Leerstofoverzicht examen 20/06/2023

In essentie komt het neer op:

- Het boek practical statistics for data scientists tot en met pagina 68 &
- Het slide-deck *05 lineaire regressie en wiskunde in ML workflow.*

De inhoud van de evaluatie is een mondeling gesprek over (een selectie van) de onderwerpen uit deze gedeeltes waarbij de student verwacht wordt van te **begrijpen** wat er speelt qua statistiek in de praktijk. Het gesprek gaat uit van een echte case waarbij we dan moeten bepalen welke statistische methodes er toegepast moeten worden om zinvolle resultaten te bekomen.

Overzicht van de te kennen onderdelen:

- Soorten data:
  - Idee hebben over verschillende bronnen van data.
  - Begrijpen welke soorten gestructureerde data er bestaan en uitleggen wat ze definieert.
- Middelpunten/centrale locaties schatten:
  - Gegeven een situatie kunnen aangeven welke centrummaat het meeste “zin” maakt.
  - Definitie mediaan/gemiddelde (alsook de gewogen & trimmed varianten) kunnen uitleggen.
  - Gegeven een grafische voorstelling van een verdeling (op basis van pdf/cdf/ccdf/bar chart/…) kunnen aanduiden waar de centrummaten ongeveer liggen.
- Variabiliteit schatten:
  - Weten wat de L_p norm te maken heeft met variabiliteit, welke maten hieruit ontstaan en effect van de keuze van p kunnen uitleggen.
  - Weten wat een percentiel is & hoe dat gebruikt kan worden om variabiliteit te duiden.
- Numerieke data verdelingen verkennen
  - De 5 usual suspects van verdelingen in het echte leven kennen en kunnen uitleggen hoe deze in het echte leven kunnen ontstaan. (In de zin van: ik geef een histogram → leg uit hoe dit is kunnen ontstaan).
  - Weten hoe een discrete verdeling gedefinieerd is (in de zin van discrete kansen) en hoe deze visueel voorgesteld kunnen worden.
  - Uitleggen wat een pdf/cdf/ccdf is en hoe je die grafisch kan voorstellen.
  - Boxplots, violin plots & histogrammen weten wat ze zijn en ze kunnen lezen.
  - Gegeven een scatterplot kunnen aangeven wat de correlatie is. Toelichten waarom correlatie belangrijk is voor data science (maar niet allesomvattend).
  - Gegeven een situatie kunnen inzien dat Bayes theorema van tel is & kunnen uitleggen hoe het hier toegepast kan worden.
- Lineaire regressie:

    - De typische ML workflow kunnen toelichten op basis van lineaire regressie. Ihb het belang van volgende zaken kunnen toelichten:
        - Loss functie.
        - Gradient descent.
        - Rapporteren van performance.
- Random sampling
    - Wat is sampling met/zonder replacement?
    - Sampling bias toelichten & herkennen.
    - Verschil sample & population mean toelichten.
    - In woorden gegeven een situatie kunnen uitleggen wat de centrale limietstelling wil zeggen en wat de voorwaarden zijn om ze te kunnen toepassen.
    - Bootstrapping kunnen uitleggen.
    - Weten hoe je zowel bootstrapping als de centrale limietstelling kan gebruiken om confidence intervals te berekenen.

De evaluatie is **fysiek in Mechelen**. We proberen ons zo goed mogelijk aan volgend tijdsschema te houden:

Tijdsslot	Persoon
17:45 – 19:00	Youssef Chatar
18:00 – 19:15	Stijn Bruggen
18:15 – 18:30	Sam Corbeel
18:30 – 18:45	Roel Helgers
18:45 – 19:00	Raf Ledeganck
19:00 – 19:15	Michael Brunclair
19:15 – 19:30	Marie Perin
19:30 – 19:45	Lieven Stassen
19:45 – 20:00	Kris Leyssens
20:00 – 20:15	Kay Warrie
20:15 – 20:30	Jan Dierckx
20:30 – 20:45	Ine Van Wassenhove
20:45 – 21:00	Cedric Engelen
21:00 – 21:15	Caroline Geukens
21:15 – 21:30	Yuri Fiten
21:30 – 21:45	Kevin Francus
21:45 – 22:00	Tom Teck
22:00 – 22:15	Hannes Dockx