ENVIHEI

Titel van de PGO-uitdaging:

Data engineering ontmoet klimaatwetenschap: CMIP6 verkennen via probleemgestuurd onderwijs

Naam van de verantwoordelijke docent

Andrés Navarro Martínez de la Casa

Universiteit / Departement / Opleiding

University of León / Department of Applied Chemistry and Physics / Bachelor's Degree in Data Engineering and Artificial Intelligence (Bacheloropleiding Data Engineering en Artificiële Intelligentie)

Cursus / Vak waarin geïmplementeerd

4e jaar / Projects in Data Engineering and Artificial Intelligence (Projecten in Data Engineering en Artificiële Intelligentie)

Implementatieperiode (datums)

10 september 2025 – 31 oktober 2025

Totale duur (weken/uren)

8 weken / 30 uur

Aantal betrokken studenten

7 studenten

Alle resultaten die binnen ENVIHEI zijn ontwikkeld, zijn beschikbaar onder Creative Commons Naamsvermelding-NietCommercieel-GelijkDelen 4.0 Internationaal.

Medegefinancierd door de Europese Unie. De geuite standpunten en meningen zijn echter uitsluitend die van de auteur(s) en weerspiegelen niet noodzakelijk die van de Europese Unie of van de Spaanse Dienst voor de Internationalisering van het Onderwijs (SEPIE). Noch de Europese Unie noch SEPIE kan hiervoor verantwoordelijk worden gesteld. [2024-1-ES01-KA220-HED-000246733]

Erasmus+ Enriching lives, opening minds. Co-funded by the European Union

P. No.: 2024-1-ES01-KA220-HED-000246733

1. Projectontwerp

1.1 Korte beschrijving van het project en de voorgestelde uitdaging

De cursus is een ervaring met probleemgestuurd onderwijs (PGO), ontworpen voor studenten data engineering. De centrale uitdaging bestaat erin de volledige levenscyclus van data engineering – inname, opschoning, analyse en modellering – toe te passen op reële klimaatproblemen. Studenten zullen gigabytes aan data uit klimaatmodellen (CMIP6) en heranalyses (ERA5) verwerken om complexe duurzaamheidsvragen te beantwoorden, zoals het detecteren van klimaatanomalieën, het classificeren van bioklimatologische zones met behulp van niet-gesuperviseerde machine learning en het evalueren van de onzekerheid van toekomstige klimaatprojecties aan de hand van multimodel-ensembles.

1.2 Verwachte leerresultaten (academische en transversale competenties)

Van de studenten wordt verwacht dat zij vaardigheid aantonen in zowel geavanceerde data engineering als klimaatwetenschappelijke toepassingen. De academische leerresultaten zijn gericht op het ontwerpen en implementeren van geautomatiseerde ETL-pipelines (extractie, transformatie, laden) voor hoogdimensionale geospatiale data. Studenten leren API's voor satelliet- en modeldata (Copernicus CDS en ESGF) te integreren met efficiënte opslagformaten. Daarnaast leren studenten technieken voor niet-gesuperviseerde machine learning toe te passen – met name dimensionaliteitsreductie (PCA) en clustering (K-Means) – om latente spatiotemporele patronen te ontdekken, terwijl zij de voorspellende onzekerheid rigoureus kwantificeren via statistische analyse van multimodel-ensembles (CMIP6).

Op transversaal vlak bevordert het project het vermogen om met onzekerheid om te gaan in besluitvormingsprocessen, een cruciale vaardigheid bij het werken met probabilistische klimaatmodellen. Studenten beheersen ook de standaarden voor reproduceerbare wetenschap door gebruik te maken van versiebeheersystemen en goede praktijken voor documentatie. Tot slot ontwikkelt de cursus sterke vaardigheden in datastorytelling, waardoor studenten complexe, multidimensionale datasets kunnen

synthetiseren tot bruikbare visuele inzichten die uiteenlopende stakeholders informeren.

1.3 Beoogde duurzaamheidscompetenties

Dit project richt zich rechtstreeks op de subdoelen van drie duurzameontwikkelingsdoelen (SDG's), volgens de richtlijnen van het Sustainable Development Solutions Network (SDSN/REDS):

- **SDG 13 - Klimaatactie (hoofddoel)**: Technische opleiding om essentiële klimaatvariabelen te analyseren en de fysische basis van klimaatverandering te begrijpen aan de hand van waargenomen en gemodelleerde data.
- **SDG 9 - Industrie, innovatie en infrastructuur**: Bevordering van wetenschappelijk onderzoek en technologische capaciteit door het gebruik van 'big data' en artificiële intelligentie toegepast op duurzaamheid.
- **SDG 4 - Kwaliteitsonderwijs**: Verwerving van geavanceerde digitale vaardigheden en bevordering van publieke toegang tot klimaatinformatie (open data/open science).

1.4 Duurzaamheidsbenadering en de manier waarop deze in het project is geïntegreerd

Duurzaamheid is geen toevoeging, maar vormt de kern van de leerstof van de cursus. Elke technische module is gekoppeld aan een duurzaamheidsvraagstuk: het berekenen van klimaatnormalen voor klimaatadaptatie en veerkrachtige infrastructuur (ASS01), het gebruik van ensembles om onzekerheid in voorspellingen te beheersen (ASS02) en het herdefiniëren van klimaatzones onder scenario's van mondiale verandering (ASS04). Deze verbanden dwingen studenten om de fysica van klimaatverandering te begrijpen, zodat zij effectieve dataoplossingen kunnen ontwikkelen.

1.5 Gebruikte methodologieën en tools (PGO-structuur, facilitatietechnieken, enz.)

De sturende uitdaging. Het project is verankerd in een centrale open vraag die het volledige leerproces richting geeft: Hoe kunnen we ruwe klimaatdata op petabyteschaal vertalen naar bruikbare inzichten voor duurzaamheid in een veranderende wereld?

Studenten treden op als data-engineers die de opdracht krijgen het klimaatsysteem te 'decoderen' om deze uitdaging aan te pakken.

PGO-structuur & scaffolding. De methodologie volgt een progressieve scaffoldingaanpak, van begeleid onderzoek naar autonoom probleemoplossen, in drie fasen (funderings-, onderzoeks- en oplossingsfase):

- **Funderingsfase (De toolkit)**. Studenten maken kennis met het ecosysteem (Linux/Conda) en leren de data-inname te automatiseren met behulp van API's. De nadruk ligt op technische competentie.
- **Onderzoeksfase (De fysica van data)**. Studenten pakken specifieke deelproblemen aan (normalen, anomalieën en ensembles), waarvoor inzicht in de onderliggende fysische concepten nodig is om de juiste code te schrijven. Deze fase verbindt statistiek met duurzaamheid.
- **Oplossingsfase (Het sluitstuk)**. Het afsluitende project 'Climate Classification System (CCS)' haalt de zijwieltjes eraf. Studenten moeten alle voorgaande modules (het ophalen van data, opschoning, statistische analyse en machine learning) integreren om een nieuwe oplossing voor de sturende uitdaging voor te stellen.

Facilitatietechnieken:

- **Just-in-time-onderwijs (JiTT)**: JiTT introduceert theoretische concepten (bijv. ensemble, anomalie, PCA) op het moment dat studenten ze nodig hebben om een programmeerprobleem op te lossen, wat de relevantie en het onthouden maximaliseert.
- **Coding dojo's**: Praktijksessies waarin de docent live programmeert om debugstrategieën en goede praktijken te demonstreren, die studenten vervolgens nabootsen en uitbreiden.
- **Peer-codereviews (in-situ)**: In plaats van traditionele beoordeling krijgen studenten feedback via 'over-de-schouder'-codereviews.

Tools & ecosysteem. De technische implementatie steunde op een open-sourcestack volgens de industriestandaard.

- **Omgeving**: Linux/WSL met Conda voor pakketbeheer.
- **Kernbibliotheken**: Python (Xarray, Pandas, Matplotlib en Cartopy) voor datamanipulatie en -visualisatie.
- **Klimaatspecifieke tools**: CDO (Climate Data Operators) voor krachtige NetCDF-verwerking en GDAL voor geospatiale rasterbewerkingen.
- **Toegang tot data**: Geautomatiseerd ophalen via de Copernicus CDS API en ESGF-nodes.
- **Samenwerking**: Git en GitHub worden gebruikt voor versiebeheer en het indienen van projecten.

1.6 Organisatie van het project (groepsvorming, tijdlijn, mijlpalen)

8 weken (30 contacturen + zelfstandig werk):

- **Week 1**: Basisprincipes, tools en data-acquisitie (API's).
- **Weken 2-4**: Klimaatstatistiek (normalen, anomalieën en ensembles). Individuele opdrachten.
- **Weken 5-7**: Geavanceerde machine learning (CCS-clustering). Studenten werkten in kleine groepen, waarbij een professionele ontwikkelomgeving werd nagebootst.
- **Week 8**: Eindintegratie en mondelinge presentaties.

1.7 Van studenten gevraagde opleveringen (type, formaat, deadlines)

Voor elke module (ASS01-ASS04) werd het volgende gevraagd:

- **Code**: Nette en gedocumenteerde Python-scripts of notebooks in de repository en/of in Moodle.
- **Technisch rapport**: Pdf-document met methodologie, visualisaties en bespreking van de resultaten.
- **Visualisaties**: Kaarten of grafieken gegenereerd met Cartopy/Matplotlib voor elke opdracht.
- **Eindpresentatie**: Mondelinge verdediging van het CCS-project. Alleen voor het samenwerkingsproject (ASS04).

Deadlines: Eén oplevering per week (zeven dagen na de sessie), behalve voor ASS04, een samenwerkingsproject van vier weken.

1.8 Betrokken externe partners (mentoren, bedrijven, experts…) en hun rol

Bij deze implementatie waren geen externe partners betrokken.

2. Projectevaluatie

2.1 Gebruikte beoordelingscriteria en -methoden

De evaluatie verliep continu en was gebaseerd op specifieke rubrics voor elke opdracht, met doorgaans de volgende weging:

- **Technische correctheid en data-analyse (40%)**: Nauwkeurigheid van berekeningen (bijv. anomalieën, seizoensgemiddelden).
- **Codekwaliteit en reproduceerbaarheid (20%)**: Structuur, commentaar en efficiënt gebruik van bibliotheken.
- **Kwaliteit van de visualisaties (20%)**: Communicatieve helderheid van kaarten en grafieken.
- **Rapport en communicatie (20%)**: Vermogen om wiskundige resultaten te synthetiseren en te interpreteren.

Details over de specifieke rubrics zijn te vinden in de GitHub-repository: https://github.com/navarro-esm/ULE_data-ai-projects/tree/main

2.2 Bewijs van het leren van studenten (voorbeelden, output, resultaten)

De eindresultaten tonen een hoog vaardigheidsniveau in het toepassen van data-engineeringtechnieken op klimaatwetenschap. Studenten maakten met succes de overstap van eenvoudige scripts naar complexe, reproduceerbare workflows met dimensionaliteitsreductie (PCA) en niet-gesuperviseerde clustering (K-Means).

Hieronder staan directe links naar de projecten die de studenten hebben ontwikkeld en die als concreet bewijs van de leerresultaten dienen:

- **Groep A (Agricultural Vulnerability)**: https://github.com/gortif00/CMIP6_Climate_Classification
- **Groep B (Solar Farm):** https://github.com/calejs00/pid
- **Groep C (Bear distribution):** https://github.com/SergioAlvesMar5/CMIP6_Climate_Classification_IvanFernandez_SergioAlves/

3. Reflecties en aanbevelingen

3.1 Belangrijkste resultaten en sterke punten van het project

Het belangrijkste resultaat is de geslaagde integratie van 'harde' data engineering met duurzaamheidswetenschap. Studenten leerden het domeinoverschrijdende karakter van klimaat te herkennen: zij zagen datasets niet langer louter als matrices van getallen en begrepen hun sociale en ecologische impact. Studenten ontwikkelden een proactieve houding, waarbij zij gestelde problemen oplosten en nieuwe projecten voorstelden die engineering en klimaat met elkaar verbinden, en zo blijk gaven van een diepgaand begrip van de leerstof.

3.2 Ondervonden uitdagingen en hoe deze werden aangepakt

De ondervonden uitdagingen waren typisch voor een nieuw ingevoerd vak, vooral omdat dit de eerste lichting studenten is die het vierde jaar van de opleiding bereikt. Als pilotprogramma met een volledig nieuwe groep studenten vereiste de cursus realtime aanpassingen van de technische workflows en het didactische tempo, om ervoor te zorgen dat de leerdoelen werden bereikt zonder de studenten te overbelasten. Concreet werden twee belangrijke technische obstakels aangepakt:

- **Heterogeniteit van omgevingen**: Het beheren van softwareomgevingen (Conda) over verschillende besturingssystemen heen (Windows, Mac, Linux) was complex. Cruciale tools zoals CDO hadden compatibiliteitsproblemen op Windows, maar deze werden opgelost door studenten te begeleiden bij het gebruik van Linux-subsystemen (WSL) of virtuele machines.

- **Stabiliteit van de datatoegang**: Externe repositories (Copernicus, ESGF) kampten af en toe met hoge latentie of onbeschikbaarheid. Om dit probleem te beperken, werden de deadlines flexibel gehouden, en lokale back-updatasets (caches) om elke blokkering van de leervoortgang van de studenten te voorkomen.

3.3 Lessen voor toekomstige implementaties

Een van de belangrijkste lessen uit deze pilotimplementatie is het belang van onmiddellijke betrokkenheid om aanvankelijke scepsis te overwinnen. Hoewel de studenten aanvankelijk aarzelden, veranderde hun perceptie aanzienlijk zodra zij met echte data begonnen te werken. Daarom zullen toekomstige edities vanaf de eerste sessie impactvolle use cases introduceren, zoals recente extreme weersgebeurtenissen, om de interesse van de studenten meteen te wekken. Bovendien benadrukte de feedback van studenten de behoefte aan meer methodologische flexibiliteit. Toekomstige iteraties zullen studenten meer vrijheid geven bij het kiezen van statistisch-wiskundige technieken om problemen op te lossen, waardoor creatievere en minder gestuurde benaderingen worden aangemoedigd zodra de basisvaardigheden onder de knie zijn.

4. Ondersteunend materiaal

4.1 Materiaal van de docenten en projectdocumentatie

Het materiaal van de docenten en de projectdocumentatie zijn beschikbaar via de volgende link:
https://drive.google.com/drive/folders/17N07p7YOHd7lprLZGN759cnp5EkDBrMi?usp=drive_link

4.2 Door studenten geproduceerde opleveringen, eindpresentaties of pitchmateriaal

Een voorbeeld van door studenten geproduceerde opleveringen is beschikbaar via de volgende link:

https://drive.google.com/drive/folders/1E7oaA-tSpN8NlM2xiN-W1QScFyMKex4I?usp=drive_link

De eindpresentaties van de studenten zijn beschikbaar via de volgende link: https://drive.google.com/drive/folders/1eSmrix7Wk8tS65pVgI1bX6tDfG6RXmAL?usp=drive_link

4.3 Afbeeldingen/video's (met toestemming)

Foto- en video-opnamen werden gemaakt met de uitdrukkelijke toestemming van de deelnemende studenten en zijn beschikbaar via de volgende link: https://drive.google.com/drive/folders/1S7-AQBQStS3EH22-l_-x_ZpvM12czDyx?usp=drive_link

4.4 Overig relevant materiaal

Al het materiaal, inclusief dit rapport, is beschikbaar via de volgende link: https://drive.google.com/drive/folders/1lOJfNdDqIVrzm_YZyGSwbdy6fbT08z4m?usp=drive_link