# Opdracht 3: Bepalen van muziekgenres

Inhoud

[Opdracht 3: Bepalen van muziekgenres 1](#_Toc151388180)

[1. Inleiding 1](#_Toc151388181)

[2. Data 1](#_Toc151388182)

[3. Opdrachtbeschrijving 1](#_Toc151388183)

[Vereisten notebook 1](#_Toc151388184)

[Opdracht 1: Feature engineering (35 punten) 2](#_Toc151388185)

[Opdracht 2: Unsupervised Learning (55 punten) 2](#_Toc151388186)

[Opdracht 3: Bevindingen en conclusie (5 punten) 2](#_Toc151388187)

[Bijlage 1: Bronvermelding en regels gebruik Generative AI (GAI) 4](#_Toc151388188)

## 1. Inleiding

In deze opdracht gebruik je unsupervised learning om te bepalen tot welk muziekgenre een geluidsfragment behoort. We hebben 105 geluidsfragmenten verzameld in de dataset “unlabeled”. Van deze geluidsfragementen is het genre bij ons bekend. In deze dataset vind je NIET alle genres terug, maar alleen een subset. Het is jullie taak om te bepalen hoeveel genres er zijn in deze dataset en om welke genres het gaat.

Daarnaast vind je 50 geluidsfragementen waarvan het genre bekend is. Deze dataset gebruik je om te bepalen welke genres de “unlabeled” geluidsfragmenten hebben.

## 2. Data

* **labeled**: hier vind je de bestanden waarvan het genre bekend is.
* **Labels\_new.csv**: hier vind je de genres van de bestanden in "labeled".
* **unlabeled**: hier vind je de bestanden waarvan de genres niet bekend zijn.
* **sample\_submission.csv**: een voorbeeldbestand voor het uploaden van je oplossingen. Het bestand bestaat uit twee kolommen:
  + filename: de naam van het geluidsframent uit de map ‘unlabeled’
  + genre: De naam van het muziekgenre. Zorg ervoor dat je precies de de genrenamen aanhoudt uit deze list: genres = ['blues', 'classical', 'country', 'disco', 'hiphop', 'jazz', 'metal', 'pop', 'reggae', 'rock']

## 3. 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: Feature engineering (35 punten)

* Ieder teamlid maakt een account op Kaggle aan en vormt een team dat deelneemt aan deze competitie: <https://www.kaggle.com/competitions/muziek-genre-clustering>
* Noteer je teamnaam, jullie namen en alle bijhorende Kaggle gebruikersnamen in het notebook.
* Maak een dataframe aan waarin je per geluidsbestand “geluidsfeatures” toevoegt. Geluidsfeatures zijn bijvoorbeeld:
  + De spectrale bandbreedte (spectral bandwidth)
  + Het spectrale zwaartepunt (spectral centroid)
  + Etc

We verwachten dat je minimaal 8 geluidsfeatures toevoegt aan dit dataframe. Waarschijnlijk zijn er meer nodig om opdracht 2 goed te kunnen uitvoeren.

* Leg in eigen woorden uit wat de betekenis is van iedere feature is en leg uit hoe de waarde van iedere feature wordt berekend / bepaald.

### Opdracht 2: Unsupervised Learning (55 punten)

* Na het toevoegen van voldoende features bepaal je hoeveel clusters er in de “unlabeled” dataset zijn. Leg hierbij duidelijk uit hoe je tot dit aantal bent gekomen.
* Voer een geschikt clustering algoritme uit en ken clusternummers toe aan ieder geluidsfragment uit de “unlabeled” dataset.
* Leg de (wiskundige) werking van het algoritme uit. Licht dit toe met een rekenvoorbeeld op basis van een kleine (fictieve) dataset.
* Vergelijk de kenmerken van de clusters met de kenmerken van de “gelabelde” dataset. Gebruik hiervoor onder andere geschikte visualisaties. Bepaal op basis hiervan welk genre bij ieder cluster hoort.
* Vervang de clusternummers door de genres die uit jullie analyse naar voren komen. Dit vormt de basis voor het bestand dat je vervolgens “submit” op Kaggle.
* Voer twee vormen van “dimensionality reduction” uit: PCA en NMF.
  + Leg de (wiskundige) werking van beide vormen uit. Toon hierbij een vereenvoudigd rekenvoorbeeld. Beschrijf de verschillen en overeenkomsten.
  + Laat per vorm zien hoe je het aantal dimensies hebt bepaald. Leg uit hoe dit werkt.
* Beide vormen genereren nieuwe features die de oorspronkelijk features vervangen. Voer met de nieuwe features opnieuw hetzelfde clustering algoritme uit. Onderzoek in hoeverre de clustering verbetert.
* Maak een eenvoudige “app” waarmee je voor ieder geluidsfragment een suggestie doet voor andere fragmenten die een luisteraar “ook leuk zou kunnen vinden”.

### Opdracht 3: Bevindingen en conclusie (5 punten)

* Geef een samenvatting van je bevindingen:
  + Welke features lijken het belangrijkst voor een goede clustering.
  + Beschrijf het effect en nut van “dimensionality reduction”.
  + Beschrijf welke gegevens je nog meer zou willen hebben om nog betere aanbevelingen te kunnen doen?

## 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>.