**Wat is Data Science?** (o.b.v. de assignments)

Bij data science gebruik je wetenschappelijke methodes, processen en algoritmes om de informatie en inzichten te extraheren die een data scientist nodig heeft. Je voegt data en statistieken samen om een bepaalde fenomeen te verklaren. Je gebruikt bij data science immers allerlei technieken, van wiskundige-, computer science tot informatie science technieken. Je past deze technieken en theorieën toe om een bepaald fenomeen of relatie te verklaren.

Dit is bijvoorbeeld bij assignment 7 en 8 gedaan met boxplots en standaardafwijkingen (wiskundige technieken en theorieën). We hebben gezien dat hoe minder data je gebruikt (alleen 10% van je rijen), hoe lager de confidence interval werd.

Verder hebben we (cor)relaties bekeken tussen verschillende variabelen bij assigment 9 en 10. Betekent dat een zwaarder gewicht van een pinguïn dat de flipper lengte ook langer is? Ja, want we zagen een positieve correlatie.

Dit deden we telkens met numerieke- én categorische data. Bij data science werk je namelijk niet alleen met getallen, maar met alle variabelen die er maar op de wereld zijn. Bijvoorbeeld bij assignment 11 en 12 keken we of een numerieke kolom impact heeft op een categorische kolom. Een voorbeeld was: de mannetjes pinguïns (categorisch: geslacht) is gemiddeld zwaarder (numeriek: gewicht) dan een vrouwtjes pinguïn. Het is dan een taak van een data scientist om hier iets mee te doen. Als de confidence interval groter is dan 95%, dan kan een data scientist al zeggen dat het bewezen is.

Daarnaast ga je bij data science theorieën voorspellen en op basis daarvan het kunnen bewijzen (inductief redeneren). Dit was bijvoorbeeld al gedaan met de confidence interval, maar dat kan ook op andere manieren, zoals een decision tree. In assignment 15 t/m 18 hebben we gewerkt met decision trees. We proberen telkens een numerieke- of categorische kolom te voorspellen op basis van de andere kolommen. Dit kan een oneindige if / else structuur worden.

Bij de eerste vier assignments, hebben we specifieke data gehaald uit een dataset. We wilden bijvoorbeeld weten hoeveel mannetjes pinguïns er zijn in een bepaalde dataset. Met data science haal je data uit data, wat bruikbaar is voor de gegeven situatie. Deze maak je vaak overzichtelijk, leesbaar en met een beetje opmaak, zodat het voor iedereen in één opslag duidelijk is wat er staat.

**Wat doet een data scientist?** (o.b.v. de assignments)

Een data scientist maakt van data een inzichtelijke report waar je wat mee kan. Ze extraheren de data die ze nodig hebben zodat zij bruikbare inzichten hebben voor verschillende doeleinden, zoals voorspellen wanneer een klant een bedrijf verlaat. Wat zijn namelijk de gewoontes van klanten vlák voordat zij vertrekken? Deze voorspellingen heb ik gedaan in assignments 15 t/m 18. Na hoeveel maanden inactiviteit heb je nog een bestaande klant? Het resultaat is dat het grotendeel van bestaande klanten nooit langer dan een halve maand inactief zijn.

Verder probeert een data scientist bepaalde theorieën te bewijzen voor verschillende doeleinden. Met genoeg data (als de zekerheid erg hoog is), dan kan je wel zeggen dat iets bewezen is. Dat was zoals met de dobbelsteen theorie. Om zeker te zijn van dat iemand een valse dobbelsteen heeft (gooit alleen x of hoger), dan moest de confidence level 95% of hoger zijn.

Ook zoekt een data scientist naar verbanden en correlaties. Zijn er op bepaalde eilanden meer vrouwtjes of mannetjes pinguïns? Dit kan ook in de consumentenwereld toegepast worden. Is er voor een bepaalde product meer mannen- of vrouwen liefhebbers? Dan kan je die product aanbevelen aan het desbetreffende geslacht (geslacht is meestal verplicht bij het aanmaken van een account, dus die informatie hebben webshops wel).

Daarnaast herkennen data scientists een patroon in ‘big data’ om zo een voorspellend algoritme te ontwikkelen. Dit wordt ook wel ‘machine learning’ genoemd. In verschillende assignments (zoals assignment 19 met clusters) hebben we voorspellingen gemaakt, zodat we de volgende set aan data gemakkelijk in een van de ‘boxjes’ kunnen zetten.

Data scientists rapporteren en delen verder ook de gegenereerde inzichten en kennissen met de mensen die er behoefte aan hebben. Ik heb af en toe wel samen gekeken met andere klasgenoten om onze uitkomsten te vergelijken. In het bedrijfsleven is dit natuurlijk dan met het management of andere betrokkenen.

Bij de eerste vier jupyter notebook assignments heb je een voorbeeld van informatie extraheren uit data (univariate analysis). We wilden bijvoorbeeld zien wat de verdeelsleutel is van de geslachten bij de pinguïns. Met barplots, pie charts en linecharts kunnen we die verdeelsleutel heel overzichtelijk zien.

Dit geld hetzelfde voor de bivariate analysis. Doordat je meerdere kolommen hebt, zie je eerder een correlatie tussen de kolommen. In assignment 9 en 10 heb je bijvoorbeeld negatieve- en positieve correlaties. Hiermee kan je mogelijk achter de reden van de correlatie komen.

Bij de multivariate assignments (15 t/m 19) voorspelde ik een numerieke- of categorische kolom op basis van de andere karakteristieken. Dit is ook een taak van een data scientist. We kijken namelijk niet alleen naar één dataset, maar naar verschillende varianten van data. We hebben de pinguïn dataset, film dataset en een bank churning dataset.

Als laatst kan een data scientist een besluit maken of tenminste advies geven op basis van de voorspellingen of ontdekte inzichten.