# Opdracht 1: Verkoopprijzen van huizen voorspellen

Inhoud

[Opdracht 1: Verkoopprijzen van huizen voorspellen 1](#_bookmark0)

1. [Inleiding 1](#_bookmark1)
2. [Data 1](#_bookmark2)
3. [Opdrachtbeschrijving 2](#_bookmark3)

[Vereisten notebook 2](#_bookmark4)

[Opdracht 1: Exploratieve Data Analyse (5 punten) 2](#_bookmark5)

[Opdracht 2: Fully-connected neuraal netwerk (20 punten) 3](#_bookmark6)

[Opdracht 3: Convolutioneel neuraal netwerk from scratch (20 punten) 3](#_bookmark7)

[Opdracht 4: Transfer learning (20 punten) 3](#_bookmark8)

[Opdracht 5: Multimodaal model (20 punten) 4](#_bookmark9)

[Opdracht 6: Bevindingen (5 punten) 4](#_bookmark10)

[Opdracht 7: Conclusie en aanbevelingen (10 punten) 4](#_bookmark11)

[Bijlage 1: Bronvermelding en regels gebruik Generative AI (GAI) 5](#_bookmark12)

## Inleiding

In deze opdracht ga je de verkoopprijs voorspellen van huizen in de VS (meestal in de staat Californië, maar ook op sommige andere locaties). Deze prijzen komen uit het jaar 2016, dus het kan zijn dat de gegevens niet actueel zijn. Deze voorspelling doe je met behulp van neurale netwerken en een dataset met foto's van de huizen en numerieke data in tabelvorm. De afbeeldingen van de woningen zijn vanuit 4 verschillende perspectieven genomen (foto van slaapkamer, badkamer, keuken en vooraanzicht). Deze foto's worden vervolgens aan elkaar geplakt tot één afbeelding per huis. Je neemt met een team deel aan een Kaggle- wedstrijd waarbij je het opneemt tegen andere teams uit de opleiding Applied Data Science & AI. De team(s) die het hoogst eindigt in de wedstrijd krijgt 5 bonuspunten bovenop het aantal behaalde punten met de opdracht.

## Data

### Map Train:

Deze map bevat 500 foto's van huizen. De afbeeldingen zijn gemaakt van telkens 1 foto van de slaapkamer, de badkamer, de keuken en het vooraanzicht van het huis. De vier foto's worden vervolgens samengevoegd tot één beeld per huis.

### train.csv

Hierin vind je de metadata die bij de foto’s horen. De kolom "Price" bevat de targetvariabele. Deze kolom geeft de verkoopprijs van het huis in Amerikaanse dollars weer. De andere kolommen geven het aantal slaapkamers, het aantal badkamers, de oppervlakte (in vierkante voet) en de lengte- en breedtegraad van de locatie weer.

### Map Test:

Deze map bevat 35 afbeeldingen van huizen. Net als in de map ‘Train’ komen ook deze afbeeldingen tot stand door de foto’s van de slaapkamer, badkamer, keuken en vooraanzicht aan elkaar te plakken.

### test.csv

Hierin vind je de metadata die bij de testfoto’s horen, maar de “Price” kolom ontbreekt. Het getrainde model pas je toe op de foto’s in map Test en 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.

### submission\_example.csv

De eerste kolom bevat de huis id’s (‘House ID’) van alle huizen in de map Test en test.csv. Nadat je je voorspellingen hebt gedaan plaats je deze in de kolom ‘Price’ van dit bestand. Vervolgens kun je dit bestand uploaden op Kaggle.

## 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 (5 punten)

* + Ieder teamlid maakt een account op Kaggle aan en vormt een team dat deelneemt aan deze competitie: [Verkoopprijzen van huizen voorspellen | Kaggle](https://www.kaggle.com/competitions/verkoopprijzen-van-huizen-voorspellen).
  + Noteer je teamnaam, jullie namen en alle bijhorende Kaggle gebruikersnamen in het notebook.
  + Voer een Exploratieve Data Analyse (EDA) uit:
    - Lees de afbeeldingen in en toon enkele afbeeldingen.
    - Laad het metadatabestand en maak relevante visualisaties om de relaties tussen de targetvariabele en de andere variabelen te onderzoeken.
  + Beschrijf wat de belangrijkste bevindingen zijn van de EDA.

## Opdracht 2: Fully-connected neuraal netwerk (20 punten)

* + In deze deelopdracht werk je uitsluitend met de tabelgegevens. Bouw en train een dense (fully-connected) netwerk om de huizenprijzen te voorspellen op basis van de tabular data.
  + Plot je modelarchitectuur met behulp van een geschikte visualisatietool.
  + Onderbouw je keuzes van het aantal lagen, het aantal neuronen per laag, de activeringsfuncties, de keuze van optimizer en het aantal trainingsepochs.
  + Beschrijf welke lossfunctie het meest geschikt is voor dit probleem. Kies een geschikte optimizer en beschrijf kort hoe deze optimizer werkt. Compile je model met de gekozen lossfunctie en optimizer, en train deze op de tabelgegevens.

## Opdracht 3: Convolutioneel neuraal netwerk from scratch (20 punten)

* + In deze deelopdracht werk je alleen met de afbeeldingen en labels, en niet met de rest van de tabeldata. Bouw een convolutioneel neuraal netwerkmodel om de prijzen te voorspellen op basis van de afbeeldingen.
  + Plot je modelarchitectuur met behulp van een geschikte visualisatietool.
  + Onderbouw je keuzes van het aantal convolutionele lagen, poollagen, convolutionele kernels per laag, het aantal fully-connected lagen en het aantal neuronen per laag.
  + Gebruik **een of meer** van de volgende technieken in je model. Beschrijf en motiveer je keuze:
    - L1 regularisatie
    - L2 regularisatie
    - Dropout
    - Batchnorm
  + Beschrijf welke vormen van data augmentation (indien van toepassing ) geschikt zijn voor dit probleem. Implementeer dit in je datapijplijn.
  + Compile je model met je gekozen lossfunctie en optimizer, en train deze op de afbeeldingen.

## Opdracht 4: Transfer learning (20 punten)

* + In deze deelopdracht werk je alleen met de afbeeldingen en labels, en niet met de rest van de tabeldata. Gebruik een vooraf getraind computervisiemodel en fine-tune dit op de beeldgegevens.
  + Plot je modelarchitectuur met behulp van een geschikte visualisatietool.
  + Geef de naam op van de modelarchitectuur die je hebt gekozen en de dataset waarop deze vooraf is getraind.
  + Noem de voordelen van het gebruik van vooraf getrainde visionmodellen en leg ze uit.
  + Beschrijf in detail de stappen die worden genomen bij het finetunen van een vooraf getraind model. Voeg details toe over welke lagen je traint en welke niet, hoeveel nieuwe lagen je hebt toegevoegd (indien aanwezig) en de details over deze nieuw toegevoegde lagen.
  + Je bent vrij om hier aanvullende technieken voor regularisatie, data augmentation,

finetuning van hyperparameters, enz. te gebruiken om je classificatiescore te verbeteren. Motiveer je keuzes.

## Opdracht 5: Multimodaal model (20 punten)

* + In deze deelopdracht ga je aan de slag met zowel de afbeeldingen als de bijbehorende metadata. Creëer een multimodaal model dat zowel 2D-beeldgegevens als 1D-tabelgegevens als inputs kan in parallel accepteren.
  + Plot je modelarchitectuur met behulp van een geschikte visualisatietool.
  + Noem het aantal convolutionele en pooling lagen in de ene tak, het aantal dense lagen in de andere tak, de strategie die wordt gebruikt om de takken te combineren,

en het aantal dense lagen dat volgt. Motiveer je keuzes.

* + Je bent vrij om transfer learning in te zetten voor de CNN branche. Indien je daarvoor kiest, beschrijf en motiveer je keuze.
  + Je bent vrij om hier aanvullende technieken voor regularisatie, data augmentation, finetuning van hyperparameters, enz. te gebruiken om je voorspellingsscore te verbeteren. Motiveer je keuzes.

## Opdracht 6: 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 7: Conclusie en aanbevelingen (10 punten)

* + Beschrijf het modelleringsproces voor deze opdracht. Wat waren de uitdagingen daarmee?
  + In deze opdracht heb je verschillende modelarchitecturen uitgeprobeerd, en verschillende technieken om de prestaties te verbeteren. Welke van deze werkte het beste? Welke van deze werkte niet goed? Waarom denk je dat dat het geval was?
  + Welk advies kun je op basis hiervan geven aan een makelaar?
  + Welk advies kun je op basis hiervan geven aan potentiële huizenkopers of huizenverkopers?

## 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/7967e](https://chat.openai.com/share/7967e049-f570-4498-b425-a1694ec14178)) [049-f570-4498-b425-a1694ec14178)](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-](https://chat.openai.com/share/7967e049-f570-4498-b425-a1694ec14178) [b425-a1694ec14178](https://chat.openai.com/share/7967e049-f570-4498-b425-a1694ec14178).