30-11-2012
* holiday: Vakantiedag of geen vakantiedag
* weathersit: Weersituatie:
  1. Helder, licht bewolkt, deels bewolkt
  2. Mistig , mistig en licht bewolkt
  3. Lichte sneeuw, lichte regen, lichte regen en onweer, zwaar bewolkt, lichte regen en zwaar bewolkt
  4. Zware regen, hagel, zware mist, sneeuw.
* temp: genormaliseerde temperatuur
* atemp: genormaliseerde gevoelstemperatuur
* hum: genormaliseerde luchtvochtigheid
* windspeed: genormaliseerde windsnelheid.

Dit bestand gebruik je om je modellen te trainen, te testen en te verbeteren. Het beste model, met de beste parameters, pas je toe op het bestand test.csv.

**test.csv:** Hierin vind je dezelfde kolommen als in train.csv maar nu over de periode van 1-12-2012 t/m 19-12-2012 en de “cnt” kolom ontbreekt. Het getrainde model pas je toe op de gegevens in dit bestand. De voorspellingen die hieruit komen gebruik je om in te leveren op Kaggle. De voorspellingen lever je in met een csv-bestand zoals submission\_example.csv.

**sample\_submission.csv:** De eerste kolom bevat de ‘date\_hour’ kolom van test.csv. Nadat je je voorspellingen hebt gedaan plaats je deze in de kolom ‘cnt’ van dit bestand. Vervolgens kun je dit bestand oploaden op Kaggle.

## 2. Opdrachtbeschrijving

Je levert als team een Jupyter notebook in dat voldoet aan onderstaande voorwaarden en waarin je onderstaande opdrachten uitvoert:

### Vereisten notebook

Notebooks die niet voldoen aan onderstaande voorwaarden worden niet nagekeken. In dit geval moet je gebruik maken van de herkansing om een cijfer voor deze opdracht te krijgen.

* Lever één net en duidelijk gestructureerd notebook in op Github.
* Structureer het notebook met markdown cellen en nummer de hoofdstukken en paragraven.
* Gebruik markdown cellen voor tekst en code cellen voor code.
* Gebruik zoveel mogelijk zelf-gedefinieerde functies en bij voorkeur OOP.
* De code voldoet aan PEP8 inclusief “comments”.
* Alle groepsgenoten begrijpen alle code en de teksten die worden ingeleverd en zijn in staat deze toe te lichten als daarom wordt gevraagd.
* Refereer voor zowel de tekst als voor de code op de juiste wijze aan gebruikte bronnen (zie bijlage 1).

### Opdracht 1: Exploratieve Data Analyse (10 punten)

* Ieder teamlid maakt een account op Kaggle aan en vormt een team dat deelneemt aan deze competitie: [Voorspel verhuuraantallen 2023 | Kaggle](https://www.kaggle.com/competitions/voorspel-verhuuraantallen-2023)
* Noteer je teamnaam, jullie namen en alle bijhorende Kaggle gebruikersnamen in het notebook.
* Voer een Exploratieve Data Analyse (EDA) uit:
  + Toon de datatypes en basisstatistieken van iedere kolom.
  + Voeg tijdserie elementen toe en gebruik passende visualisaties om relevante patronen zichtbaar te maken.
  + Onderzoek de relaties tussen de onafhankelijke variabelen enerzijds en de afhankelijke variabele anderzijds. Gebruik hierbij passende visualisaties.
  + Beschrijf wat de belangrijkste bevindingen zijn van de EDA.
* Sommige regels en/of kolommen zijn soms niet relevant voor machine learning of bevatten onjuiste of missende waardes. Schoon, zo nodig, de data op.

### Opdracht 2: Time series feature engineering (35 punten)

* Er zijn diverse soorten tijdseries features die je kunt toevoegen aan de data. Denk hierbij onder andere aan trend, seizoenspatronen en cycli.
* Voeg tijdserie features toe aan de dataset en onderzoek wat de invloed is van deze features, samen met de andere features, op de voorspellingen.
* De volgende zaken zijn hierbij duidelijk zichtbaar **onderzocht** en **uitgelegd**:
  + - Onderzoek naar seizoenspatronen met Fourier analyse.
    - Onderzoek naar trends
    - Onderzoek naar cycli en autocorrelatie
    - Onderzoek naar andere tijdsgebaseerde features.
* Deze opdracht hangt sterk samen met de opdracht 3 (modelleren). Je zult na het toevoegen van features moeten gaan modelleren om te achterhalen wat de invloed van de features is. Daarna zul je waarschijnlijk weer teruggaan naar deze opdracht om nieuwe features uit te proberen. Dit is een iteratief proces. Zorg ervoor dat je uitwerking duidelijk maakt hoe jullie tot de uiteindelijke keuze voor de features en model(len) zijn gekomen.

### Opdracht 3: Modelleren (45 punten)

* Onderzoek welke modellen de beste resultaten geven in combinatie met de gekozen features.
* Neem in dit onderzoek in ieder geval de volgende zaken mee:
  + Modellen waarbij je time series features combineert met reguliere features. Lineaire regressie en een ensemble model moeten onderdeel van deze modellen zijn. Pas lineaire regressie toe en geef een duidelijke (wiskundige) uitleg waarbij je onder andere de termen loss function en regularisatie gebruikt. Ook bij het ensemble model geef je een heldere (wiskundige) uitleg. De keuze voor de overige modellen is vrij maar dient wel onderbouwd te worden. We verwachten dat je minimaal 5 modellen toepast en uitlegt.
  + Twee specifieke tijdseriemodellen waarvan één SARIMA(X) is en de andere vrije keuze is. Pas beide modellen toe en leg de werking uit.
  + Hybride modellen: Dit is een aanpak waarbij je met het eerste model voorspellingen doet en vervolgens met een tweede model de zogenaamde residuals voorspelt. Pas een hybride model toe en leg de werking uit.
* Pas cross-validation en hyperparameter tuning toe op ieder model.
* Pas ieder model toe op de testset en upload de resultaten op Kaggle.

### Opdracht 3: Bevindingen (5 punten)

* Geef een samenvatting van de uitkomsten van het modelleren.
  + Geef een beknopt overzicht van de resultaten.
  + Welke model presteert het beste? Bij welke parameters?
  + Toon je scores op Kaggle en laat zien wat de resultaten waren van je verbeteringen op je score op Kaggle.

### Opdracht 4: Conclusie en aanbevelingen (5 punten)

* + Bepaal welke features het meeste invloed hebben op je voorspellingen
  + Beschrijf in hoeverre dit aansluit op de EDA.
  + Welk advies kun je op basis hiervan geven aan de verhuurder?
  + Welk model zou je adviseren aan de verhuurder? Waarom? Is dit per definitie het model met de beste voorspellingen of zijn er ook andere overwegingen? Zo ja, welke?

### Bijlage 1: Bronvermelding en regels gebruik Generative AI (GAI)

* Generative AI is een toegestane bron naast de reguliere bronnen, zoals wetenschappelijke artikelen. Eerder ingeleverd eigen werk is ook een bron.
* Als een stuk tekst of code is gebaseerd op een bron dan dient hieraan te worden gerefereerd volgens de APA-stijl. Als dit niet is gedaan dan wordt dat gezien als fraude.
* Teksten gebaseerd op bronnen worden geparafraseerd, oftewel in eigen woorden beschreven.
* Teksten mogen alleen in incidentele gevallen (maximaal 5% van de totale tekst) letterlijk worden overgenomen uit een bron. Een vrijwel letterlijke overname wordt ook gezien als een letterlijke verwijzing.
* Als er letterlijke tekst wordt overgenomen dan wordt dat zichtbaar gemaakt in APA-stijl.
* Code mag voor maximaal 50% van de totale code letterlijk worden overgenomen uit bronnen als ernaar wordt gerefereerd in APA-stijl. Houd hierbij wel rekening met de volgende zaken:
  + De code dient te voldoen aan de PEP8 richtlijnen.
  + De stijl van de code dient uniform te zijn. Bijvoorbeeld, de stijl van naamgeving en ‘comments’ moet consistent zijn in het hele notebook.
  + Ook code waarin kleine wijzigingen zijn aangepast worden gezien als letterlijk overgenomen.
* Als GAI de bron is dan dient de volgende informatie te worden verstrekt:
  + de naam van het language-model, versie en datum (meestal zal dit ChatGPT zijn, maar er zijn er nog meer)
  + het nummer van de prompt, dit maakt het makkelijk voor de docent om in de referentielijst de juiste prompt te vinden
  + de titel van de chat, in het geval van ChatGPT te vinden aan de linkerzijde
  + een link naar het gesprek, in het geval van ChatGPT is een link voor het gehele gesprek te vinden rechtsboven. Het is hierbij belangrijk dat de student op ‘Share your name’ klikt, zodat de gebruikersnaam van het account zichtbaar is voor de docent.
* Bronvermelding in het notebook in het geval van code:
  + Student gebruikt Prompt-referenties en neemt die op in een Jupyter notebook in zowel markdown-cellen als in comments bij de code waar ChatGPT bij is gebruikt.
* De code-cel bevat een comment met het promptnummer en de prompttitel en wordt als volgt geformat: #Prompt 1: Typen Vulkanen
* De markdowncel bevat de details van de vermelding, als volg geformat: [ChatGPT, 2023. Prompt 1: Typen vulkanen](<https://chat.openai.com/share/7967e049-f570-4498-b425-a1694ec14178)>.
* Bronvermelding in het notebook inhet geval van tekst:
  + Naast de schild- en stratovulkanen, bestaan er ook caldera’s, spleetvulkanen en lava plateaus (ChatGPT, 2023. Prompt 1: Typen vulkanen)
* Het notebook dat je inlevert bevat een referentielijst in APA-stijl met daarin alle bronnen waaraan je in tekst en code hebt gerefereerd. In het geval van de eerder gebruikte referentie aan ChatGPT ziet dat er als volgt uit.
  + ChatGPT, 2023. Prompt 1: Typen vulkanen. <https://chat.openai.com/share/7967e049-f570-4498-b425-a1694ec14178>.