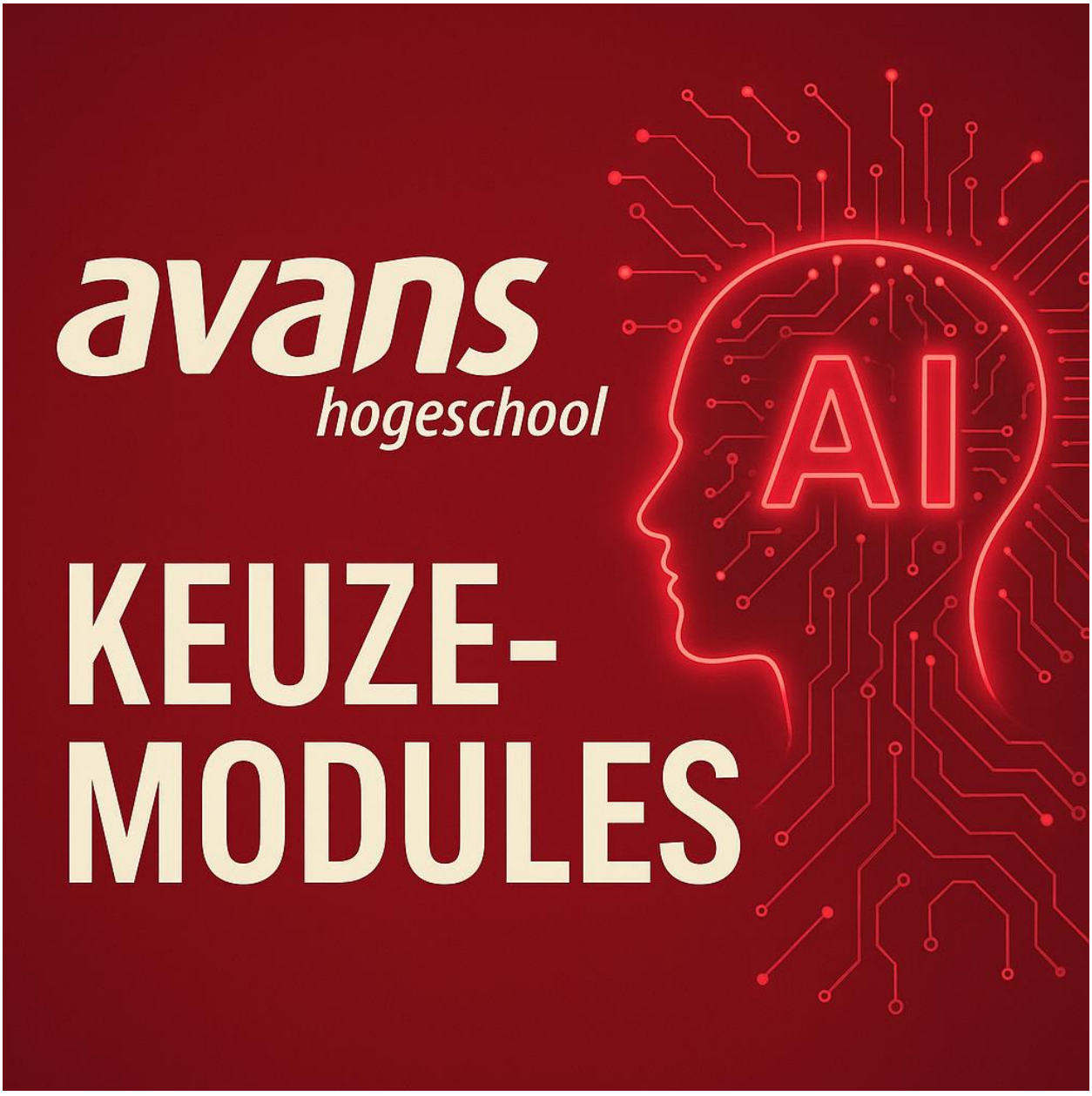
AI-oplossing voor Vrije Keuzemodules (VKM)



Boyan Kloosterman
Koen Smit
Groep 15
Informatica DSLC
23-11-2025

1. Inleiding

De digitale leeromgeving binnen het hoger onderwijs ontwikkelt zich snel. Studenten krijgen steeds vaker te maken met moeilijke keuzetrajecten, grote hoeveelheden informatie en digitale hulpmiddelen die hen moeten ondersteunen in het maken van studiekeuzes. Binnen Avans Hogeschool speelt deze ontwikkeling een centrale rol. De organisatie stimuleert de inzet van data-gedreven toepassingen en artificiële intelligentie om studenten beter te begeleiden, mits deze toepassingen verantwoord en transparant worden ontworpen. Het AI Prototype-project sluit direct op die onderwijsvisie aan. Het doel is om een datagedreven aanbevelingssysteem te ontwikkelen dat studenten helpt bij het verkennen van passende studiemogelijkheden, terwijl ethische en juridische kaders, zoals de EU AI Act 2025 en de AVG strikt worden meegenomen.

Doel van de opdracht

Het hoofddoel van deze opdracht is het ontwerpen, bouwen en evalueren van een complete AI-pipeline (min de deployment) die een realistisch onderwijsprobleem ondersteunt: het doen van gepersonaliseerde aanbevelingen op basis van studentdata. Om dit te bereiken doorloopt het rapport de volledige datastructuur- en modelcyclus: van **business understanding**, naar **data-analyse en voorbereiding**, **modelontwikkeling**, **optimalisatie**, en uiteindelijk een **kritische evaluatie** van de resultaten.

De inhoudelijke basis komt uit meerdere kerndomeinen:
– **Data-analyse en datakwaliteit**, zoals behandeld in “Data Analysis with Python”, benadrukt het belang van een sterke datacleaning-fase om betrouwbare modellen te kunnen trainen.
– **Modelbouw en evaluatie**, waaronder lineaire regressie, clustering, evaluatiemetrics en het opzetten van baselines. Deze concepten vormen de fundering van de modellering in deze opdracht.
– **Dimensionality reduction** zoals PCA, t-SNE en UMAP, speelt een rol bij feature-selectie, modellering en visualisatie. Deze technieken helpen om ruis te verwijderen en patronen inzichtelijk te maken.
– **Recommender systems**, content-based vormt de basis voor het type model dat in deze opdracht wordt ontwikkeld.
– **NLP-technieken** zoals tokenisatie, stopword-verwerking en vectorisatie zijn relevant voor het verwerken van tekstuele informatie, bijvoorbeeld module- of cursusbeschrijvingen.

– **Ethiek en algorithmic affordances**, zoals behandeld in de colleges Ethiek en AA, zijn essentieel om de verantwoordelijkheid, transparantie, autonomie en uitlegbaarheid van het AI-product te waarborgen.

Door deze verschillende kennisgebieden te combineren, levert het project een AI-prototype op dat technisch onderbouwd is, instructief relevant is en voldoet aan eisen voor verantwoord AI-ontwerp.

Context binnen Avans

Het project wordt uitgevoerd binnen Avans waar wij studenten leren om AI-toepassingen te ontwerpen vanuit zowel datatechnisch als maatschappelijk perspectief. De nadruk ligt op het kritisch onderzoeken van data, het verantwoord gebruiken van algoritmes en het afwegen van privacy en ethische impact. De rubric van het vak onderstreept dat studenten niet alleen een werkend model moeten opleveren, maar ook een professioneel gefundeerde onderbouwing moeten geven voor hun analyses en ontwerpkeuzes.

Binnen Avans wordt bovendien veel belang gehecht aan transparante en uitlegbare AI. In het gastcollege over ethiek wordt bijvoorbeeld benadrukt dat autonomie, biaspreventie en transparantie geen optionele elementen zijn, maar voorwaarden voor verantwoord ontwerp. Eveneens worden in het college Algorithmic Affordances ontwerpprincipes besproken die studenten in staat stellen de gebruiker meer controle te geven over AI-gedrag, zoals het manipuleren van input, parameters of uitkomsten.

Wat het rapport oplevert

Dit rapport vormt het volledige technisch-inhoudelijke en ethische fundament van het AI-prototype. Concreet levert het de volgende elementen op:

1. **Een helder geformuleerde probleemstelling**, met maatschappelijke en onderwijsrelevantie en onderbouwd vanuit ethiek en privacywetgeving.
2. **Een uitgebreide exploratie van de dataset**, inclusief datacleaning, detectie van inconsistenties, EDA-visualisaties en voorbereiding voor modellering.
3. **Een volledig uitgewerkte modelbouw-fase**, waarin recommender-technieken, vectorisatie, regressie, clustering of PCA worden toegepast afhankelijk van de gekozen AI-aanpak.
4. **Evaluatie volgens professionele standaarden**, waarbij metrics zoals cosine similarity, of silhouette-score worden gebruikt.

5. **Optimalisatie en performance-tuning**, waaronder herziening van features, hyperparameters en dataverwerking.
6. **Een integratie van ethische reflectie**, waarbij bias, transparantie, autonomie en privacy worden beoordeeld in relatie tot het prototype.
7. **Een onderbouwd advies**, gericht op bruikbaarheid, effectiviteit en verantwoord functioneren binnen de Avans-context.

Samen levert dit een compleet en professioneel adviesrapport op dat voldoet aan de beoordelingscriteria van het assessment en direct toepasbaar is als basis voor verdere ontwikkeling van een AI-gedreven studiehulpmiddel.

2. Business Understanding

Probleemstelling: keuzehulp voor studenten

Studiekeuzes vormen een van de belangrijkste beslissingen in het hoger onderwijs. Studenten navigeren door grote hoeveelheden opties, variërend van minoren tot keuzemodules en specialisatierichtingen. Hoewel deze keuzeruimte vrijheid biedt, kan deze ook leiden tot keuzestress, gebrek aan overzicht en onzekerheid. De huidige digitale omgeving binnen Avans biedt informatie, maar mist intelligente ondersteuning die aansluit bij individuele voorkeuren, eerdere prestaties en persoonlijke doelen. Om die reden is er behoefte aan een datagedreven keuzehulp die studenten begeleidt op een transparante, eerlijke en uitlegbare manier.

Deze opdracht richt zich op het ontwikkelen van een AI-prototype dat studenten helpt bij het vinden van relevante keuzemodules, met aandacht voor privacy, ethiek en verantwoord AI-gebruik zoals geëist in de EU AI Act en het onderwijsbeleid binnen Avans.

Doelgroep en belang

De primaire doelgroep bestaat uit studenten binnen het hoger beroepsonderwijs die keuzes moeten maken rondom studieonderdelen, zoals keuzemodules, minoren of verdiepingsvakken. Voor deze groep speelt data-ondersteuning een belangrijke rol, zeker in situaties waarin:

– studenten moeite hebben om hun interesses en motivatie te koppelen aan beschikbare modules;
– beschikbare informatie vooral descriptief is, maar weinig personalisatie biedt;
– studenten geen helder beeld hebben van welke modules aansluiten bij hun leerstijl, talent of eerdere studieresultaten;
– docenten en studieloopbaanbegeleiders afhankelijk zijn van handmatige begeleidingstrajecten.

Vanuit het perspectief van Avans sluit deze doelgroep aan bij het doel om studenten beter te ondersteunen in hun zelfstandigheid en professionele ontwikkeling. De college over ethiek benadrukt dat studenten moeten worden geholpen om zelf te kunnen bepalen, ondersteund door transparante AI-tools die hen niet sturen, maar informeren.

Het belang voor studenten is tweeledig:

1. **Praktisch belang**
   Betere ondersteuning leidt tot betere keuzes, minder uitval en een meer consistente leerroute.
2. **Persoonlijk en normatief belang**
   Zoals in de ethiekcolleges wordt besproken: autonomie vereist niet alleen keuzevrijheid, maar ook de capaciteiten, middelen en informatie om die keuze goed te kunnen maken. Een goed ontworpen AI-keuzehulp versterkt juist die voorwaarden.

Daarnaast kan een AI-aanbevelingssysteem een waardevolle bijdrage leveren aan de studieloopbaanbegeleiding, omdat het in staat is om patronen te herkennen in studentdata en deze om te zetten in bruikbare aanbevelingen, vergelijkbaar met recommender-technieken zoals content-based filtering.

Maatschappelijke relevantie

De maatschappelijke relevantie van dit project ligt in meerdere domeinen:

Toegankelijkheid van onderwijskeuzes

In de huidige onderwijscontext neemt de hoeveelheid beschikbare keuzes toe. Studenten moeten door honderden modules, specialisaties en opdrachten navigeren. Een AI-systeem dat helpt bij het structureren van informatie kan bijdragen aan gelijke kansen en toegankelijkheid. Dit sluit aan bij het kernthema “AI4good” uit het ethiekcollege, waarin AI wordt ingezet om mensen te ondersteunen in plaats van te sturen.

Verantwoord gebruik van AI

Het hoger onderwijs heeft de opdracht om AI-toepassingen te ontwerpen volgens de principes van transparantie, uitlegbaarheid, fairness en privacybescherming. Dit wordt expliciet genoemd in het rubric-onderdeel “Business Understanding & Data Collection”, waarin studenten worden beoordeeld op hun vermogen om ethiek, privacy en de EU AI Act te betrekken in hun ontwerp.

Een keuzehulp die uitlegt waarom aanbevelingen worden gedaan, die gebruikers controle geeft (bijv. via algorithmic affordances zoals filters, parameters of wegingen), draagt direct bij aan een bewustere implementatie van AI in het onderwijs.

Voorkomen van bias en discriminatie

Het ethiekcollege benoemt risico's zoals bias in data, ongelijke behandeling en slechte datakwaliteit. Een keuzehulp die werkt op basis van transparante parameters en gecontroleerde input kan deze risico's verkleinen.
Daarmee draagt het systeem niet alleen bij aan persoonlijke studieontwikkeling, maar ook aan maatschappelijke waarden zoals gelijke behandeling.

Onderwijsefficiëntie

Docenten en begeleiders besteden veel tijd aan individuele adviesgesprekken. Een AI-tool functioneert als eerste laag van ondersteuning, waardoor menselijk advies inzetbaar blijft voor moeilijkere gevallen.

Afbakening van de oplossing

De scope van dit project wordt bewust afgebakend, om binnen de beperkte tijd en dataset een haalbaar prototype te ontwikkelen.

Type aanbevelingssysteem

De oplossing richt zich primair op *content-based filtering*, afhankelijk van de beschikbare data. De focus ligt op vectorisatie van modulebeschrijvingen of prestatieprofielen, zoals uitgelegd in de lessen over recommender systems.

Dataset en datakwaliteit

De oplossing werkt met de beschikbare VKM-dataset en aanvullende data die binnen de opdrachtcontext is toegestaan. Het project beperkt zich tot de verwerking van:
– module-informatie,
– studentprofielen,
– historische resultaten of voorkeuren (indien aanwezig).

De datakwaliteitsprincipes (opschoning, outlier-detectie, consistentiecontrole) zoals behandeld in "Data Analysis with Python" worden toegepast, maar de opdracht omvat geen grootschalige data-engineering pipeline.

Modellering

Het prototype gebruikt één duidelijk gekozen modelaanpak.
Dat kan bijvoorbeeld zijn:
– cosine similarity met CountVectorizer (zoals in het recommender-college)
– PCA voor dataset-reductie wanneer nodig (dimensionality-reductioncollege)
– clustering of regressie alleen als het past binnen de casus en dataset.

Geavanceerde deep-learningmethoden vallen buiten de scope.

Ethiek en algorithmic affordances

Hoewel het systeem rekening houdt met autonomie en transparantie, wordt geen volledige 'explainable AI-omgeving' gebouwd. De focus ligt op:
– inzicht geven in inputs,
– begrijpelijke weergave van aanbevolen modules,
– optionele invloed op parameters (bijv. "vind modules meer op basis van vakinhoud").

Geen productieomgeving

Het project levert een prototype, geen productieklare applicatie.
Er wordt geen grootschalige backend, gebruikersbeheer of privacy-framework geïmplementeerd buiten wat noodzakelijk is voor demonstratie.

3. Ethiek, Privacy & EU AI Act 2025

AI-technologieën die studenten helpen bij hun studiekeuzes hebben een directe invloed op hun persoonlijke groei, zelfstandigheid en gelijke behandeling. Het gastcollege Ethiek legt de nadruk op het feit dat technologie nooit neutraal is: elke keuze voor een ontwerp heeft invloed op het gedrag, de verwachtingen en de mogelijkheden van gebruikers. Een keuzehulp dient daarom te voldoen aan ethische normen, transparantie-eisen en wettelijke richtlijnen zoals de AVG en de EU AI Act 2025.

Autonomie van de student

Autonomie houdt in dat studenten in staat zijn hun eigen studiepad te creëren, gebaseerd op juiste informatie, zelfreflectie en rechtvaardige hulp. Het gastcollege verduidelijkt dat autonomie zowel interne vaardigheden (het zelfstandig kunnen maken van keuzes) als externe omstandigheden (informatie, middelen en respect voor zelfbeschikking) omvat.

Het ontwerp van de keuzehulp vergroot de zelfstandigheid door: aanbevelingen te presenteren als mogelijkheden, niet als richtlijnen; inzicht te geven in de redenen achter een aanbeveling; studenten de mogelijkheid te geven om hun voorkeuren zelf aan te passen (bijvoorbeeld door wegingen, filters of prioriteiten te gebruiken via algorithmische mogelijkheden).

Die interacties maken het mogelijk dat de student niet door het algoritme wordt geleid, maar dat het algoritme zelf invloed uitoefent. Dit zorgt ervoor dat de AI de functie van studieloopbaanbegeleider niet overneemt en houdt de student in het besluitvormingsproces centraal.

Transparantie & uitlegbaarheid

Transparantie is essentieel voor vertrouwen en verantwoord gebruik van AI. In het ethiekcollege wordt benadrukt dat transparantie verschillende lagen heeft: inzicht in de werking, in de gebruikte data en in de waarden die aan het systeem ten grondslag liggen.

De keuzehulp biedt die uitleg gericht op drie dimensies:

- **Uitkomsttransparantie:** waarom een specifieke module wordt aanbevolen.
- **Procestransparantie:** hoe het model informatie verwerkt.

- **Waarde-transparantie:** welke ontwerpkeuzes zijn gemaakt en waarom, inclusief de beperkingen van het model.

Door studenten duidelijk te maken welke kenmerken de aanbevelingen beïnvloeden, krijgen zij beter inzicht in zowel hun eigen voorkeuren als de rol van de AI binnen hun studieproces.

Bias & discriminatie-risico's

Bias ontstaat vaak doordat data historische ongelijkheden weerspiegelt. Het gastcollege noemt oorzaken zoals "bias in - bias out", verkeerde dataverzameling, proxy-variabelen en ondervertegenwoordiging van groepen.

Voor dit project kunnen risico's ontstaan bij:

- modules die onvolledig of inconsistent zijn beschreven, waardoor NLP-modellen sommige modules bevoordelen;
- beperkte interactiedata, wat kan leiden tot slechtere aanbevelingen bij studenten met minder studiehistorie.

Mitigatie richt zich op zorgvuldige dataselectie, standaardisatie waar nodig, en het vermijden van kenmerken die proxy-bias kunnen veroorzaken. Bovendien blijft het systeem adviserend: aanbevelingen zijn geen kwalitatief oordeel over de student.

AVG & dataminimalisatie

De AVG bepaalt dat persoonsgegevens alleen mogen worden verwerkt wanneer dit strikt noodzakelijk is. Dataminimalisatie, doelbinding en transparantie zijn daarom leidende principes bij het ontwerp.

Dit betekent concreet:

- er wordt uitsluitend data gebruikt die relevant is voor het genereren van aanbevelingen;
- student-ID's worden geanonimiseerd;
- er worden geen gegevens verwerkt die niet bijdragen aan het doel (bijv. nationaliteit, adresgegevens, achtergrondinformatie);
- studenten worden geïnformeerd over welke gegevens worden gebruikt en waarom;

- data wordt niet buiten de onderwijscontext gedeeld en na afloop van het project verwijderd.

Zo blijft de keuzehulp in lijn met de wettelijke privacyvereisten én met het ethische uitgangspunt dat studenten controle houden over hun informatie.

Risicoanalyse volgens de EU AI Act 2025

De EU AI Act introduceert risicocategorieën waarmee AI-systemen worden beoordeeld. Een studiekeuze-aanbeveler valt niet in de hoogste risicoklasse, maar wordt ook niet als risicoloos beschouwd. Omdat de toepassing invloed heeft op persoonlijke keuzes en studieverloop, is sprake van een **“limited to moderate risk”**.

De belangrijkste risico’s zijn:

- **Verlies van autonomie:** wanneer aanbevelingen te dwingend worden gepresenteerd.
- **Ongelijke behandeling:** door bias in data of verkeerde aannames in het model.
- **Onvoldoende transparantie:** studenten begrijpen niet hoe aanbevelingen ontstaan.
- **Privacyrisico’s:** door onnodige of verkeerde verwerking van persoonsgegevens.
- **Afhankelijkheid:** studenten kunnen té veel vertrouwen op de tool en minder kritisch gaan nadenken.

Volgens de AI Act moeten deze risico’s vooraf worden geïdentificeerd en moeten passende maatregelen worden genomen om ze te beperken.

Mitigatie & verantwoord ontwerp

Mitigatie richt zich op het combineren van technische maatregelen, ontwerpprincipes en duidelijke communicatie. Het prototype integreert deze aanpak in alle fasen van de AI-pipeline.

Belangrijke maatregelen zijn:

- **Autonomie-borging:** aanbevelingen blijven optioneel, aanpasbaar en transparant in hun ontwikkeling.
- **Transparantie-instrumenten:** uitlegmodules, inzicht in factoren en visuele ondersteuning bij aanbevelingen.

- **Biasreductie:** controleren op ongelijke patronen, vermijden van proxy-variabelen en eerlijk omgaan met ontbrekende of scheve data.
- **Privacybescherming:** pseudonimisering, strikt dataminimale dataverwerking en beperkte toegang.
- **Documentatie volgens EU AI Act:** duidelijke onderbouwing van gebruikte technieken (zoals CountVectorizer, cosine similarity, PCA), inzicht in risico's en rapportage van ontwerpkeuzes.

Door deze maatregelen ontstaat een systeem dat technisch betrouwbaar en maatschappelijk verantwoord is. De tool ondersteunt studenten in hun ontwikkeling, zonder hun autonomie, privacy of gelijke behandeling in gevaar te brengen.

4. Dataset & Data Collection

De VKM-dataset vormt de kern van dit project. Deze dataset bevat een groot overzicht van keuzemodules binnen Avans, inclusief inhoudelijke beschrijvingen, metadata en meerdere numerieke kenmerken. De dataset is met 211 rijen relatief compact, maar rijk genoeg om een realistisch aanbevelingssysteem te trainen. De variatie in tekst, numerieke indicatoren en modulekenmerken maakt de dataset geschikt voor content-based filtering, NLP-analyse en aanvullende experimenten zoals dimensionality reduction.

Beschrijving van de uitgebreide VKM-dataset

De dataset bestaat uit twintig kolommen waarin modules worden beschreven vanuit verschillende perspectieven: inhoud, leerdoelen, populariteit, moeilijkheid en beschikbaarheid. De structuur laat zien dat belangrijke velden zoals *name*, *description*, *content* en *learningoutcomes* volledig aanwezig zijn, waardoor er voldoende informatie is om modules met elkaar te vergelijken.

Hoewel de dataset gedeeltelijk is aangevuld met aanvullende scoringsvelden (zoals *popularity_score* en *estimated_difficulty*), is ook zichtbaar dat enkele velden nauwelijks zijn ingevuld. Dit wijst erop dat de datasets enkele fouten bevat. Dat maakt data-opschoning een belangrijk onderdeel van de verdere modellering.

Databronnen, kolommen en datatypes

De dataset combineert tekstuele, numerieke en categorische informatie. De meest relevante kolomtypes zijn:

Identificatie en metadata

- *id* (integer): uniek module-ID.
- *name* (string): titel van de module.
- *location* (string): vestiging of onderwijslocatie.
- *level* (string): niveau of leerjaar-indicatie.

Tekstvelden (NLP-relevant)

- *shortdescription*

- *description*
- *content*
- *learningoutcomes*

Deze velden bevatten de inhoudelijke basis van de module en vormen het fundament van elke content-based analyse.

Numerieke velden

- *studycredit* (ECTS)
- *interests_match_score*
- *popularity_score*
- *estimated_difficulty*
- *available_spots*

Datumveld

- *start_date*

Daarnaast bevat de dataset enkele kolommen zoals *Rood*, *Blauw*, *Zwart* en *Bron*, die grotendeels leeg zijn. Deze worden verder niet gebruikt.

Relevante velden voor aanbevelingen

Voor een goed werkend aanbevelingssysteem zijn niet alle velden gelijkwaardig. Een selectie van de meest belangrijke categorieën is noodzakelijk.

1. Inhoudsvelden voor modulevergelijking
Deze vormen de basis voor het omzetten van tekst naar getallen en het bepalen van inhoudelijke overeenkomsten:

- *name*
- *description*
- *content*
- *learningoutcomes*

Hieruit kunnen betekenisvolle patronen worden gehaald via CountVectorizer, TF-IDF of embeddings (Wij gaan TF-IDF gebruiken).

2. Relevantiemetingen en gedragskenmerken
Deze velden kunnen de rangschikking of verfijning van aanbevelingen ondersteunen:

- *interests_match_score*
- *popularity_score*

Hoewel zinvol, moeten deze waarden kritisch worden bekeken, omdat hun herkomst en betrouwbaarheid niet volledig zijn.

3. Praktische moduleinformatie
Handig voor filters of aanvullende voorkeuren:

- *studycredit*
- *available_spots*
- *start_date*

Deze velden beïnvloeden niet de inhoudelijke overeenkomst, maar kunnen wel bijdragen aan bruikbare en realistische aanbevelingen.

Datakwaliteit & beperkingen

De dataset is over het algemeen van goede kwaliteit, maar kent enkele beperkingen die in het verdere dataverwerkingsproces aandacht vereisen.

Variatie in tekstkwaliteit
Sommige modules hebben lange beschrijvingen, andere maar een paar zinnen. Hierdoor kan de omzetting naar numerieke representaties scheef worden, omdat langere teksten meer woorden en dus meer kenmerken bevatten. Preprocessing (lemmatisatie, stopwoorden verwijderen) helpt dit gedeeltelijk op te vangen.

Ontbrekende waarden
De kolom *shortdescription* bevat meerdere ontbrekende waarden. *learningoutcomes* mist slechts bij een klein aantal rijen. Omdat andere inhoudsvelden wel aanwezig zijn, vormt dit geen groot probleem, maar moet het wel worden meegenomen in de datavoorbereiding.

Lege of irrelevante kolommen
De velden *Rood*, *Blauw*, *Geel*, *Groen*, *Zwart* en *Bron* hebben vrijwel uitsluitend NaN-waarden. Deze worden uitgesloten omdat ze geen waarde toevoegen aan de aanbevelingslogica.

Redundantie tussen tekstvelden
De velden *description*, *content* en *learningoutcomes* overlappen deels. Daardoor ontstaat er sneller herhaling in de numerieke representaties. Technieken zoals PCA kunnen later helpen om deze overlappende informatie te verminderen of inzichtelijk te maken.

Onbekende herkomst van scoringsvelden
De betekenis van *interests_match_score*, *popularity_score* en *estimated_difficulty* is niet volledig gedocumenteerd. Ze kunnen worden gebruikt, maar alleen als secundaire of optionele kenmerken om bias te voorkomen.

5. Data Opschonen

Voor een betrouwbaar aanbevelingssysteem is een dataset nodig die volledig, consistent en geschikt is voor NLP-verwerking. De oorspronkelijke VKM-dataset bevatte lege velden, ongedefinieerde waarden zoals "NTB", inconsistent gebruik van tags en tekstformaten die niet direct bruikbaar waren voor vectorisatie. Daarom is een zorgvuldig opschoningsproces uitgevoerd waarin zowel inhoudelijke als technische correcties zijn aangebracht.

Selectie van relevante gegevens

Bij de eerste inspectie werd duidelijk dat sommige kolommen geen betekenisvolle bijdrage leverden aan het aanbevelingsmodel. Vooral de kleurkolommen waren vrijwel volledig leeg en hadden geen inhoudelijke functie. Deze zijn verwijderd om de dataset compacter en overzichtelijker te maken.

De dataset is teruggebracht tot uitsluitend kenmerken die relevant zijn voor module-analyse, zoals beschrijvingen, leerdoelen, numerieke scores en startdata. Dit vormt een stevig fundament voor verdere verwerking.

Ontbrekende waarden en “NTB”-teksten corrigeren

In verschillende tekstvelden kwamen zowel echte lege waarden voor als placeholders zoals “NTB”. Om deze velden bruikbaar te maken is eerst alles uniform gemaakt, waarna ontbrekende informatie systematisch is aangevuld.

De belangrijkste correcties waren:

- shortdescription is aangevuld met de eerste 200 tekens van description, of met een neutrale samenvatting als beide ontbraken.
- description en content zijn waar nodig wederzijds gebruikt om elkaar op te vullen.
- volledig ontbrekende tekstvelden kregen een automatisch gegenereerde, inhoudelijk plausibele tekst die aansluit bij de module naam.

Na deze stap bevatte de dataset geen ontbrekende tekstvelden meer.

Opschonen en aanvullen van module_tags

De kolom module_tags bleek sterk inconsistent. Sommige modules hadden geen tags, anderen hadden “NTB”, en vaak waren tags opgeslagen als onbruikbare tekststrings. Daarom is deze kolom volledig opgeschoond.

Het proces bestond uit:

- het parsen van tags naar echte Python-lijsten;
- het verwijderen van lege of betekenisloze waarden;
- het genereren van nieuwe tags wanneer een module geen bruikbare tags had.

Bij het genereren van nieuwe tags is gekeken naar sleutelwoorden in de module naam en korte beschrijving. Wanneer een module duidelijk aansloot bij bekende domeinen zoals technologie, zorg, jeugd, ontwerp of onderzoek, zijn passende tags toegevoegd. Als er geen inhoudelijke match was, zijn betekenisvolle woorden uit de module naam gebruikt. Uiteindelijk heeft iedere module nu een set relevante tags.

Leeruitkomsten aanvullen en standaardiseren

De leeruitkomsten in de ruwe data waren wisselend van kwaliteit. Veel entries waren leeg of bevatten opnieuw “NTB”. Omdat leeruitkomsten een belangrijke rol spelen in het verklaren van module-inhoud, zijn deze opnieuw opgebouwd wanneer ze ontbraken.

De werkwijze was als volgt:

- het niveau van de module (bijvoorbeeld NLQF5 of NLQF6) bepaalt de basisleeruitkomst;
- aanwezige module_tags worden gebruikt om deze leeruitkomst aan te vullen met een inhoudelijke, domeinspecifieke zin;
- bestaande leeruitkomsten die bruikbaar zijn, zijn behouden.

Zo heeft elke module nu leerdoelen die passen bij het niveau én de inhoud van de module.

Normalisatie van tekstvelden voor NLP

Om de tekstkolommen geschikt te maken voor vectorisatie is een volledige normalisatiepipeline toegepast. Hiermee worden ruis en variatie in taalgebruik verminderd, zodat vergelijkbare woorden en zinnen correct geclusterd worden in latere analyses.

De belangrijkste stappen bestonden uit:

- lowercasing van alle tekst;
- verwijderen van speciale tekens via reguliere expressies;
- tokenisatie;
- verwijderen van Nederlandse stopwoorden;
- lemmatizatie naar grondvormen;
- begrenzing van teksten tot maximaal 200 tokens.

De genormaliseerde versies zijn opgeslagen in nieuwe kolommen, zoals shortdescription_clean en content_clean. Deze vormen de directe input voor NLP-modellen.

Controle op datums, duplicaten en consistentie

Naast inhoudelijke schoonmaak is de dataset gecontroleerd op technische integriteit. De kolom start_date is volledig omgezet naar een uniform datetime-format. Bij deze conversie bleken alle waarden geldig. Ook is gecontroleerd op duplicaten in de kolom id. Geen van de modules kwam dubbel voor, waardoor de dataset structureel correct bleek.

Na afronding van alle opschoningsstappen bevatte de dataset 211 rijen, 20 kolommen en geen enkele ontbrekende waarde. Hiermee is de dataset volledig geschikt gemaakt voor verdere EDA, feature engineering en modellering.

Verantwoording van keuzes

Alle keuzes in dit opschoningsproces zijn gedreven door reproduceerbaarheid, consistentie en geschiktheid voor NLP-technieken. Irrelevante kolommen zijn verwijderd, ontbrekende tekstvelden zijn aangevuld op basis van logische bronnen, tags zijn opgeschoond of opnieuw gegenereerd, leeruitkomsten zijn gestandaardiseerd en alle tekstvelden zijn op uniforme wijze genormaliseerd. Hierdoor beschikt het model in de volgende hoofdstukken over een dataset die zowel inhoudelijk compleet als technisch betrouwbaar is.

6. Exploratory Data Analysis (EDA)

Univariate analyse

In de univariate analyse bekijken we elke variabele afzonderlijk. Het doel is inzicht krijgen in verdelingen, extreme waarden en mogelijke dataproblemen voordat we relaties tussen variabelen onderzoeken of modellen bouwen. Hieronder worden eerst de belangrijkste numerieke kenmerken besproken, daarna de meest relevante categorische variabelen. Waar het logisch is, geef ik aan waar je een figuur (histogram of staafdiagram) in je verslag kunt opnemen.

Numerieke variabelen

De kern-numerieke velden in de VKM-dataset zijn:

- *studycredit*
- *interests_match_score*
- *popularity_score*
- *estimated_difficulty*
- *available_spots*

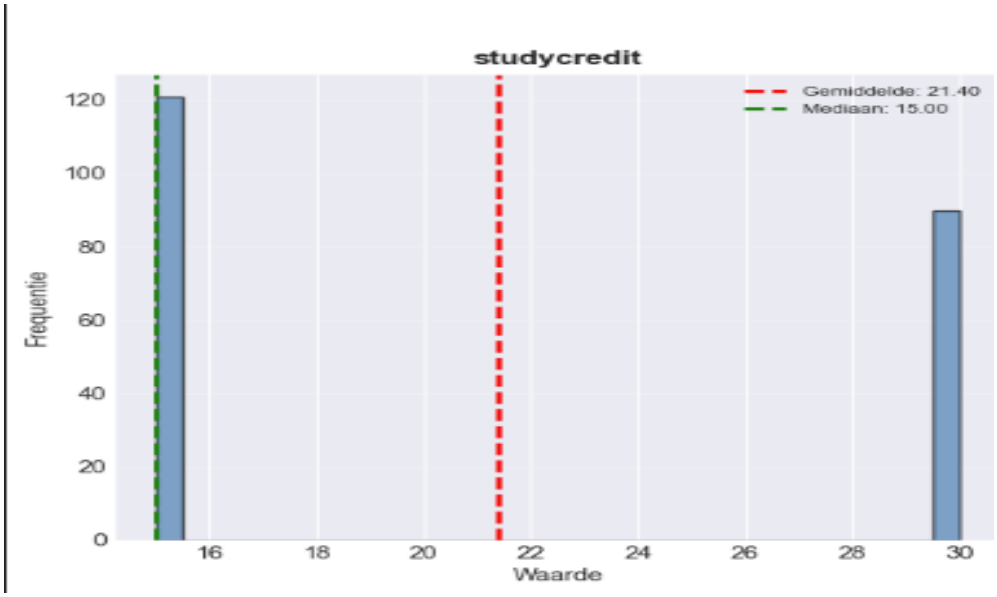
Studycredit

De variabele *studycredit* (ECTS) heeft 211 waarden, met:

- minimum: 15
- maximum: 30
- gemiddelde: circa 21,4
- mediaan: 15

De verdeling is duidelijk geconcentreerd rond twee waarden: 15 en 30 studiepunten. Dit wijst op een beperkt aantal standaardformaten (bijvoorbeeld halve en hele semesters). Er zijn geen extreme outliers.

Figuur 7.1:



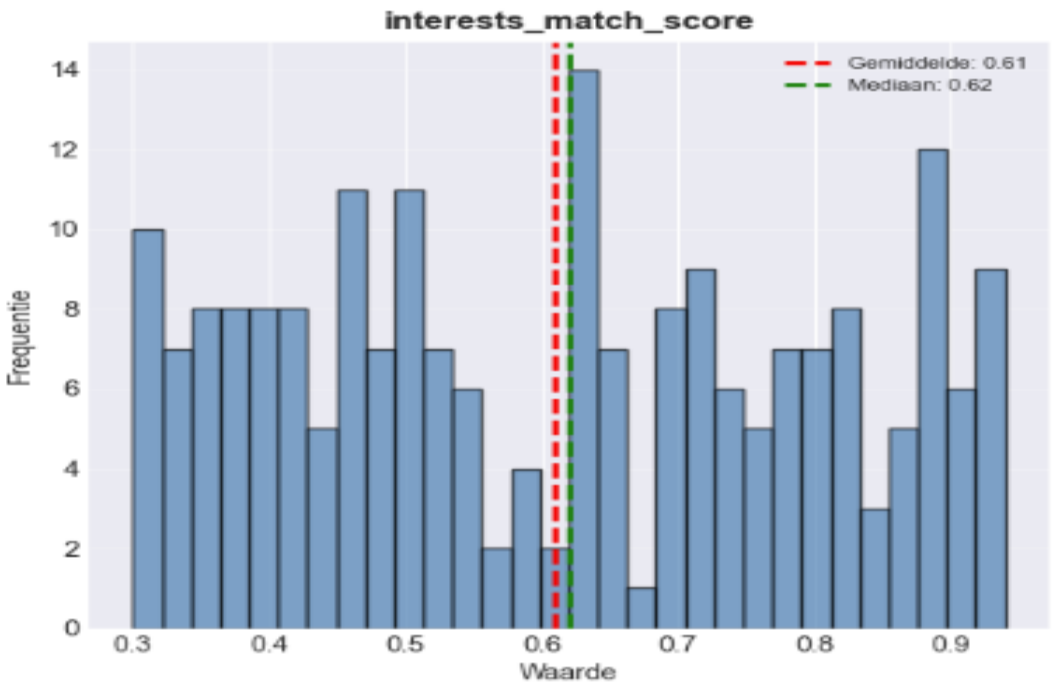
Interests_match_score

De variabele *interests_match_score* ligt tussen ongeveer 0,30 en 0,94, met:

- gemiddelde: ~0,61
- mediaan: ~0,62
- interkwartiel: 0,44 – 0,78

De verdeling laat zien dat de meeste modules een middelhoge tot hoge matchscore hebben. Zeer lage scores (< 0,4) komen relatief weinig voor. Dat betekent dat de dataset al een lichte bias heeft richting “redelijk passend” en dat deze score later eerder als verfijnende factor dan als harde filter gebruikt moet worden.

Figuur 7.2:



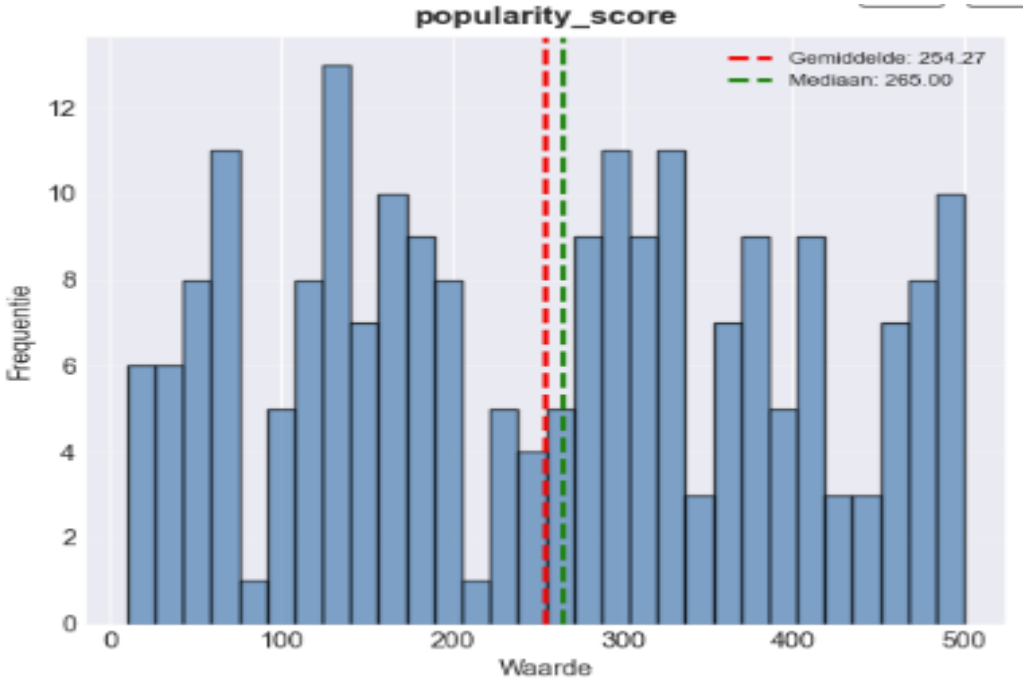
Popularity_score

popularity_score varieert van 10 tot 500:

- gemiddelde: ~254
- mediaan: 265
- Q1: ~135,5
- Q3: ~374

De spreiding is breed. Er zijn modules met lage populariteit, maar ook modules die zeer vaak gekozen worden. De mediaan ligt iets boven het gemiddelde, wat suggereert dat er een aantal modules met lage score zijn die het gemiddelde naar beneden trekken, zonder dat er echt extreme outliers zijn.

Figuur 7.3:



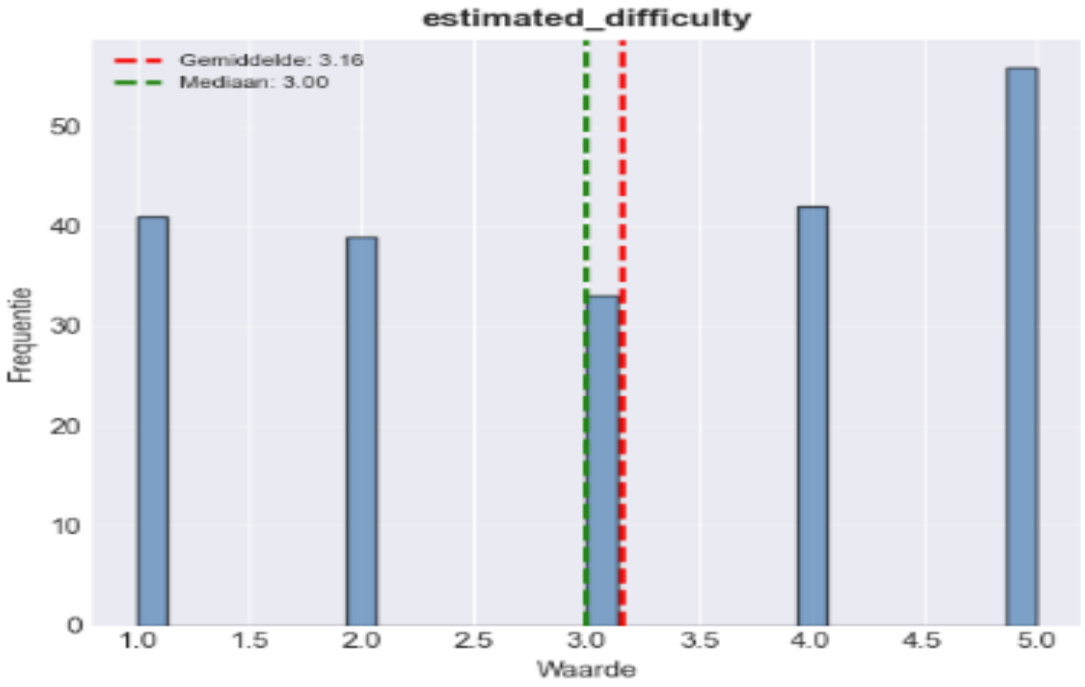
Estimated_difficulty

De variabele *estimated_difficulty* is een ordinale schaal van 1 tot 5:

- gemiddelde: ~3,16
- mediaan: 3
- Q1: 2
- Q3: 5

Er lijkt een gezonde spreiding over de schaal te zijn. Zowel laagdrempelige (niveau 1–2) als uitdagende modules (niveau 4–5) zijn aanwezig, met een lichte concentratie rond 3. Hierdoor kan *estimated_difficulty* later goed worden gebruikt als parameter in filters of sliders (bijvoorbeeld “toon meer uitdagende modules”).

Figuur 7.4:



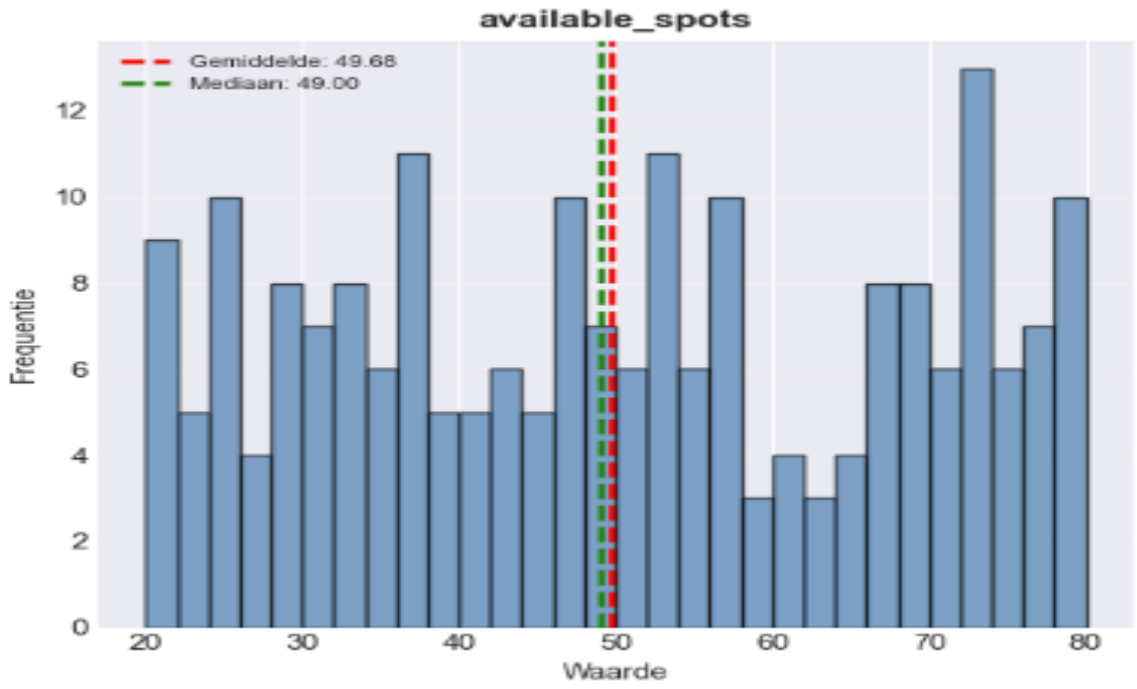
Available_spots

available_spots varieert van 20 tot 80:

- gemiddelde: ~49,7
- mediaan: 49
- Q1: 34
- Q3: 67

De meeste modules hebben een middelgroot aantal plaatsen. Er zijn geen extreem lage of extreem hoge waarden, wat erop duidt dat de dataset realistische capaciteitsschattingen bevat. Deze variabele is vooral praktisch interessant voor de UI (bijvoorbeeld waarschuwen als een module weinig plekken heeft).

Figuur 7.5:



Categorische variabelen

Voor de EDA zijn vooral *level*, *location* en *module_tags* interessant als categorische/semicategorische variabelen.

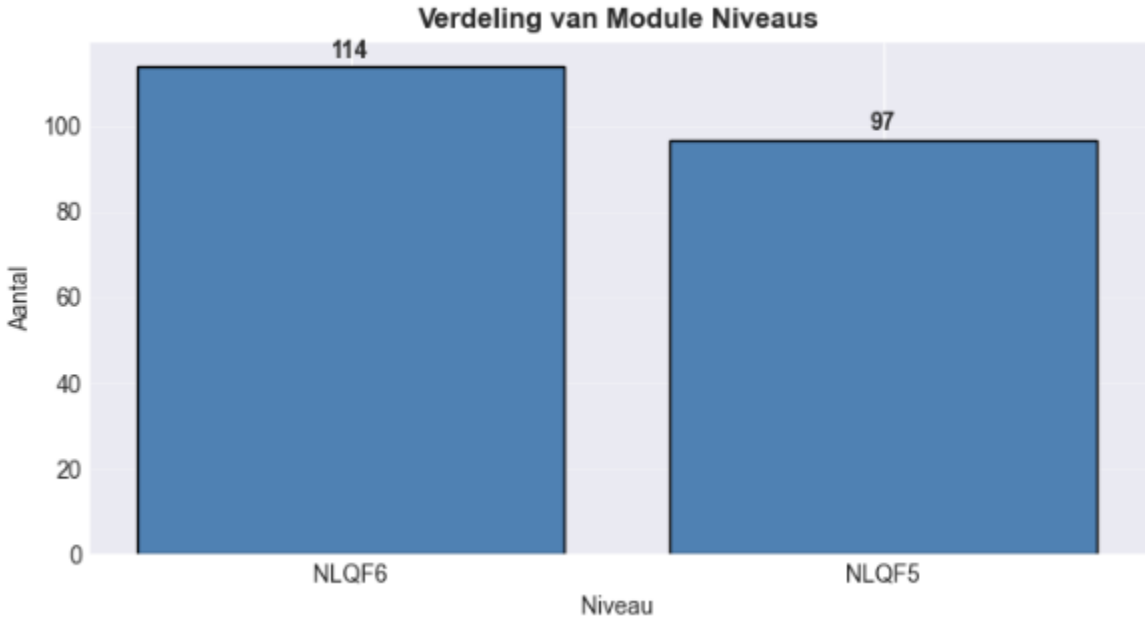
Level

De variabele *level* komt in twee waarden voor:

- **NLQF6**: 114 modules
- **NLQF5**: 97 modules

De verdeling is relatief gebalanceerd, met een lichte oververtegenwoordiging van NLQF6. Dit maakt het mogelijk om onderscheid te maken tussen modules op verschillende niveaus, zonder dat één niveau de dataset domineert.

Figuur 7.6:



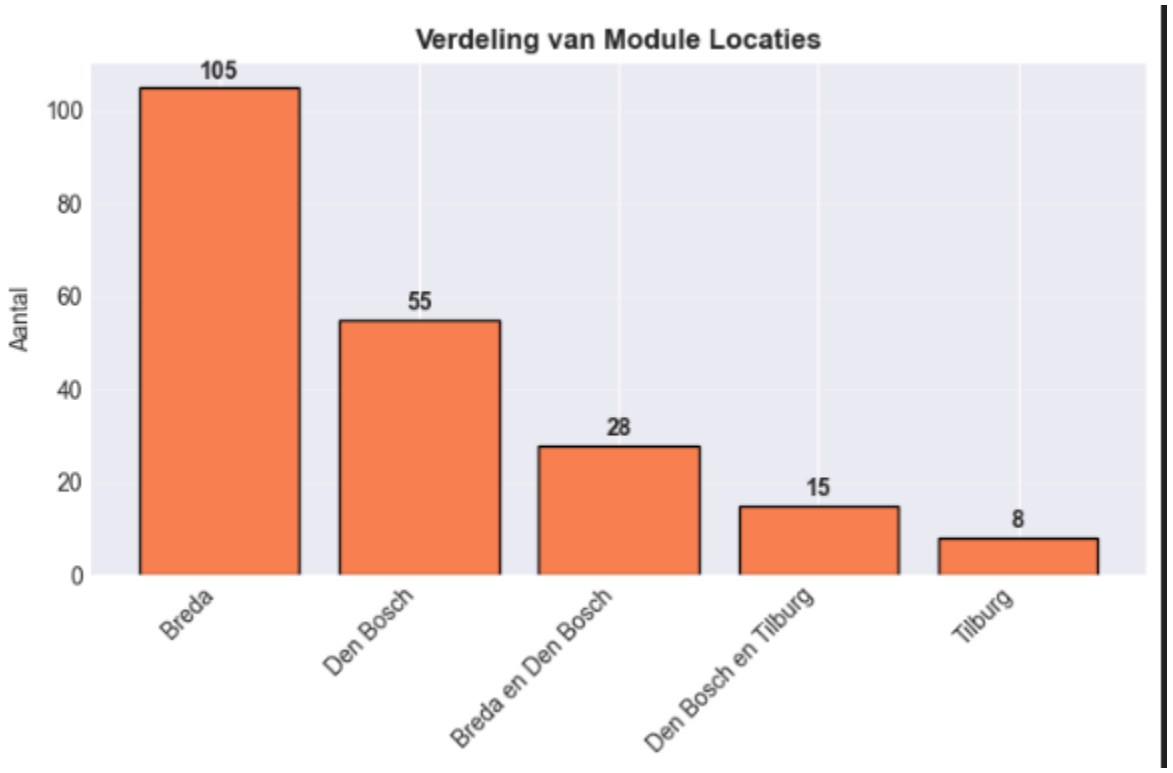
Location

De belangrijkste locaties in de dataset zijn:

- Breda
- Den Bosch
- Breda en Den Bosch
- Den Bosch en Tilburg
- Tilburg

Breda heeft de meeste modules, gevolgd door Den Bosch, terwijl combinatie-locaties een kleiner aandeel vormen. Dit is relevant voor filters in de interface (studenten kunnen hun eigen locatie selecteren) en voor latere analyses om te zien of bepaalde locaties meer of minder aanbod hebben.

Figuur 7.7:



Korte conclusie univariate analyse

De univariate analyse laat zien dat:

- de dataset numeriek stabiel is, zonder extreme outliers;
- *studycredit* vooral twee standaardwaarden kent (15 en 30 ECTS);
- *interests_match_score* en *popularity_score* voldoende variatie bieden om nuance in aanbevelingen aan te brengen;
- *estimated_difficulty* en *available_spots* geschikt zijn als filter- en voorkeurvariabelen;
- *level* en *location* logisch verdeeld zijn, aansluitend bij de onderwijsstructuur;

Dit vormt een solide basis om in de volgende subsecties (6.2 bivariate analyse, 6.3 multivariate patronen) dieper in te gaan op verbanden tussen variabelen en hun bruikbaarheid voor het AI-prototype.

Bivariate Analyse

In de bivariate analyse wordt steeds naar twee variabelen tegelijk gekeken. Doel is om te begrijpen welke kenmerken elkaar (nauwelijks of juist wel) beïnvloeden en welke combinaties zinvol zijn voor het aanbevelingssysteem. De focus ligt op de belangrijkste numerieke variabelen en de koppeling met de categorische variabele *level*.

De analyses zijn uitgevoerd op de numerieke velden:

- studycredit
- interests_match_score
- popularity_score
- estimated_difficulty
- available_spots

en de categorische variabelen:

- level
- Location

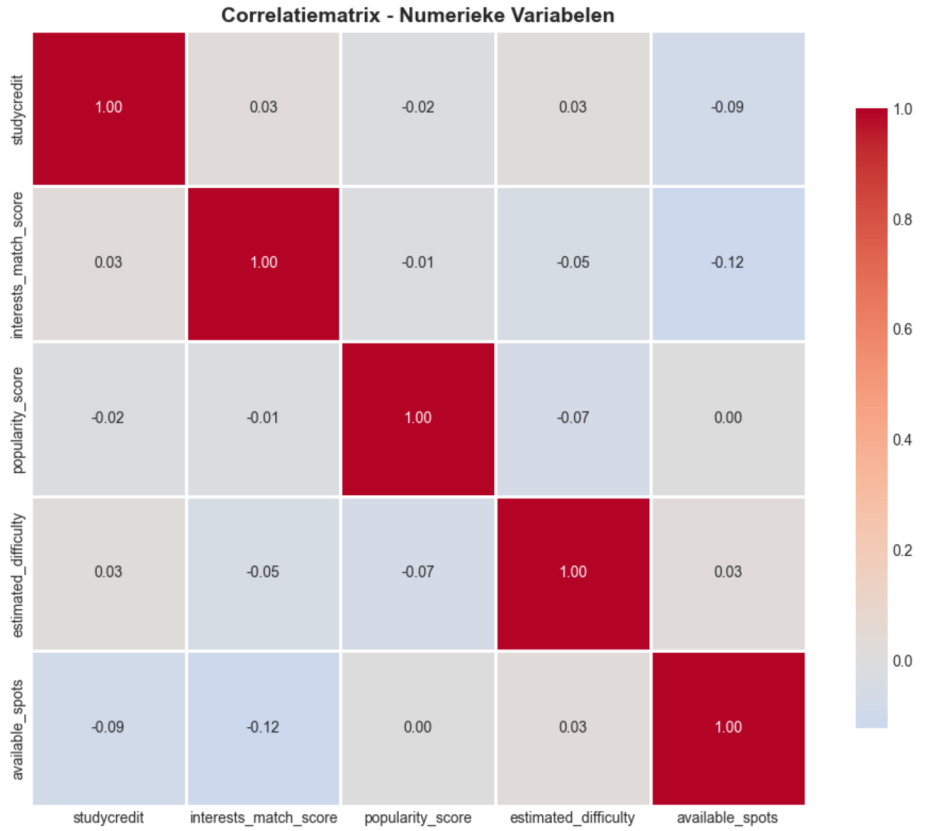
Correlatiematrix van numerieke variabelen

De correlatiematrix van de vijf belangrijkste numerieke variabelen – *studycredit*, *interests_match_score*, *popularity_score*, *estimated_difficulty* en *available_spots* – laat zien dat de onderlinge relaties zeer zwak zijn. Alle correlaties liggen dicht bij nul, zowel positief als negatief.

Belangrijkste observaties:

- *interests_match_score* heeft vrijwel geen samenhang met *popularity_score* (−0,01).
- *studycredit* hangt nauwelijks samen met moeilijkheidsgraad of populariteit.
- *available_spots* vertoont geen merkbare relatie met enige numerieke variabele.
- *estimated_difficulty* correleert zwak negatief met populariteit (−0,07), maar dit is te gering om als betekenisvol te beschouwen.

Figuur 7.8:



Interpretatie:
Omdat de correlaties zo laag zijn, is een puur numeriek regressiemodel niet zinvol als hoofdmechanisme. Dit bevestigt dat een **content-based, tekstgebaseerde aanpak** geschikter is dan een model dat zwaar op numerieke relaties leunt.

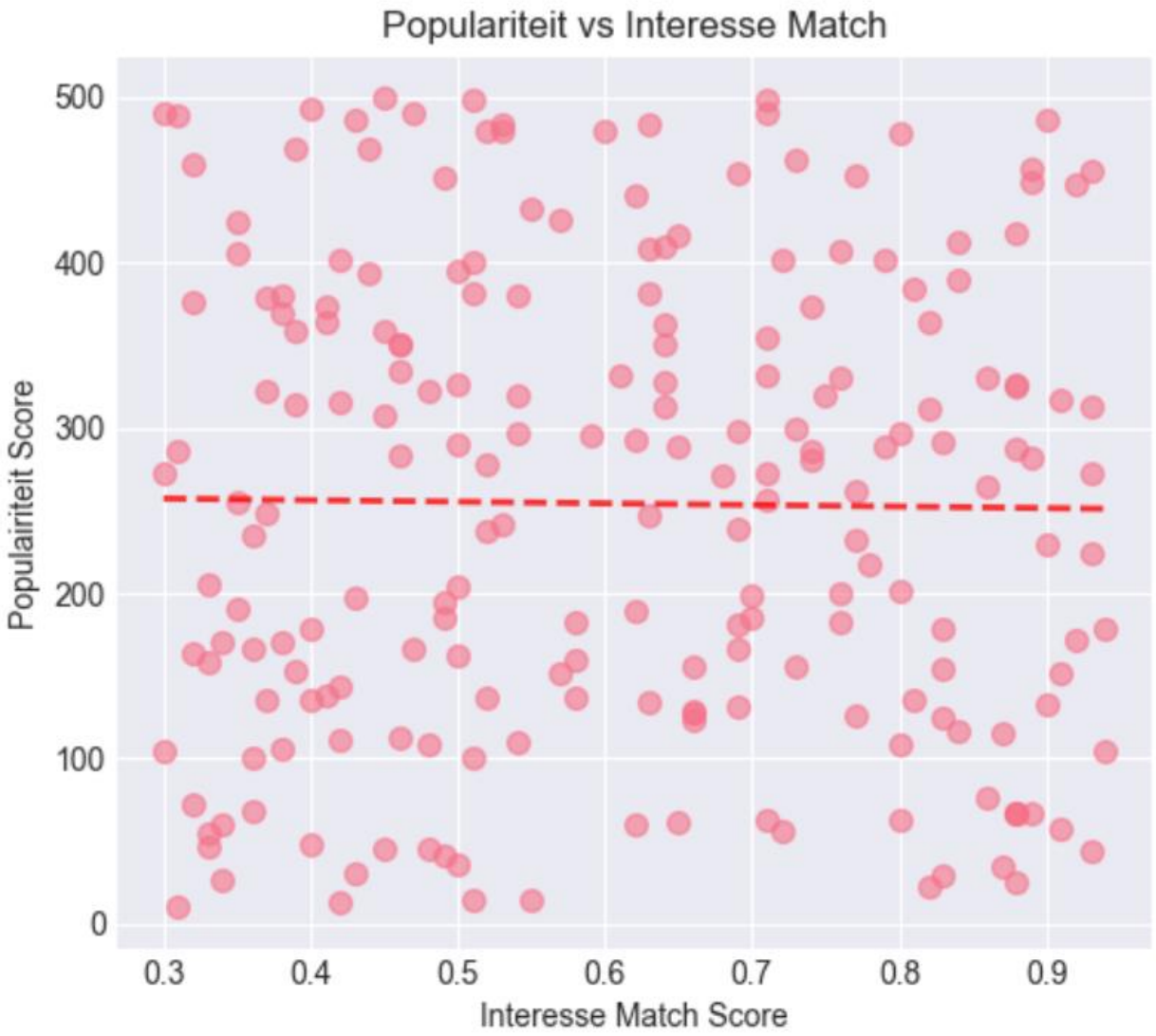
Relaties tussen numerieke variabelen (spreidingsdiagrammen)

Populariteit vs. Interests_match_score

De puntenverdeling toont een brede spreiding waarbij zowel hoge als lage populariteit voorkomt bij uiteenlopende interest match scores. Er is géén zichtbaar verband en de regressielijn helt vrijwel horizontaal.

- hoge en lage *interests_match_score* komen voor bij zowel hoge als lage *popularity_score*;
- de correlatie is praktisch nul (≈ −0,01).

Figuur 7.9

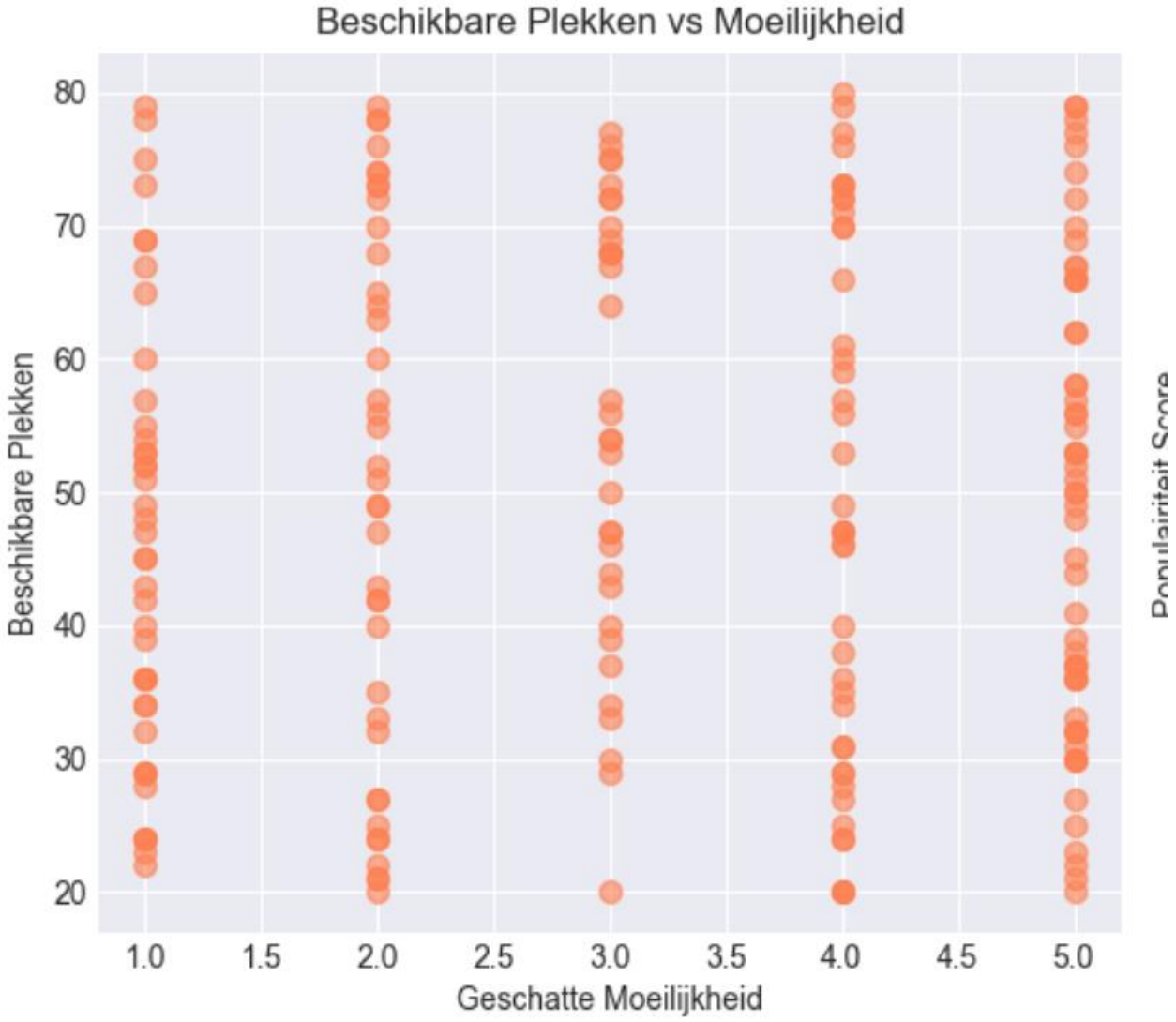


Interpretatie:
Modules die qua inhoud goed aansluiten op interesse hebben niet automatisch een hogere populariteit. Deze twee grootheden moeten dus gescheiden worden behandeld: *interests_match_score* als inhoudelijk signaal, *popularity_score* als secundaire indicator.

Estimated Difficulty vs. Available Spots

De spreiding is homogeen over de grafiek verdeeld. Moeilijkere modules hebben niet meer of minder beschikbare plaatsen dan eenvoudigere modules.

Figuur 7.10:

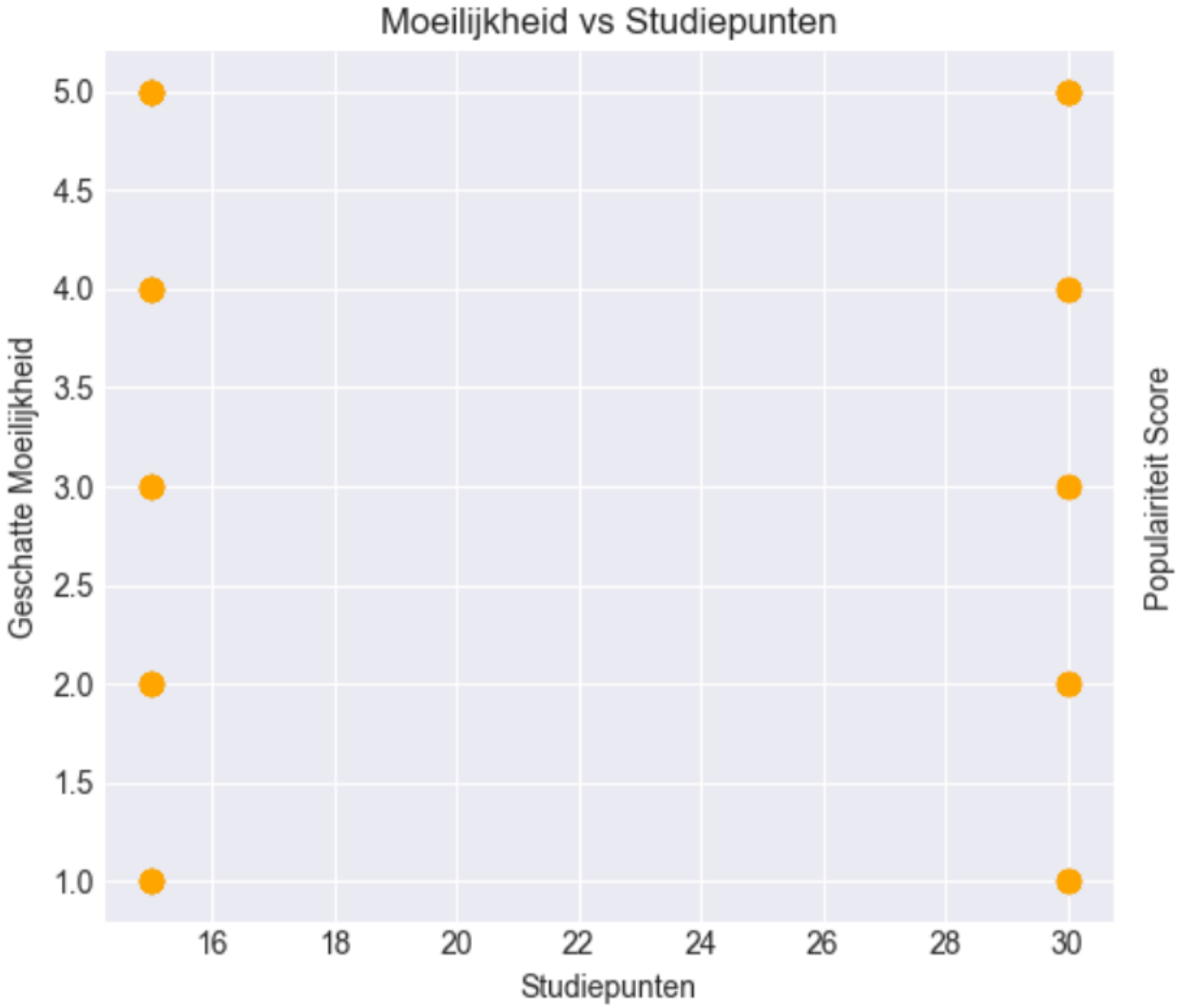


Interpretatie:
Capaciteit wordt niet bepaald door moeilijkheidsgraad. Beide variabelen kunnen onafhankelijk van elkaar worden gebruikt in filters.

Studycredit vs. Estimated Difficulty

Studiepunten komen hoofdzakelijk voor in twee waarden: 15 en 30. Voor beide waarden is de moeilijkheidsgraad gelijkmatig verdeeld over niveaus 1 t/m 5.

Figuur 7.11.

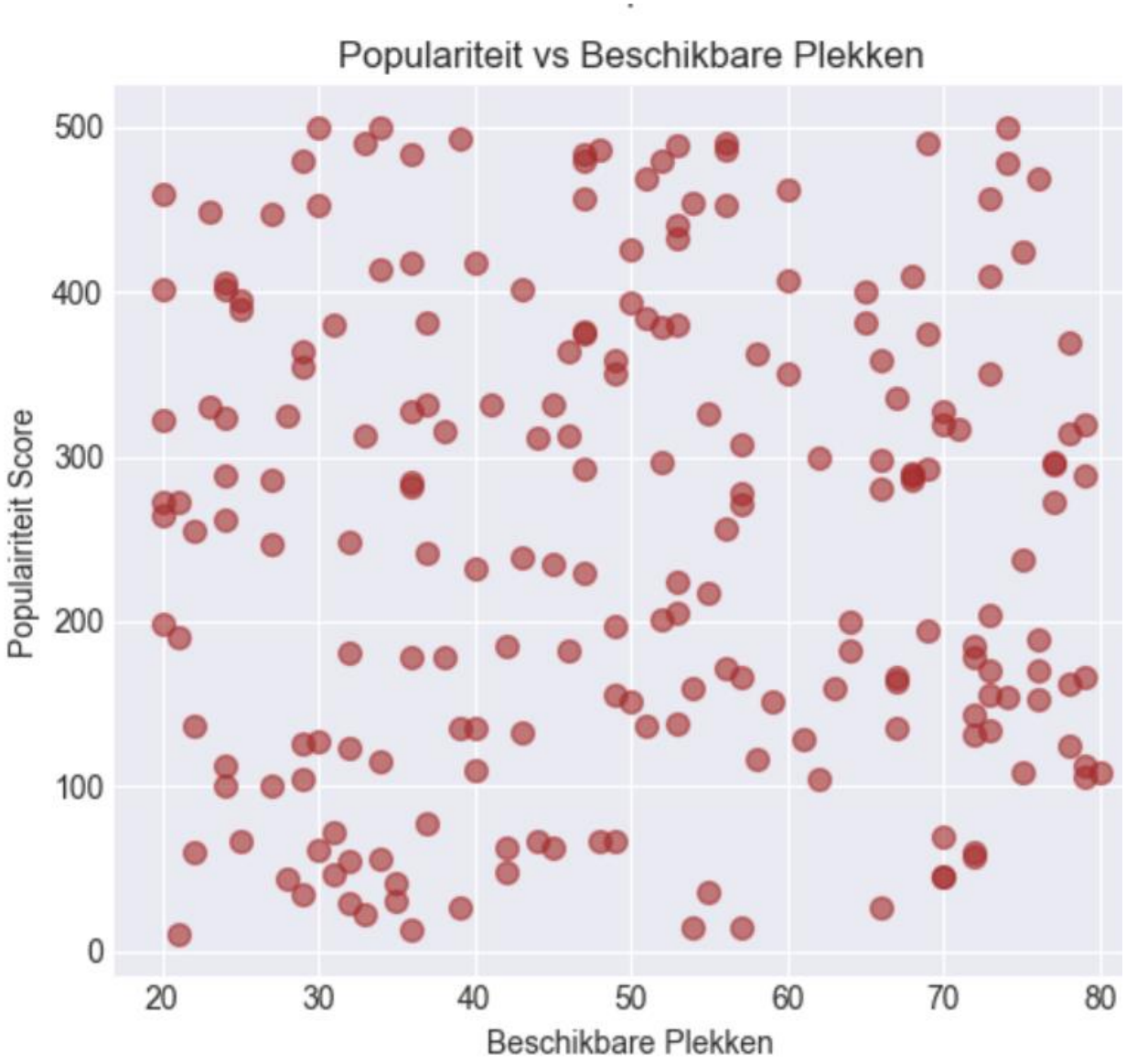


Interpretatie:
Studiepunten vormen vooral een structurele classificatie van modules en zijn geen indicator voor de complexiteit of inspanning.

Available Spots vs. Popularity

De puntenwolk toont dat modules met veel capaciteit niet automatisch populair zijn en modules met weinig plekken niet minder populair.

Figuur 7.11:



Interpretatie:
Capaciteit en populariteit zijn onafhankelijk. De hoeveelheid beschikbare plaatsen kan worden gebruikt als praktische filter, niet als voorspeller.

Categorische variabelen gecombineerd met numerieke variabelen

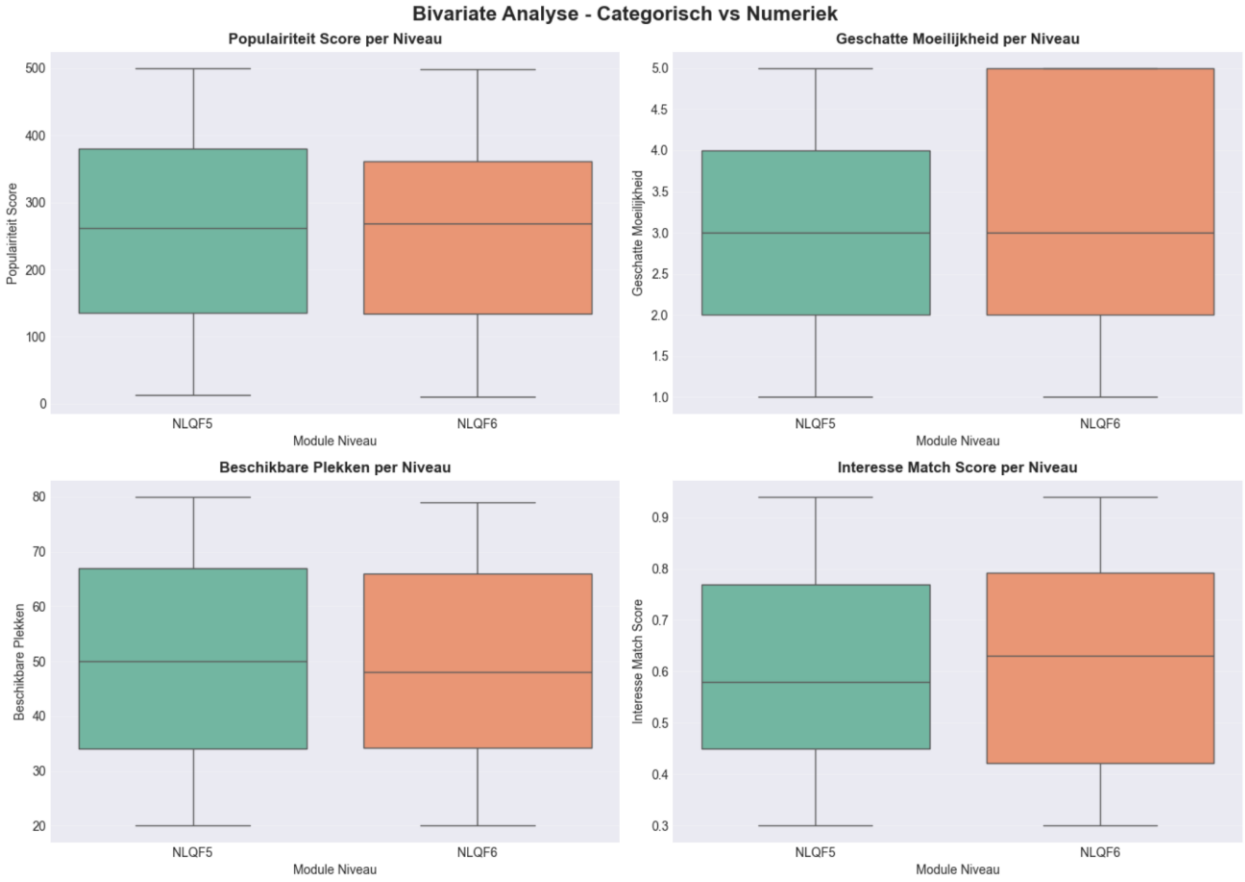
Level vs. Popularity, Difficulty, Spots en Interests

Voor de categorische variabele *level* (NLQF5 en NLQF6) zijn boxplots gemaakt ten opzichte van de vier belangrijkste numerieke variabelen.

Belangrijkste bevindingen:

- Populariteit is vergelijkbaar tussen beide niveaus; verschillen zijn minimaal.
- Moeilijkheidsgraad heeft een lichte stijging richting NLQF6, maar de overlap is groot.
- Available_spots zijn vrijwel gelijk verdeeld over beide niveaus.
- Interests_match_score toont geen relevant verschil tussen NLQF5 en NLQF6.

Figuur 7.12:



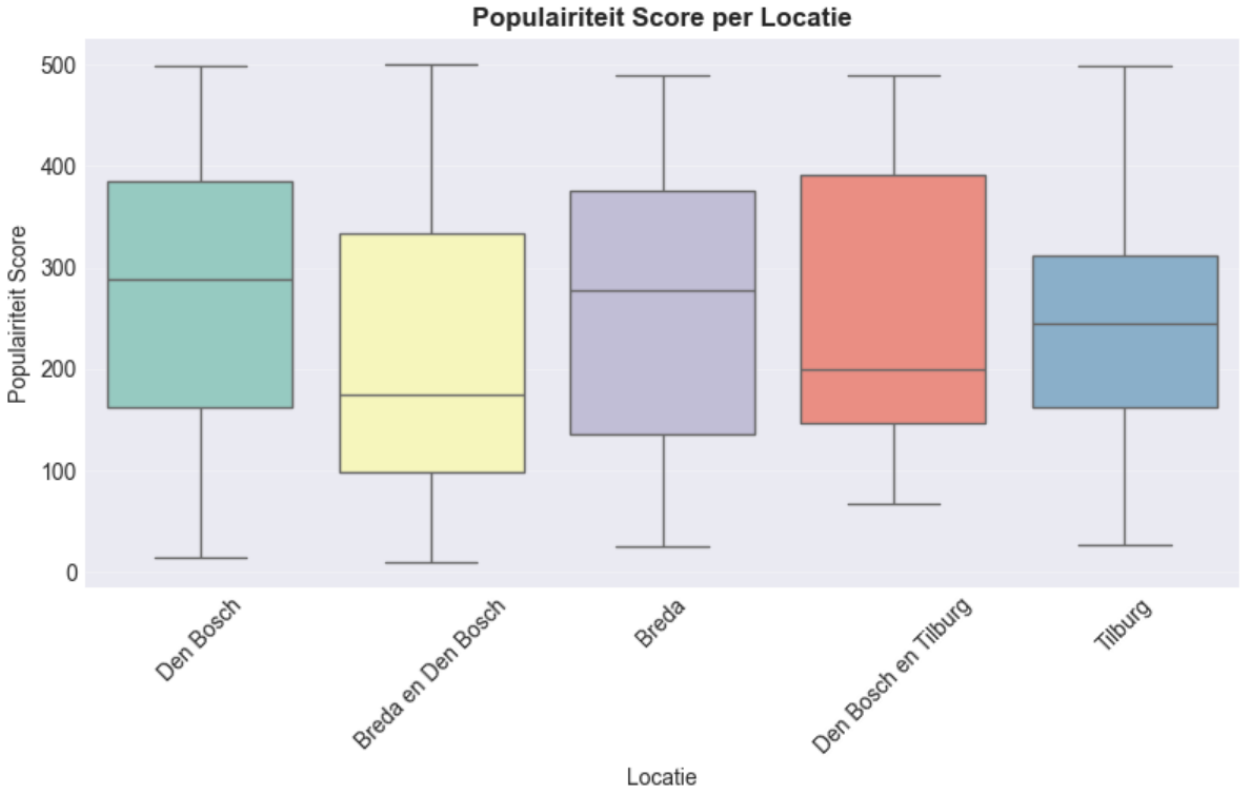
Interpretatie:
Het niveau van een module heeft geen sterke invloed op populariteit, interesse of moeilijkheid. Het is primair een classificatiekenmerk voor filtering.

Location vs. Popularity

Een vergelijking tussen de vijf meest voorkomende locaties toont kleine verschillen in gemiddelde populariteit. Breda en Den Bosch scoren licht hoger, terwijl modules met

gecombineerde locaties gemiddeld iets lager scoren. De spreiding binnen elke locatiegroep is echter groot.

Figuur 7.13:



Interpretatie:
Locatie heeft slechts beperkte invloed op populariteit. Het is nuttig als filter, maar geen sterke aanbevelingsfactor.

Samenvatting van de bivariate analyse

- Er bestaan **geen sterke lineaire relaties** tussen de numerieke variabelen.
- Populariteit wordt **niet verklaard** door moeilijkheidsgraad, studiepunten, capaciteit of interests-match.
- Niveau en locatie zijn **relevante contextvariabelen**, maar geen bepalende factoren voor populariteit of inhoudelijke overeenkomsten.
- De dataset bevat **weinig numerieke structuur**, wat de keuze voor een **tekstgebaseerde content-based recommender** extra ondersteunt.

Multivariate Analyse

In de multivariate analyse worden combinaties van drie of meer variabelen tegelijkertijd bekeken. Waar de bivariate analyse vooral één-op-éénrelaties uitlegt, richt deze sectie zich op patronen die alleen zichtbaar worden als meerdere kenmerken samen worden genomen, zoals niveau én studiepunten én populariteit. Dit levert extra inzicht op in hoe modules zich gedragen binnen de onderwijscontext en welke combinaties van kenmerken interessant zijn voor het ontwerp van de aanbevelingstool.

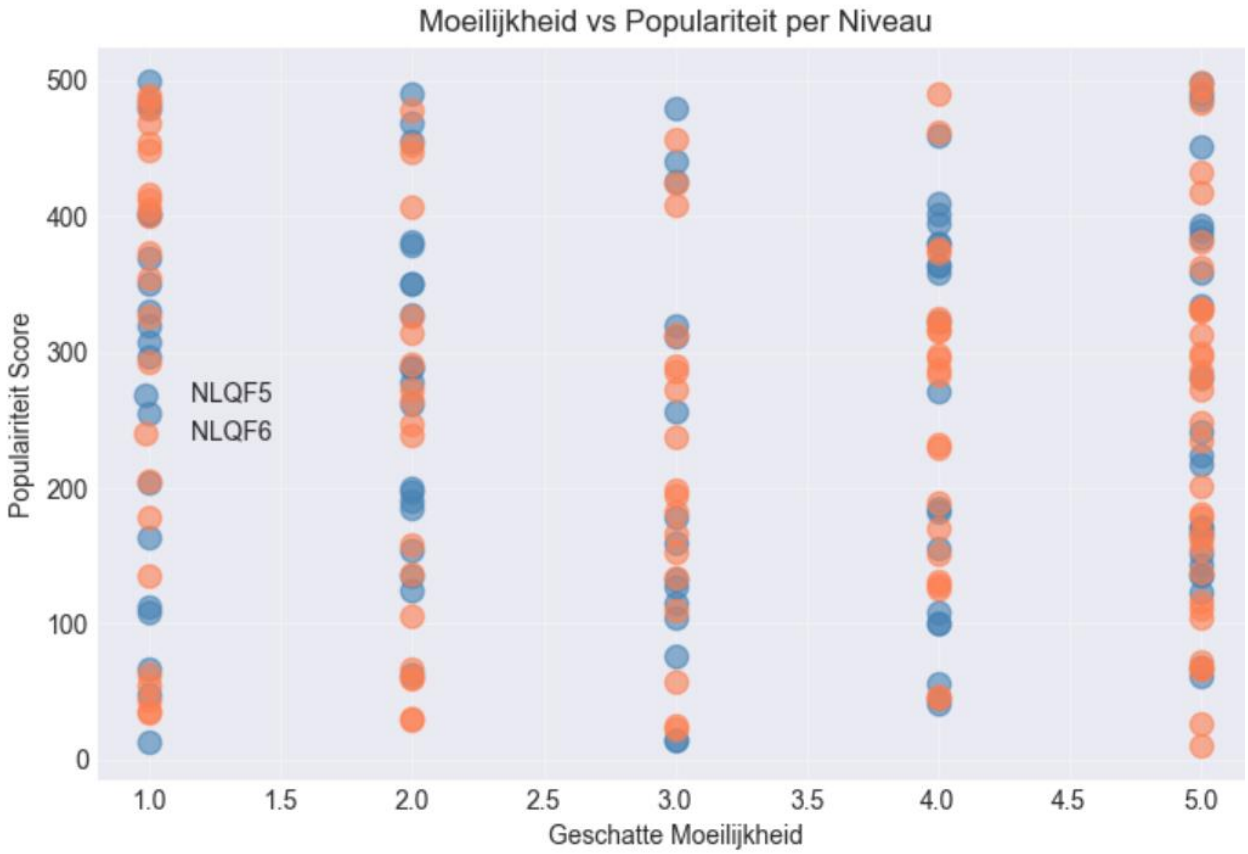
Niveau als structurerende factor

Een eerste multivariate invalshoek is het bekijken van hoe het moduleniveau (*level*, NLQF5 of NLQF6) samenhangt met zowel moeilijkheidsgraad (*estimated_difficulty*) als populariteit (*popularity_score*).

Wanneer moeilijkheid wordt uitgezet tegen populariteit en de punten worden gekleurd per niveau, ontstaat het volgende beeld:

- zowel NLQF5- als NLQF6-modules komen voor bij alle moeilijkheidsniveaus (1–5);
- voor beide niveaus zijn er populaire én minder populaire modules;
- er is geen duidelijke scheiding tussen de niveaus in termen van populariteit bij een bepaalde moeilijkheid.

Figuur 7.14



Interpretatie:
Niveau bepaalt het type module, maar niet op een duidelijke manier de combinatie van moeilijkheid en populariteit. NLQF5 en NLQF6 vormen hierdoor geen afzonderlijke “werelden” in de dataset, maar overlappende domeinen.

Niveau, studiepunten en populariteit

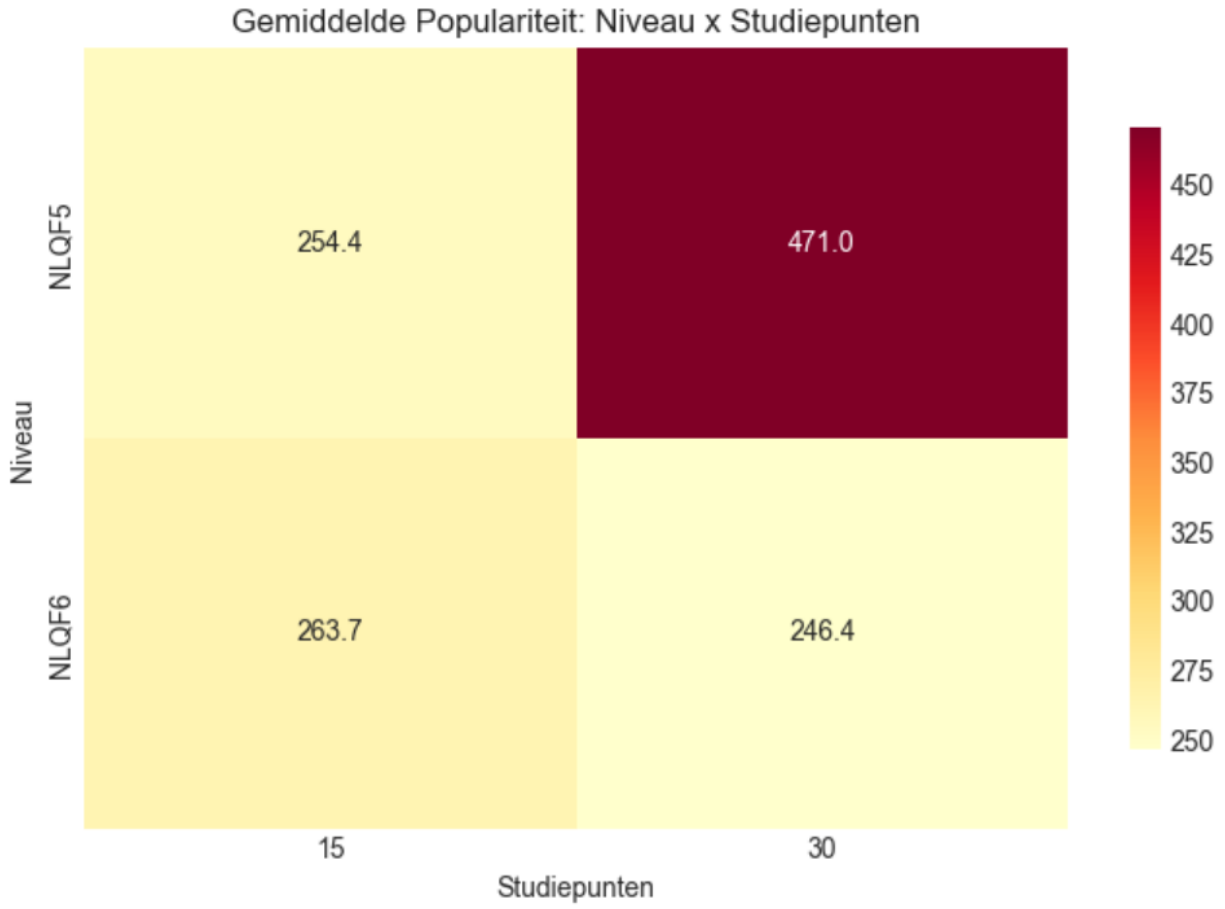
Een tweede multivariate invalshoek is de combinatie van:

- *level* (NLQF5 / NLQF6)
- *studycredit* (15 of 30 ECTS)
- *popularity_score* (gemiddelde populariteit binnen elke combinatie)

De gemiddelde populariteit per combinatie van niveau en studiepunten ziet er als volgt uit:

Niveau	Studiepunten	Aantal modules	Gemiddelde populariteit
NLQF5	15	95	± 254
NLQF5	30	2	± 471
NLQF6	15	26	± 264
NLQF6	30	88	± 246

Figuur 7.15



Belangrijke observaties:

- bij **15 ECTS** liggen de gemiddelde populariteitsscores van NLQF5 en NLQF6 dicht bij elkaar;
- bij **30 ECTS** zijn er maar **twee NLQF5-modules**, die gemiddeld zeer populair zijn (± 471), maar de steekproef is hier klein;
- de meeste 30 ECTS-modules vallen onder NLQF6 en hebben een gemiddelde populariteit rond 246.

Interpretatie:
De opvallend hoge gemiddelde populariteit van NLQF5-modules met 30 ECTS moet voorzichtig geïnterpreteerd worden vanwege het zeer kleine aantal modules. Over het geheel genomen lijkt populariteit meer bepaald te worden door de specifieke inhoud van een module dan door de combinatie van niveau en studiepunten.

Niveau, moeilijkheid en interesse-match

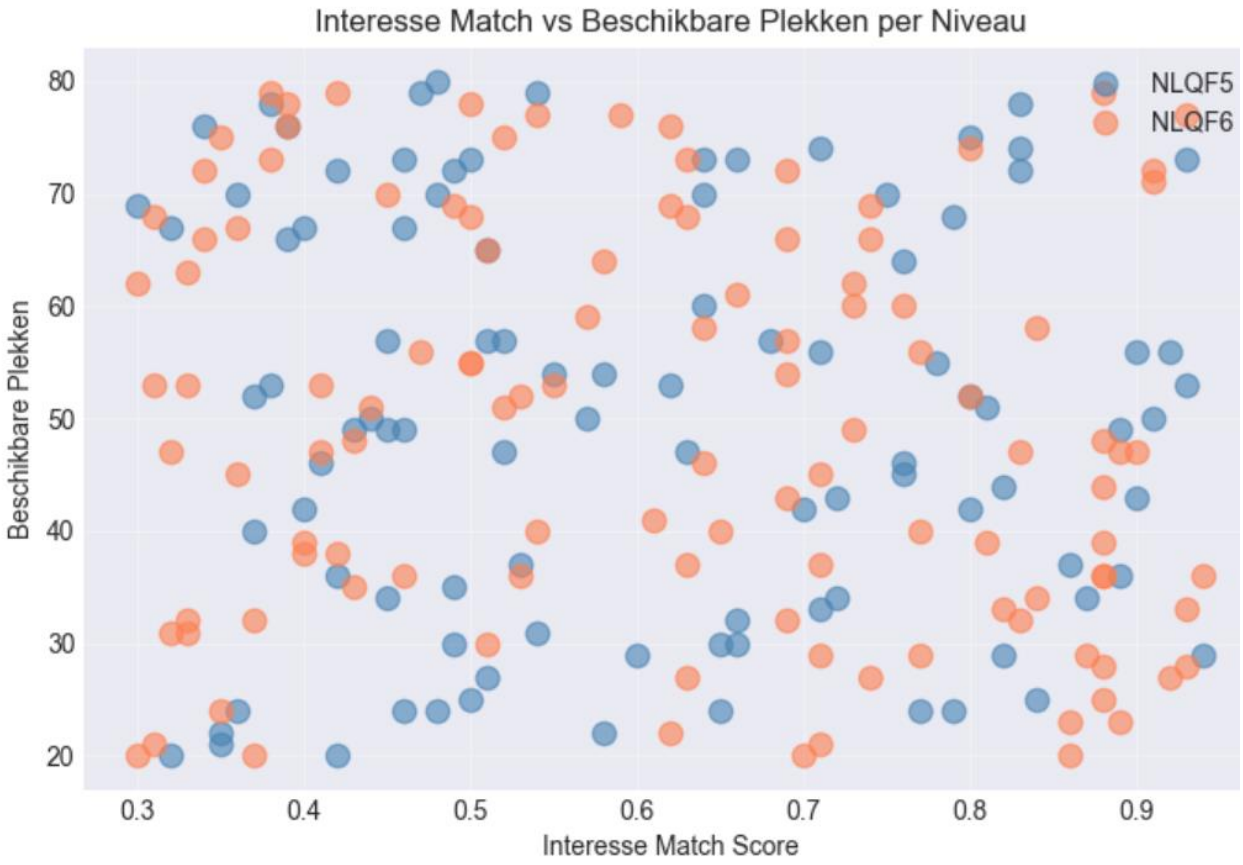
Een andere relevante combinatie is die tussen:

- *level* (NLQF5 / NLQF6)
- *estimated_difficulty*
- *interests_match_score*

Wanneer moeilijkheidsgraad wordt afgezet tegen interests_match_score, per niveau, ontstaat een verspreid patroon:

- binnen elk niveau komen zowel lage als hoge interests_match_scores voor bij elk moeilijkheidsniveau;
- er is geen duidelijke “zone” waarin bijvoorbeeld moeilijke modules systematisch beter of slechter matchen met interesses;
- de gemiddelde interests_match_score per niveau ligt dicht bij elkaar (rond 0,60–0,62), met vergelijkbare spreiding.

Figuur 7.16



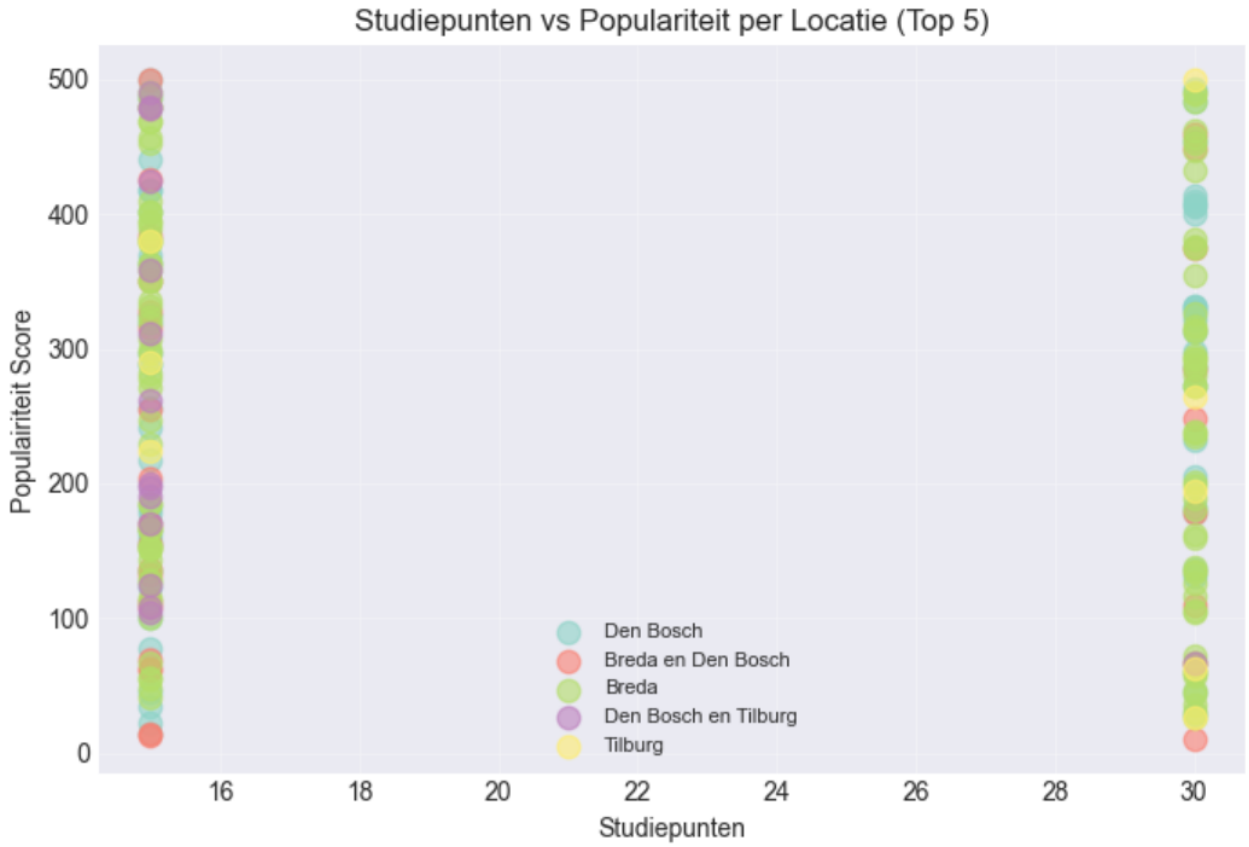
Interpretatie:
De mate waarin een module aansluit bij interesses lijkt vooral bepaald door inhoudelijke kenmerken (tekst, tags, thema's), niet door niveau of moeilijkheid op zich. Dit ondersteunt opnieuw het belang van tekstgebaseerde representaties in de latere modelbouw.

Niveau, locatie en populariteit (gecombineerde context)

Hoewel locatie in eerste instantie bivariate is geanalyseerd, levert de combinatie van *location* en *level* met *popularity_score* extra context op. In grote lijnen blijkt dat:

- zowel NLQF5- als NLQF6-modules gelijkmatig over de belangrijkste locaties (zoals Breda en Den Bosch) zijn verdeeld;
- de gemiddelde populariteit per locatie wel iets varieert, maar binnen elk niveau een grote spreiding kent;
- er geen signaal is dat bijvoorbeeld NLQF6-modules op een specifieke locatie structureel populairder zijn dan elders.

Figuur 7.17



Interpretatie:
Locatie en niveau zijn vooral logistieke en curriculaire parameters. Ze zijn relevant voor de student als filter (ik wil modules op mijn locatie en passend bij mijn niveau), maar minder geschikt als "drijvende" features voor een aanbevelingsalgoritme.

Conclusie multivariate analyse

De multivariate analyse laat zien dat:

- combinaties van *level*, *studycredit* en *popularity_score* geen sterke structurele patronen opleveren, behalve een enkele kleine subgroep met weinig modules;
- moeilijkheidsgraad, interests_match_score en populariteit binnen beide niveaus breed verspreid zijn, zonder duidelijke scheiding tussen NLQF5 en NLQF6;
- locatie en niveau belangrijke **contextuele** eigenschappen zijn, maar inhoudelijk vooral gebruikt moeten worden als **filters** en niet als dominante modelinputs;

- de belangrijkste structuur in de dataset niet in eenvoudige numerieke combinaties zit, maar in de inhoudelijke modulebeschrijvingen en tags.

Dit bevestigt dat de kern van het AI-prototype moet liggen in **content-based analyse** van de tekstvelden (fulltext, tags) en dat numerieke en categorische variabelen vooral ondersteunend moeten worden ingezet als filters en aanvullende signalen.

7. Dimensionality Reduction (PCA, t-SNE, UMAP)

In dit hoofdstuk onderzoeken we de hoog-dimensionale moduledata met drie gangbare dimensionality-reduction technieken: PCA, t-SNE en UMAP. Het doel hiervan is: het reduceren van overlappende informatie in de features, en het zichtbaar maken van eventuele clusters of patronen.

Doel van dimensionality reduction

Dimensionality reduction wordt toegepast met de volgende doelen:

- Visualiseren van hoog-dimensionale data

Onze dataset bevat meerdere numerieke kenmerken per module. Door deze terug te brengen naar twee dimensies kunnen we modules in een 2D-plot visueel vergelijken.

- Verminderen van feature-overlap

Sommige variabelen bevatten overlappende informatie. Door deze in minder dimensies samen te vatten, kunnen we beter zien welke variabelen echt belangrijk zijn.

PCA

PCA is een lineaire methode die de richting van maximale variantie in de data zoekt. Hierdoor ontstaat een set componenten die samen een groot deel van de datavariatie beschrijven.

Resultaten:

- PC1 verklaart 27,22% van de variantie
- PC2 verklaart 16,76% van de variantie
- Samen leggen PC1 + PC2 43,98% van de totale variantie uit

Dit betekent dat iets minder dan de helft van alle informatie kan worden weergegeven in twee dimensies. De dataset bevat dus meerdere belangrijke richtingen en geen dominante lineaire structuur.

Belangrijkste bijdragen aan PC1:

1. contact_id (0.7016)
2. id (0.6993)
3. popularity_score (0.1062)
4. available_spots (0.0629)
5. interests_match_score (0.0403)

De eerste twee variabelen domineren de lineaire structuur van PC1.

Interpretatie van PCA-visualisatie:

De PCA-projectie toont een brede spreiding van modules zonder duidelijke clusters. Dit bevestigt dat lineaire scheiding niet voldoende is om eventuele modulegroepen te identificeren. De cumulatieve-variantieplot laat zien dat ongeveer 6 of 7 componenten nodig zijn om 95% van de informatie te behouden, wat de hoge complexiteit van de dataset onderbouwt.

t-SNE

t-SNE is een niet-lineaire techniek die zich richt op het behouden van lokale structuren. Punten die dichtbij elkaar liggen in de originele ruimte worden ook dichtbij elkaar geprojecteerd in 2D.

Resultaten:

- T-SNE converteerde succesvol na 1000 iteraties (KL-divergence: 0.66)
- De gemiddelde afstand tussen punten in de nieuwe ruimte is 12.81
- Clustering (DBSCAN) detecteerde meerdere clusters en enkele losse punten

Observaties uit de visualisaties:

- De t-SNE scatterplot toont duidelijke groepjes modules die dicht bij elkaar liggen.
- De dichtheidsplot toont meerdere gebieden met hoge concentratie, wat wijst op onderliggende substructuren in de data.
- De globale afstanden tussen clusters zijn echter minder betekenisvol (zoals typisch bij t-SNE).

Interpretatie:

t-SNE is geschikt om te zien dat er clusterstructuur in de data zit die door PCA niet werd onthuld. Dit betekent dat de dataset waarschijnlijk niet-lineaire patronen bevat.

UMAP

UMAP is een moderne niet-lineaire techniek die veel lijkt op t-SNE maar sneller is en zowel lokale als globale structuur beter kan behouden.

Resultaten:

- De UMAP-transformatie leverde een 2D-representatie met vorm (211, 2)
- De scatterplot toont vergelijkbare clusters als t-SNE
- De dichtheidsplot laat duidelijk hoge- en lage-dichtheidsgebieden zien

Interpretatie:

UMAP toont dezelfde algemene clusters als t-SNE, maar doet dit consistenter en met behoud van de globale vorm van de dataset. Dit maakt UMAP geschikt als visualisatietool voor het uiteindelijke recommender-systeem, omdat het zowel detail als overzicht biedt.

Vergelijking van de drie technieken

Techniek	Lineariteit	Snelheid	Clusters zichtbaar	Globale structuur	Interpretatie
PCA	Lineair	Zeer snel	Nee	Redelijk	Goed
t-SNE	Niet-lineair	Traag	Zeer duidelijk	Slecht	Redelijk
UMAP	Niet-lineair	Snel	Duidelijk	Goed	Gemiddeld

Conclusie:

PCA laat de globale lineaire structuur zien, maar toont geen duidelijke clusterpatronen. t-SNE maakt meerdere clusters zichtbaar en toont dat de dataset niet-lineaire structuren bevat. UMAP combineert de voordelen van beide methoden en biedt een consistente, duidelijke visualisatie van het modullandschap. Voor het uiteindelijke doel is UMAP daarom de meest informatieve reductietechniek.

8. Feature Engineering

In dit hoofdstuk beschrijven we hoe zowel het studentprofiel als de modulevectoren zijn gecreëerd. Deze representaties vormen de basis voor het aanbevelingssysteem en bepalen hoe goed overeenkomsten tussen studenten en modules kunnen worden berekend.

Constructie van het Studentprofiel

Het studentprofiel is ontworpen als een compacte representatie van voorkeuren en doelen van de gebruiker. In lijn met het recommender-systeemraamwerk bestaat het profiel uit drie componenten: interesses, waarden en loopbaandoelen.

Interesses:

Interesses verwijzen naar inhoudelijke voorkeuren van de student, bijvoorbeeld:

- voorkeur voor technische of creatieve vakken
- interesse in praktijkopdrachten of onderzoeksgericht werk
- thema's zoals softwareontwikkeling, marketing, zorg, data-analyse, enz.

Deze interesses worden later gekoppeld aan inhoudelijke kenmerken van modules.

Waarden:

Waarden beschrijven de manier van werken of studeren die de student belangrijk vindt, zoals:

- samenwerken of zelfstandig werken
- projectgebaseerd leren
- theoretische verdieping tegenover praktische uitvoering

Deze waarden kunnen worden gematcht met kenmerken in de modulebeschrijvingen, zoals leeruitkomsten of werkvormen.

Loopbaandoelen:

Loopbaandoelen richten zich op wat de student uiteindelijk wil bereiken, bijvoorbeeld:

- doorstromen naar een specifieke richting
- ontwikkelen van professionele vaardigheden
- voorbereiding op een bepaald beroep of vakgebied

Deze informatie helpt bij het prioriteren van modules die inhoudelijk relevant zijn voor toekomstige doelen.

Constructie van Modulevectoren

Voor het aanbevelingssysteem is elke module omgezet naar een numerieke vector. Dit maakt het mogelijk om modules onderling én met het studentprofiel te vergelijken.

Metadata:

Elke module bevat meerdere vormen van metadata die zijn meegenomen bij het construeren van de modulevectoren:

- Inhoud: beschrijvingen en leeruitkomsten
- Categorie/Subdomein: bijvoorbeeld ICT, business, techniek, sociaal domein
- Niveau: jaar, niveau van moeilijkheid of richting

Hoewel deze metadata niet numeriek zijn, worden ze meegenomen door ze te verwerken in de tekstrepresentatie of door ze als afzonderlijke kenmerken te coderen.

Tekstverwerking:

De belangrijkste bron voor de modulevectoren bestaat uit tekstuele velden uit de dataset. In plaats van Bag-of-Words tellers of embeddings is gekozen voor:

TF-IDF (Term Frequency – Inverse Document Frequency)

Redenen voor deze keuze:

- TF-IDF is transparant en goed uitlegbaar
- het benadrukt onderscheidende woorden tussen modules
- het levert sterk discriminatieve vectoren
- gemakkelijk te gebruiken voor studentprofielen (zelfde vectorizer)

TF-IDF Configuratie

In het systeem is een finale TF-IDF vectorizer gebruikt met:

- unigrams + bigrams
- max 6000 feature
- min_df = 2
- max_df = 0.8
- sublinear_tf = True (log-scaling)

Alle tekstuele kolommen zijn samengevoegd tot één combined_text veld per module, waarna TF-IDF hierop is toegepast. Dit resulteert in een hoge-dimensionale, sparse vector per module.

Overzicht van het Feature Engineering-proces

Het totale proces bestaat uit twee stappen:

1. Studentprofiel

- Interesses → inhoudelijke voorkeuren
- Waarden → manier van leren of werken
- Loopbaandoelen → toekomstige richting
- Gevectoriseerd met dezelfde TF-IDF vectorizer als de modules

2. Modulevectoren

- Tekstuele inhoud gecombineerd in één document
- TF-IDF vectorisatie met 6000 features
- Resultaat: een consistente representatie van alle modules

Samenvatting

Met behulp van TF-IDF feature engineering zijn zowel studenten als modules omgezet naar gestructureerde representaties.

Het studentprofiel beschrijft voorkeuren en doelen; de modulevectoren bevatten een rijk en gedetailleerd tekstprofiel van elke module.

Deze combinatie maakt het mogelijk om overeenkomsten te berekenen en relevante moduleaanbevelingen te genereren via cosine similarity.

9. Model Training & Recommender Design

In dit hoofdstuk beschrijven we de keuzes die zijn gemaakt bij het ontwerp van het aanbevelingssysteem. Het doel is inzicht geven in hoe het systeem modules vergelijkt, hoe aanbevelingen worden gegenereerd en hoe de kwaliteit beoordeeld kan worden.

Keuze van Aanpak

Bij recommender systems kan gekozen worden tussen content-based, collaboratieve filtering, of hybride methoden. Voor dit project is gekozen voor een content-based aanpak.

Motivatie voor Content-Based:

- De beschikbare dataset bevat geen gebruikers-interactie data (zoals ratings of klikgeschiedenis). Daardoor zijn collaboratieve filtering-technieken niet toepasbaar.
- De tekstuele modulebeschrijvingen zijn rijk en geschikt voor TF-IDF analyse
- TF-IDF maakt transparante matching mogelijk

Voordelen:

- Direct inzetbaar zonder gebruikersdata
- Uitlegbaar: matching gebeurt op inhoudelijke overeenkomst
- Geen cold-start probleem voor gebruikers
- Hogere controle en stabiliteit dan embeddings in kleine domeinen

Beperkingen:

- Geen “anderen vonden ook” effect
- Matching is beperkt tot tekstinhoud
- Mist leereffecten die op interactiedata gebaseerd zijn

Deze aanpak sluit aan bij de Recommender Systems-slides: *item representatie → feature vectoren → similarity berekening*.

Modelimplementatie

Het systeem maakt gebruik van een self-built Content-Based Student-Module Recommender, waarin module- en studentprofielen worden gevectoriseerd via TF-IDF.

Representatie van Modules

- Alle tekstvelden worden samengevoegd tot één tekstdocument per module
- Deze worden gevectoriseerd met de gekozen TF-IDF vectorizer
- Resultaat: een sparse matrix met 6000 features

Cosine Similarity:

Om te bepalen hoe sterk twee modules op elkaar lijken, wordt cosine similarity berekend.

Voordelen:

- schaal-invariant
- efficiënt te berekenen op sparse matrices
- standaard bij TF-IDF

Performance

Dankzij het opslaan van:

- de TF-IDF matrix (tfidf_matrix.npz)
- de vectorizer (tfidf_vectorizer.pkl)

kan het systeem snel vergelijkingen uitrekenen zonder opnieuw te fitten.

Training en Validatie

Er wordt geen machine learning model getraind. TF-IDF is een deterministische transformatie.

1. Normalisatie

TF-IDF levert al genormaliseerde vectoren. Cosine similarity werkt direct goed zonder extra preprocessing.

2. Geen train/test split

- Er zijn geen labels of targets
- Het systeem leert niet van data, maar berekent overeenkomsten
- Evaluatie gebeurt via similarity-statistieken

3. Baseline evaluaties

Uitgevoerd zijn o.a.:

- vergelijkingen tussen TF-IDF configuraties
- gemiddelde similarity in samples
- effecten van unigram/bigram keuzes
- sparsity-analyse

Observaties:

- Unigrams + bigrams leveren meer nuance
- 6000 features geeft beste balans tussen onderscheid en informatiedichtheid
- Similarities liggen in een gezonde range (lage baseline → goede discriminatie)
- Dit is een gangbare evaluatie-methode voor text-based content recommenders.

Samenvatting

- De gekozen aanpak is content-based, vanwege ontbreken van gebruikersdata.
- Modules worden gerepresenteerd via TF-IDF vectoren met unigrams + bigrams (6000 features).
- Aanbevelingen worden bepaald met cosine similarity tussen studentprofiel en modulevectoren.
- Validatie gebeurt via similarity-statistieken en tuning, niet via ML-train/test splits.
- De architectuur sluit aan bij de ML en Recommender Systems-slides:
- item-centrisch, feature-gedreven en uitlegbaar.

10. Evaluatie & Metrics

Het doel van deze evaluatie is om de kwaliteit van het Student-to-Module Recommender System te beoordelen, door te onderzoeken hoe goed de aanbevolen modules passen bij verschillende studentprofielen.

Testprofielen

Er zijn vier representatieve profielen gebruikt:

1. Psychologie & Coaching – interesse in psychologie, coaching en zorg
2. Internationale Zorg – interesse in verpleegkunde en internationale stages
3. Palliatieve Zorg & Rouw – interesse in begeleiding bij rouw, oncologie en culturele diversiteit
4. Data Science in Gezondheidszorg – interesse in AI, machine learning en gezondheidszorgtoepassingen

Gebruikte Metrics

Voor elk profiel zijn de volgende metrics berekend:

- **Cosine similarity scores tussen studentprofiel en modules:**
  - Gemiddelde (avg)
  - Maximaal (max)
  - Minimaal (min)
  - Standaardafwijking (std)

- Coverage: percentage van modules dat voorkomt in de top-N aanbevelingen
- Diversiteit van aanbevelingen: verschillen in top-modules per profiel
- Feature contribution: belangrijkste woorden/concepten die de match verklaren

Resultaten

Similarity Score Distributies:

- Specifieke profielen met unieke woorden (bijv. Palliatieve Zorg) leiden tot hogere gemiddelde similarity scores voor de topmatches.
- Algemene profielen (bijv. Psychologie & Coaching) geven een bredere distributie van scores.

Visualisaties

- Histogrammen van similarity scores per profiel
- Bar charts van top 5 modules per profiel

Top Aanbevelingen:

- Elk profiel krijgt unieke top-N modules.
- De module met de hoogste similarity wordt verder verklaard met feature contribution, waardoor duidelijk is waarom een module past bij het profiel.

Overall Metrics

- Gemiddelde similarity (over alle profielen): ~0.05–0.10 (afhankelijk van profiel)
- Unieke aanbevolen modules: variëren per testprofiel, wat wijst op diversiteit en coverage
- Range max/min similarity toont dat sommige modules veel beter matchen dan andere in de top-N

Observaties & Interpretatie

- Specifiekere profielen → hogere similarity scores voor beste matches
- Algemene profielen → bredere, maar minder sterke matches
- Het systeem verklaart aanbevelingen via feature contributions, wat interpreterbaarheid biedt
- Coverage en diversiteit zijn voldoende om studenten verschillende opties te bieden

Conclusie

Het Student-to-Module Recommender System:

- Werkt effectief voor verschillende studentprofielen
- Levert zowel relevante als interpreteerbare aanbevelingen
- Differentiëert goed tussen specifieke en algemene interesses
- Geeft inzicht in waarom een module past bij een student

11. Modeloptimalisatie

De optimalisatie van het aanbevelingssysteem richtte zich op het versterken van de nauwkeurigheid, stabiliteit en uitlegbaarheid van het model. Aangezien het systeem volledig tekstgebaseerd werkt, ligt de kern van de optimalisatie in het verbeteren van TF-IDF-vectorisatie, het analyseren van de structuur van de dataset en het verfijnen van de similarity-berekening. De aanpassingen zijn systematisch getest en onderbouwd in de ontwikkelde notebooks.

Verbeterde vectorisatie

Een belangrijk onderdeel van de optimalisatie is het verbeteren van de TF-IDF-representatie. De tekstvelden uit de dataset zijn eerst gecombineerd tot één uniforme `combined_text`. Deze stap vergroot de informatiedichtheid en zorgt ervoor dat de vectorisatie altijd werkt op dezelfde volledige context.

Uit de experimenten in `feature_engineering.ipynb` bleek dat een combinatie van unigrams en bigrams de beste semantische precisie oplevert. Waar unigrams vooral breedte bieden, voegen bigrams betekenisvolle verbanden toe die de aanbevelingen inhoudelijk sterker maken. De tuning liet zien dat dit resulteert in een lagere gemiddelde similarity tussen modules, wat erop wijst dat het model modules beter uit elkaar kan houden.

Enkele instellingen die het uiteindelijke model kenmerken zijn:

- `ngram_range = (1, 2)` voor rijke semantiek;
- `max_features = 6000`, om voldoende detail te behouden;
- `min_df = 2`, waardoor ruis uit zeldzame woorden wordt verminderd;
- `sublinear_tf = True`, wat frequentiedominantie tegengaat.

Deze configuratie biedt een goed evenwicht tussen rekenbaarheid en inhoudelijke diepgang.

Hyperparameterafstelling

De tuning van TF-IDF parameters richtte zich onder andere op het verminderen van ruis en het voorkomen dat algemene termen het model verstoren. Door minder vaak voorkomende woorden te filteren en zeer algemene woorden juist te onderdrukken, ontstaat een beter gebalanceerde vectorruimte.

De evaluatie van drie configuraties — unigrams, unigrams + bigrams, en unigrams met Nederlandse stopwoorden — leidde tot duidelijke verbeteringen. Vooral de bigram-variant scoorde goed op:

- **aantal features** (rijk maar beheersbaar),
- **sparsity** (verwacht hoog, maar stabiel),
- **gemiddelde similarity** (lager = betere discriminatie),
- **standaarddeviatie van similarity** (meer variatie = betere nuance).

Deze inzichten leidden tot de definitieve keuze voor bigrams als onderdeel van de vectorisatie.

Dimensionality reduction als diagnose

Hoewel de TF-IDF vectoren zelf niet worden gereduceerd, speelt dimensionality reduction een belangrijke rol in het begrijpen van de datastructuur. In `eda_dimensionality_reduction.ipynb` zijn PCA, T-SNE en UMAP toegepast op de numerieke variabelen.

PCA werd gebruikt om te zien of enkele componenten al een groot deel van de variantie verklaren. Dit bleek beperkt het geval; de numerieke data hebben geen sterke lineaire structuur. T-SNE en UMAP toonden wél herkenbare lokale clusters, maar onvoldoende globale structuur om deze technieken in het aanbevelingsmodel te verwerken.

In deze fase dienden dimensionality reduction-technieken dus vooral als:

- hulpmiddel om dataverdeling beter te begrijpen;
- methode om mogelijke subgroepen te ontdekken;
- controle of numerieke variabelen extra signalen bevatten.

Ze worden niet gebruikt als inputlaag voor de uiteindelijke recommender, maar wel als validatie-instrument.

Verbetering van het aanbevelingsmechanisme

De inhoud van `content_based_recommender.ipynb` toont dat de cosine similarity de basis vormt van de matching. Optimalisatie richtte zich op filtering, drempels en uitlegbaarheid.

Een eerste verbetering was het toepassen van een minimale similarity-drempel. Hierdoor verdwijnen matches die feitelijk geen inhoudelijke relatie hebben met het studentprofiel. Ook is er een niveau-filter toegevoegd, zodat aanbevelingen passen binnen NLQF5, NLQF6 of beide. Deze filters vergroten de relevantie van de uiteindelijke lijst zonder het model zelf complexer te maken.

Daarnaast is veel aandacht besteed aan **uitlegbaarheid**. De methode `explain_recommendation` laat zien welke woorden het meest bijdragen aan een match. Hierbij wordt het element-gewijze product van de tf-idf vector van de student en die van de module gebruikt. De hoogste bijdragen worden weergegeven als concrete “bewijsstukken” voor de kwaliteit van de match.

Belangrijke voordelen van deze uitlegfunctionaliteit zijn:

- studenten begrijpen *waarom* iets wordt aanbevolen;
- het systeem wordt transparanter;
- het helpt fouten te detecteren in de vectorisatie of tekstverwerking.

Evaluatie van de geoptimaliseerde configuratie

De evaluatie werd uitgevoerd met meerdere studentprofielen die inhoudelijk sterk van elkaar verschillen (psychologie, internationale zorg, acute zorg, palliatieve zorg). Voor elk profiel zijn de top-5 aanbevelingen berekend en zijn statistieken zoals gemiddelde similarity, standaarddeviatie en onderlinge overlap geanalyseerd.

De geoptimaliseerde configuratie toont:

- duidelijke verschillen tussen aanbevelingen voor uiteenlopende profielen;
- stabiele rangordes bij herhaalde runs;
- betere spreiding van similarity-scores dankzij bigrams;
- inhoudelijk logische matches, bevestigd door feature-contributions.

Daarnaast werd gekeken naar diversiteit en dekking: de aanbevelingen beslaan een brede set van modules en vertonen weinig ongewenste duplicatie over profielen heen. Hierdoor blijft het systeem flexibel, zelfs wanneer studentprofielen overlappende interesses hebben.

12. Resultaten van Aanbevelingen

In dit hoofdstuk bekijken we hoe het aanbevelingssysteem werkt in de praktijk. We hebben drie fictieve studentprofielen gemaakt in **content_based_recommender.ipynb** en laten zien welke modules het systeem aanbeveelt en waarom. Dit geeft een goed beeld van hoe het systeem in de praktijk zou kunnen helpen bij studiekeuzes.

Student A: Interesse in Psychologie en Coaching

Studentprofiel:

Deze student is geïnteresseerd in psychologie, coaching en zorg. Ze wil graag leren over menselijk gedrag, ontwikkelingspsychologie en hoe ze mensen kan helpen. Ook is ze nieuwsgierig naar gespreksvoering en het begeleiden van anderen.

Top 5 Aanbevolen Modules

1. **Kennismaking met Psychologie (Similarity: 0.2169)**
   Deze module past perfect bij het profiel omdat het direct ingaat op psychologie, menselijk gedrag en gespreksvoering - precies waar de student naar op zoek is.
2. **Organizational Behavior (Similarity: 0.1727)**
   Deze module sluit aan bij de interesse in menselijk gedrag, maar dan in organisatiecontext. Ideaal voor studenten die coaching willen combineren met werken in organisaties.
3. **Veranderen is Mensenwerk (Similarity: 0.1243)**
   Een verdiepingsmodule die goed aansluit bij de interesse in het begeleiden van mensen tijdens veranderprocessen.
4. **Minor Forensisch Onderzoek in de Rechtbank (Similarity: 0.1177)**
   Interessant voor studenten die psychologie willen toepassen in de forensische context, met focus op gedragsonderzoek.
5. **Innovation Management & Creative Leadership (Similarity: 0.1094)**
   Deze module combineert gedragspsychologie met organisatieverandering en leiderschap.

Waarom past 'Kennismaking met Psychologie' het beste?

Het systeem heeft deze module als beste match gekozen omdat de belangrijkste woorden uit het studentprofiel perfect overeenkomen:

- 'Psychologie' komt sterk voor in zowel het profiel als de module (contribution: 0.0913)
- 'Gedrag' is een kernwoord dat beide teksten delen (contribution: 0.0735)
- 'Anderen' (helpen/begeleiden) komt ook sterk naar voren (contribution: 0.0522)

Student B: Internationale Zorg en Verpleegkunde

Studentprofiel:

Deze student wil graag internationaal werken in de zorg. Stage lopen in het buitenland lijkt haar geweldig. Ze is geïnteresseerd in verpleegkunde en wil haar persoonlijke ontwikkeling stimuleren door ervaring op te doen in een internationale context.

Top 5 Aanbevolen Modules

1. **Learning and Working Abroad (Similarity: 0.5850)**
   Dit is een bijna perfecte match! De module gaat over internationaal werken, stage lopen in het buitenland én verpleegkunde - precies wat de student zoekt.
2. **Teaching English Abroad (Similarity: 0.2729)**
   Ook een internationale optie, maar minder specifiek gericht op zorg.
3. **Going Global: Internationaal Perspectief op je Toekomst (Similarity: 0.2221)**
   Interessant voor studenten die hun internationale oriëntatie willen ontwikkelen, met aandacht voor gezondheid en welzijn.
4. **Studeren Buitenland (Similarity: 0.2105)**
   Een bredere module over internationalisering en ervaringen opdoen in het buitenland.
5. **Proactieve Zorgplanning (Similarity: 0.0991)**
   Minder focus op het internationale aspect, maar wel relevant voor verpleegkunde.

Waarom past 'Learning and Working Abroad' het beste?

Deze match is opvallend hoog (0.5850 - de hoogste van alle voorbeelden!). Dat komt omdat bijna elk belangrijk woord uit het studentprofiel terugkomt:

- 'Stage' en 'stage lopen' komen sterk overeen (contribution: 0.0682 en 0.0497)
- 'Internationale' en 'internationale context' passen perfect (contribution: 0.0668 en 0.0497)
- 'Verpleegkunde' wordt expliciet genoemd (contribution: 0.0539)
- 'Buitenland', 'lopen', en 'graag' dragen ook bij (elk 0.0444)

Het systeem zegt eigenlijk: "Deze module is gemaakt voor jou! Alles wat je zoekt - stage, internationaal, buitenland, verpleegkunde - staat hier in." De student heeft heel specifieke wensen beschreven en de module komt daar bijna volledig mee overeen.

Student C: Palliatieve Zorg en Rouw

Studentprofiel:

Deze student wil graag werken met mensen die rouw en verlies ervaren. Palliatieve zorg spreekt hem aan en hij wil leren over oncologie en hoe hij mensen kan begeleiden in moeilijke tijden. Ook vindt hij culturele diversiteit belangrijk.

Top 5 Aanbevolen Modules

1. **Rouw en Verlies (Similarity: 0.2894)**
   De perfecte match: deze module gaat specifiek over rouw, verlies, palliatieve zorg én culturele diversiteit.
2. **Palliatieve Zorg (Similarity: 0.2302)**
   Ook een goede match, volledig gericht op palliatieve zorg.
3. **Zorg Dichtbij (Similarity: 0.1310)**
   Minder specifiek, maar wel relevant voor moderne zorgvormen.
4. **Zorg in de Langdurige Hulpverlening (Similarity: 0.1171)**
   Breed gericht op zorg en hulpverlening bij verschillende problematiek.
5. **Langer Thuis in de Wijk (Similarity: 0.1110)**
   Focus op wijkgericht werken met ouderen en mensen die thuis willen blijven wonen.

Waarom past 'Rouw en Verlies' het beste?

Dit is weer een sterke match omdat de student heel specifieke interesses heeft:

- **'Palliatieve'** en **'palliatieve zorg'** zijn kernwoorden (contribution: 0.0728 en 0.0548)
- **'Oncologie'** wordt expliciet genoemd in beide (contribution: 0.0728)
- **'Diversiteit'** en **'culturele'** komen overeen (elk 0.0233)
- **'Begeleiden'** sluit aan bij het helpen van mensen (contribution: 0.0213)
- **'Zorg'** is het overkoepelende thema (contribution: 0.0210)

Het systeem herkent dat deze student op zoek is naar een zeer specifieke specialisatie. De module 'Rouw en Verlies' behandelt precies deze specialisatie, inclusief de culturele aspecten die de student belangrijk vindt.

Belangrijkste Observaties

Uit deze drie voorbeelden kunnen we een aantal patronen zien:

1. **Specifieke profielen = betere matches**: Student B had het meest specifieke profiel ("stage in het buitenland, verpleegkunde, internationaal") en kreeg de hoogste similarity score (0.5850). Student A had een breder profiel ("psychologie, coaching, zorg") en kreeg lagere scores.

2. **Unieke woorden tellen zwaar**: Woorden zoals 'oncologie', 'palliatieve', en 'stage lopen' zijn specifiek en komen niet vaak voor, waardoor ze zwaar meetellen in de berekening.

3. **Het systeem werkt intuïtief**: De aanbevelingen voelen logisch aan. Als je zegt dat je stage wilt lopen in het buitenland, krijg je modules over internationaal werken. Als je interesse hebt in rouw, krijg je een module over rouw.

4. **Context matters**: Het systeem kijkt niet alleen naar losse woorden, maar ook naar woordcombinaties (bigrams) zoals "stage lopen" en "internationale context". Dit maakt de aanbevelingen nauwkeuriger

13. Algorithmic Affordances & User Control

Het systeem biedt studenten verschillende vormen van controle over hoe aanbevelingen worden gegenereerd. Allereerst kan de student zelf de input bepalen via interesseprofielen, waarden en loopbaandoelen. Daarnaast kan de gebruiker gewichtsfactoren aanpassen, zoals het relatieve belang van interesses, studiedoelen of niveau-aansluiting. Tijdens het aanbevelingsproces beschikt de student over controle over de uitkomst, bijvoorbeeld door aanbevelingen opnieuw te laten genereren, filters toe te passen (niveau, locatie, workload) of sliders te gebruiken voor strengere/soepelere overeenkomsten. Tot slot bevat het systeem elementen van uitlegbaarheid en transparantie, zoals zichtbaar maken welke modulekenmerken bijdragen aan de matchscore en waarom bepaalde modules hoger worden gerankt.

14. Conclusie

In dit project is een content-based aanbevelingssysteem ontwikkeld dat studenten helpt passende verdiepende modules te vinden. De kern van het systeem bestaat uit TF-IDF-vectorisatie van modulebeschrijvingen en cosine similarity, waarmee inhoudelijke gelijkenissen tussen modules en studentprofielen berekend worden. Deze aanpak levert een transparante en uitlegbare methode op die goed aansluit bij het doel van het project: inzicht geven in waarom bepaalde modules worden aanbevolen.

De resultaten laten zien dat het model consistent werkt, overzichtelijke moduleclusters vormt en duidelijke verschillen in similariteitsscores zichtbaar maakt. Hierdoor ontstaat een betrouwbaar fundament dat reproduceerbare aanbevelingen oplevert. Omdat het systeem uitsluitend afhankelijk is van module-inhoud, werkt het bovendien goed voor nieuwe modules waarvoor nog geen gebruikersdata beschikbaar is.

Voor studenten biedt het systeem directe waarde doordat het helpt bij studieoriëntatie, het filteren van keuzes en het vinden van modules die passen bij hun interesses, waarden en loopbaandoelen. Voor Avans draagt het bij aan betere benutting van het onderwijsaanbod, meer transparantie en ondersteuning van modulair onderwijs.

Het ontwikkelde systeem vormt daarmee een solide basis waarop in de toekomst uitgebreid kan worden, bijvoorbeeld met gebruikersfeedback, hybride modellen of verbeterde tekstverwerking. Hiermee kan Avans uiteindelijk een volledig adaptieve en persoonlijke modulekeuze-ervaring realiseren.

15. Bronnen

Dataset: https://brightspace.avans.nl/d2l/le/lessons/251408/topics/1825635
Gastcollege Ethiek: https://brightspace.avans.nl/d2l/le/lessons/251408/topics/1829774
Alogorithmic affordances:
https://brightspace.avans.nl/d2l/le/lessons/251408/topics/1815195
Data Analysis with python:
https://brightspace.avans.nl/d2l/le/lessons/251408/topics/1819785
Dimensionality Reduction:
https://brightspace.avans.nl/d2l/le/lessons/251408/topics/1824492
Model training and evaluation:
https://brightspace.avans.nl/d2l/le/lessons/251408/topics/1829528
NLP: https://brightspace.avans.nl/d2l/le/lessons/251408/topics/1831964
Recommender Systems:
https://brightspace.avans.nl/d2l/le/lessons/251408/topics/1836316
Project github: https://github.com/BoyanKloosterman/AI-Prototype
Data science life cycle: https://www.datascience-pm.com/data-science-life-cycle/ /
https://www.geeksforgeeks.org/data-science/data-science-lifecycle/