

Datawetenschap (I002440)

Cursusomvang (nominale waarden; effectieve waarden kunnen verschillen per opleiding)

Studiepunten 5.0

Studietijd 150 u

Aanbodsessies en werkvormen in academiejaar 2024-2025

A (semester 2)

Nederlands

Gent

groepswerk

werkcollege

hoorcollege

Lesgevers in academiejaar 2024-2025

Verwaeren, Jan

LA26

Verantwoordelijk lesgever

Meys, Joris

LA26

Medewerker

Aangeboden in onderstaande opleidingen in 2024-2025

[Bachelor of Science in de bio-ingenieurswetenschappen](#)

stptn

aanbodsessie

5

A

[Master of Science in Bioinformatics \(afstudeerrichting Bioscience Engineering\)](#)

5

A

Onderwijstalen

Nederlands

Trefwoorden

Data-wrangling, data-exploratie, datavisualisatie, datacommunicatie, R, exploratieve hoogdimensionale data-analyse.

Situering

Dit opleidingsonderdeel is een inleidende cursus tot *data science* waarin studenten theoretische kennis wordt bijgebracht over data-analyse, maar ook sterk wordt ingezet op het verwerven van praktische vaardigheden voor het analyseren van data in de programmeertaal R. Het uitgangspunt van deze cursus is dat data in allerlei maten en vormen kunnen voorkomen, gaande van zelf uitgevoerde, gestructureerde observaties of metingen in een eigen onderzoek, tot online bronnen die minder gestructureerd of betrouwbaar zijn, ontbrekende waarden kunnen bevatten, etc. Binnen deze cursus worden de inzichten en tools aangereikt om op een betrouwbare, reproduceerbare en transparante manier informatie te extraheren uit deze data. In een eerste, eerder beschrijvend deel van de cursus maken de studenten kennis met de levenscyclus van data in een wetenschappelijke context, de opbouw en exploratie van laagdimensionale datatables met bijzondere aandacht voor visualisatietechnieken, data-cleaning en data-wrangling. In een tweede, meer wiskundig/technisch deel wordt verder ingegaan op het gebruik van resamplingtechnieken, methoden voor exploratie van hoogdimensionale datasets en (eenvoudige) predictieve modellering. Ten slotte wordt het gebruik van relationele databanken in de context van datawetenschap ingeleid en wordt in de context van data-bewaring aandacht besteed aan ontwikkelingen inzake de bescherming van persoonsgegevens.

Inhoud

De leerinhouden worden onderverdeeld in vijf delen. Elk deel bevat een combinatie van theoretische inzichten en kadering, aangevuld door hands-on PC-labs in R.

DEEL 1: Inleiding tot datawetenschap en het gebruik van R

- Inleiding tot data en de data-analyse cyclus
- Introductie tot R (belangrijkste datatypes i.f.v. data-analyse, controlestructuren, built-in en user-defined functies, vectorisatie, data-intake)

DEEL 2: Univariate en bivariate exploratieve data-analyse

- Datatables: meetschalen, datacollectieprincipes en -terminologie, data-opslag

(Goedgekeurd)

en datatoegang

- Univariate exploratieve analyse: frequentieverdelingen, steekproefpercentielen, centrum- en spreidingsmaten, standaardiseren, resamplingtechnieken (resampling uit populatie en bootstrapping), omgaan met missing values
- Bivariate exploratieve analyse: bivariate frequentieverdelingen, enkelvoudige lineaire regressie, R^2 , lokale lineaire regressie en scatterplot smoothers, Spearman correlatie, Pearson correlatie, Cramér's V, bivariate outliers

DEEL 3: Datamanipulatie en -visualisatie

- Principes van reproduceerbare en transparante data-analyse en communicatie
- Datamanipulatie (voor data-preparation): selecteren, filteren, aggregeren, join-operaties, wide en long formaten, tidy data. De implementatie van deze principes in tidyverse wordt gebruikt bij de practica.
- Datavisualisatie: eigenschappen van goede data-diagrammen, de layered grammar of graphics. De implementatie van deze principes in ggplot2 wordt gebruikt bij de practica.

DEEL 4: Modelgebaseerde en multivariate data-analyse

- Curve fitting: verliesfuncties, methode van Newton en gradient descent voor het minimaliseren van verliesfuncties, de algoritmen van Gauss-Newton en Levenberg-Marquardt
- Inleiding tot dimensionaliteitsreductietechnieken
- Inleiding tot predictieve modellering: k nearest neighbours, performantie-evaluatie en kennismaking met tree-based methods

DEEL 5: Relationale databanken en persoonsgegevens

In functie van de opleiding wordt een beknopte inleiding gegeven tot het gebruik van relationele databanken en SQL. In de ruimere context van data-bewaring wordt aandacht gegeven aan recente ontwikkelingen inzake bescherming van persoonsgegevens.

Begincompetenties

Datawetenschap bouwt verder op bepaalde eindcompetenties van opleidingsonderdelen 'Calculus', 'Lineaire algebra' en 'Wetenschappelijk programmeren'; of de eindcompetenties werden op een andere manier verworven.

Eindcompetenties

- 1 Verschillende types en vormen waarin data voorkomen herkennen.
- 2 R kunnen gebruiken, op een intermediate level, als een programmeeromgeving voor data-analyse.
- 3 Data die beschikbaar zijn in courante text-based dataformaten kunnen inlezen, combineren en omzetten naar een vorm die verdere analyse toelaat (data intake en data-wrangling).
- 4 Een exploratieve univariate en bivariate exploratieve analyse uitvoeren (inclusief data-visualisatie).
- 5 Overzichtelijke samenvattende figuren en tabellen, van een relevante opmaak voorzien, genereren die voldoen aan de eisen voor wetenschappelijke communicatie.
- 6 Een gepaste verliesfunctie en optimalisatie-algoritme kiezen voor het oplossen van curve-fitting problemen.
- 7 Dimensionaliteitstechnieken toepassen om inzicht te krijgen in hoog-dimensionele datasets.
- 8 Selecteren van een geschikte predictieve modelleringstechniek voor het oplossen van een voorspellingsprobleem.
- 9 Een eenvoudige relationele databank kunnen bevragen.

Creditcontractvoorwaarde

Toelating tot dit opleidingsonderdeel via creditcontract is mogelijk na gunstige beoordeling van de competenties

Examencontractvoorwaarde

Dit opleidingsonderdeel kan niet via examencontract gevolgd worden

Didactische werkvormen

Groepswerk, Werkcollege, Hoorcollege

Toelichtingen bij de didactische werkvormen

Tijdens de hoorcolleges worden de fundamentele theoretische concepten aangebracht. De PC-klasoefeningen bestaan uit een 10-tal hands-on practica. De

(Goedgekeurd)

studenten werken in het groepswerk samen aan een overkoepelende data-analyse opdracht.

Studiemateriaal

Type: Syllabus

Naam: Syllabus datawetenschap

Richtprijs: Gratis of betaald door opleiding

Optioneel: nee

Taal : Nederlands

Aantal pagina's : 300

Oudst bruikbare editie : 2024

Beschikbaar op Ufora : Ja

Online beschikbaar : Ja

Beschikbaar in de bibliotheek : Nee

Beschikbaar via studentenvereniging : Nee

Referenties

Benjamin S. Baumer, Daniel T. Kaplan and Nicholas J. Horton. (2017) Modern Data Science with R. CRC Press, 578p.

John L. Faundeen, Thomas E. Burley, et al. (2013) The United States Geological Survey Science Data Lifecycle Model, USGS Open-File Report 2013-1265.

Vakinhoudelijke studiebegeleiding

De lesgever beantwoordt vragen i.v.m. de theorie na afspraak of in de marge van de hoorcolleges, de assistenten beantwoorden vragen i.v.m. de oefeningen en practica en er is ondersteuning via Ufora.

Evaluatiemomenten

periodegebonden en niet-periodegebonden evaluatie

Evaluatievormen bij periodegebonden evaluatie in de eerste examenperiode

Schriftelijke evaluatie met open vragen

Evaluatievormen bij periodegebonden evaluatie in de tweede examenperiode

Schriftelijke evaluatie met open vragen

Evaluatievormen bij niet-periodegebonden evaluatie

Werkstuk

Tweede examenkans in geval van niet-periodegebonden evaluatie

Examen in de tweede examenperiode is mogelijk

Toelichtingen bij de evaluatievormen

Theoretische kennis (25%) wordt geëvalueerd tijdens het geslotenboekexamen, praktische kennis m.b.t. data-analyse wordt geëvalueerd tijdens het openboekexamen op PC (65%). Het verslag dat wordt ingediend wordt beoordeeld door de lesgever en telt mee voor 10% van het eindresultaat.

Studenten krijgen in de tweede zitting de kans om de taak uit de eerste zitting te herwerken, op basis van de verkregen feedback. Op basis van de herwerkte versie wordt een nieuwe score toegekend.

Eindscoreberekening

De eindscore wordt bepaald op basis van drie onderdelen (aandeel in eindscore vermeld tussen haakjes): gesloten boek theorie-examen (25%), open boek PC-examen (65%) en verslag (10%)