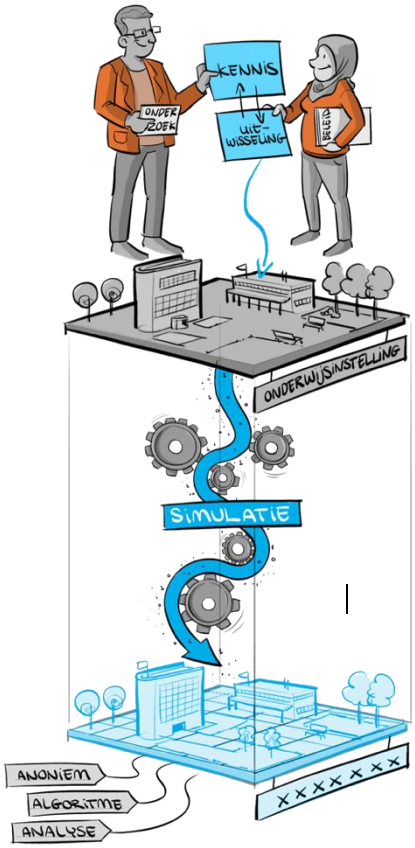
Versnellingsplan
Onderwijsinnovatie
met ICT

veilig en betrouwbaar
benutten van studiedata

Simulatiedataset



Wat is de simulatiedataset?

De simulatiedataset is een synthetische dataset die is gebaseerd op tien jaar aan echte studentdata van een universiteit. De simulatie is zo uitgevoerd dat deze data op geen enkele manier meer terug te herleiden is tot de oorspronkelijke gegevens en vormt daarmee een fictieve universiteit.
De simulatiedataset kun je gebruiken om analyses of algoritmes buiten je eigen instelling te delen. Ook kun je de simulatiedataset gebruiken om mee te oefenen of algoritmes te testen.
We ontwikkelen drie simulatiedatasets. De eerste is een simulatie van instellingsdata van een universiteit, de tweede is een simulatie van instellingsdata van een hogeschool en de derde is een simulatie van data uit leermanagementsystemen. De eerste set is nu klaar.

Welk probleem lost het op?

Landelijk doen hoger onderwijsinstellingen steeds meer om studiedata te benutten. Wanneer een data-analist een analyse of algoritme wil delen met een vakgenoot van een andere onderwijsinstelling, loopt de analist tegen een barrière aan. Het is namelijk niet mogelijk om analyses buiten de eigen instelling te delen zonder hiervoor de studiedata van de eigen instelling te gebruiken, maar dit is vanwege privacyreglementen niet toegestaan.

De originele karakteristieken, onderlinge correlaties en patronen uit de data blijven behouden in het simulatie-resultaat. De synthetische data is enkel gebaseerd op een kansverdeling van de unieke variabelen en bevat dus alleen synthetische gegevens. De simulatiedataset bevat geen persoonsgegevens van bestaande of oud-studenten. Hierdoor kan de data-analist deze gebruiken om alsnog kennis te delen met externe collega's.

Wat kan ik met de simulatiedataset?

De simulatiedataset download je gratis op de website van het Versnellingsplan. Het bestaat uit een .csv bestand met de studiedata, een R script om simulatiedata te reproduceren en een R package om zelf simulatiedata te genereren.

Uit de simulatiedataset ontleen je dezelfde statistische resultaten als uit de originele data. Je kunt deze sets dus gebruiken om mee te oefenen of testen, of om analyses extern te delen. Je kunt er ook voor kiezen om zelf een simulatiedataset te genereren met het R package die gebaseerd is op data van jouw eigen onderwijsinstelling.

Wat zijn de reacties?

"De EUR heeft een samenwerkingsverband met het Erasmus Medisch Centrum en met de TU Delft, gericht op samenwerking in onderzoek met betrekking tot gezondheidsgegevens, zoals bijvoorbeeld COVID-19 data. Momenteel leggen we vast hoe alle betrokken partijen - niet in de laatste plaats burgers zelfvertrouwen kunnen hebben in de doelen van onze samenwerking en in de (ethische) waarborgen die we inbouwen in de manier van samenwerken. Rondom zorgdata en onderzoeksdata heb je allerlei, terechte, juridische barrières. Een van de oplossingen om hier goed gehoor aan te geven, is gebruik te maken van synthetische data."

Marlon Domingus, functionaris gegevensbescherming bij de Erasmus Universiteit Rotterdam.

Waar kan ik meer informatie vinden?

Versnellingsplan.nl

Hier vind je de simulatiedataset zelf en verdere toelichting op het project.