# Opdracht 101

Data science is een interdisciplinair vakgebied dat zich bezighoudt met het extraheren van inzichten, kennis en informatie uit gegevens, vaak met behulp van geavanceerde technieken uit de wiskunde, statistiek, informatica en machine learning. Data science omvat het verzamelen, opschonen, analyseren en visualiseren van gegevens om patronen en trends te identificeren, en het ontwikkelen van modellen en algoritmen om voorspellingen en aanbevelingen te doen op basis van die gegevens. Het doel van data science is om waardevolle inzichten te leveren die bedrijven en organisaties kunnen gebruiken om betere beslissingen te nemen en hun prestaties te verbeteren.

In onder andere opdracht 13 heb ik een analyse van mijn IMDB-dataset uitgevoerd. Deze opdracht doelde vooral op het vinden van verbanden in gegevens. Aan het begin van de analyse creëerde ik een zogenoemde ‘contingency table’ om ratio’s te kunnen weergeven tussen de ‘certificate’ en de ‘genre’ kolommen en dus eventueel verbanden te kunnen weergeven. Om nog duidelijker eventuele verbanden te tonen heb ik er ook een staafdiagram bij gemaakt. Tot slot heb ik ook nog een ‘chi-squared test’ uitgevoerd om de kans op correlatie of samenhang te berekenen. Hoe hoger dit percentage zou zijn hoe meer kans op een correlatie en dus een verband tussen de twee kolommen. Het resultaat was uiteindelijk ook een zeer hoog percentage wat zeker duidt op een verband tussen de twee kolommen. Ik heb dus uiteindelijk dus door te analyseren proberen verbanden tussen gegevens te vinden. Dit soort verbandenanalyses kan ook superbelangrijk zijn voor organisaties en bedrijven, omdat zij door het analyseren van gegevens en eventuele verbanden te vinden hun beleid en beslissingen op de opgedane kennis aan kunnen passen om uiteindelijk betere besluitvorming te realiseren. De informatie verkregen van een analyse kan dus zeker nuttig zijn voor een bedrijf, omdat zij hiermee nieuwe inzichten opdoen en met deze inzichten bepaalde aanbevelingen of voorspellingen kunnen doen. Kortom was deze opdracht zeker gelinkt aan data science, omdat ik gegevens analyseerde en op de nieuwe opgedane informatie conclusies kon trekken.

Daarnaast heb ik het ‘Explorative Analysis’ traject uitgevoerd. De voornaamste onderwerpen die we hebben behandeld binnen de overkoepelende term ‘Data Science’ zijn ‘Univariate analysis’, ‘Bivariate Analysis’ en ‘Multivariate analysis’. Deze termen ben ik tegengekomen in verscheidene opdrachten en ze hebben allemaal een sterk verband met data science. Dit verband wordt uitgebreider beschreven in de volgende alinea’s.

Univariate analyse is een methode die wordt gebruikt om één variabele per keer te analyseren. Het is een analyse dat tot doel heeft de kenmerken van een enkele variabele te samenvatten en te beschrijven. In opdracht 7 en 8 heb ik mij vooral beziggehouden met het uitvoeren van univariate analyses op de pinguïns en mijn gekozen IMDB-datasets. Dit deed ik door zogenoemde ‘distributions’ te maken en te analyseren om zo weer meer inzicht te krijgen in de datasets en eventuele verbanden te vinden.

Bivariate analyse is een methode waarbij de relatie tussen twee variabelen wordt geanalyseerd. Het is een analyse dat tot doel heeft om te onderzoeken of er een verband bestaat tussen twee variabelen en, zo ja, hoe sterk dit verband is. In opdracht 9 t/m 14 heb ik mij vooral beziggehouden met het uitvoeren van Bivariate analyses op de pinguïns en mijn gekozen IMDB-datasets. Dit deed ik door onder andere contingency tables te maken en chi-squared tests te maken en deze resultaten te analyseren om zo meer informatie te verkrijgen over de datasets.

Multivariate analyse is een methode waarbij de relatie tussen drie of meer variabelen wordt geanalyseerd. Het is een analyse dat tot doel heeft om de complexe relatie tussen meerdere variabelen te begrijpen en te modelleren. In opdracht 15 t/m 18 heb ik mij vooral beziggehouden met het uitvoeren van Multivariate analyses op de pinguïns en mijn gekozen IMDB-datasets. Dit deed ik door onder andere een ‘decision tree model’ te maken om zo voorspellingen te kunnen doen over de data door een ‘set’ te trainen en deze te vergelijken met een ongetrainde set. Door deze resultaten te analyseren heb ik vanuit een ander perspectief de dataset kunnen bekijken.

Uiteindelijk heb ik dus door het uitvoeren van al deze typen analyses een andere kijk gekregen op mijn datasets. Door vanuit deze perspectieven te kijken en verbanden te vinden binnen een grote hoeveelheid informatie heb ik me actief beziggehouden met data science, omdat ik gegevens op verschillende manieren analyseerde en verwerkte om uiteindelijk aanbevelingen en voorspellingen te kunnen doen. Data science houdt dus vooral in dat een individu of organisatie actief bezig is met het analyseren van gegevens om voorspellingen en aanbevelingen te kunnen doen op basis van gevonden verbanden in datasets. Hiermee kan de besluitvorming worden verbeterd en dus de toekomst van het bedrijf verbeterd worden. De definitie van data science is zo dus te achterhalen uit de opdrachten die ik heb gemaakt.

Data science en business intelligence zijn gerelateerde maar verschillende vakgebieden die beide gegevens gebruiken om besluitvorming in een zakelijke context te ondersteunen. Data science richt zich op het ontwikkelen van voorspellende modellen, terwijl business intelligence zich richt op het bieden van inzichten en het presenteren van informatie om besluitvorming te ondersteunen. Hieruit kun je dus concluderen dat doelgerichte analyses op wiskundige modellen (data science) leiden tot een betere besluitvorming (bussiness intelligence).

Data science is onlosmakelijk verbonden met AI, omdat veel van de technologieën die worden gebruikt bij het ontwikkelen van AI-systemen zijn gebaseerd op machine learning-technieken die zijn ontwikkeld door data scientists. Data science is daarom een ​​cruciaal onderdeel van het ontwikkelen van AI-systemen die in staat zijn om menselijke taken uit te voeren en zichzelf te verbeteren. Een mooi voorbeeld hiervan is hoe ik in een eerder gemaakte opdracht code heb moeten schrijven die voorspelt welke categorie behoord tot een rij data door te kijken naar de eigenschappen.