## [SLIDE 1]

Dag iedereen  
Robin en Ik zullen ons project over ‘Spaceship Titanic’ voorstellen aan jullie.

## [SLIDE 2]

We beginnen met een korte introductie, daarna overlopen we ons data onderzoek in PowerBi, hoe we de dataset opgekuist hebben met behulp van Python, wat we kunnen afleiden uit de correlatiematrix en hoe we ons model getraind hebben. We sluiten af met onze conclusie over dit project

## [SLIDE 3]

Wat hield ons project in?

We zijn het jaar 2912 en hebben een boodschap ontvangen van ‘Spaceship Titanic’. Tijdens een onoplettendheid is het schip een aanvaring gehad met een wormgat waarbij ± de helft van de bemanning verloren is geraakt.

Het was onze taak om op basis van de bestaande data te voorspellen wie er een grotere kans heeft om geteleporteerd te worden

## [SLIDE 4]

Voor we konden beginnen aan het opkuisen van de data, bekeken we de dataset eens van dichterbij. Zo konden we al een eerste inschatting maken van welke velden er veel invloed zouden hebben op het resultaat.

Hier zijn enkele grafieken die te maken hebben met de overlevingskansen van de passagiers.

## [SLIDE 5]

Als we de vertrek planeten bekijken, zien we dat Aarde het populairst was. Het aantal verloren passagiers bleef wel constant voor elke planeet. Er waren ook 201 passagiers waarvan deze info niet gekend was

## [SLIDE 6]

Als we de destinaties bekijken, zien we dat Trappist de populairste bestemming was. Opnieuw is aantal verloren passagiers constant voor elke planeet. Van 182 passagiers was de eindbestemming niet gekend.

## [SLIDE 7]

Hier vinden we onze eerste interessante grafiek. We zien hier duidelijk dat mensen in Croysleep een grotere overlevingskans hebben dan mensen die niet in cryosleep reizen. Ook hier merken we dat er 200 onbekende keuzes aanwezig waren.

## [SLIDE 8]

Als we de leeftijd bekijken zien we iets opmerkelijk. Als een persoon nog zeer jong is, is zijn overlevingskans opmerkelijk groter. Ook zien we dat mensen tussen de 200 en 300 jaar het meest reizen.

## [SLIDE 9]

We vonden het ook interessant om eens te kijken of een duurdere reis gelijk stond met een veiligere reis. Dit bleek dan toch ook 50/50 te zijn.

## [SLIDE 10]

Blanke waarden in een dataset zijn niet interessant omdat ze geen informatie bevatten die het model kan gebruiken om voorspellingen te doen.

Als er te veel ontbrekende waarden zijn, kan dit een negatieve invloed hebben op de nauwkeurigheid van het model.

## [SLIDE 11]

Voor de kolommen "homeplanet" en "destination" kozen we er voor om elke planeet om te zetten in een uniek nummer. Dit is handig omdat veel machine learning algoritmes alleen met numerieke data kunnen werken.

Vervolgens berekenen we het gemiddelde van de gecodeerde waarden om de ontbrekende planeten in te vullen. Met andere woorden, we vervangen de ontbrekende waarden door het gemiddelde van de nummers die de planeten in die kolom vertegenwoordigen.

## [SLIDE 12]

Voor de leeftijd hebben we besloten om de blanke waarden simpelweg te vervangen door de gemiddelde leeftijd van alle passagiers.

## [SLIDE 13]

Bij cryosleep hebben we belist om alle ontbrekende waarden te vervangen door ‘False’ omdat deze het meest voorkomt en dus het minst zal doorwegen als we die er bij tellen.

## [SLIDE 14]

Hier hadden we 2 scenario’s. Als de passagier koos voor cryosleep werdt de onbrekende waarde ingevulde door 0.

Dit omdat de passagier tijdens die toestand onmogelijk gebruik kon maken van deze diensten. Als de passagier niet in cryosleep toestand was, namen we het gemmidelde gebruik van de andere passagiers.

## [SLIDE 15-16]

Uit de corrolatie matrix kunnen we afleiden dat cryosleep positief gecorolleerd is met de overlevingskans, en diensten negatief gecorolleerd.

## [SLIDE 17]

Dit betekend dat passagiers in cryosleep een grotere overlevingskans hadden dan de passagiers die wakker waren. We kunnen ook afleiden dat het gebruik van diensten negatief gecorolleerd is met de overlevingskans.

## [SLIDE 18]

Hoe minder er gebruik werd gemaak van een dienst, hoe groter de overlevingskans was. Dit viel ook te verwachten aangezien mensen in cryosleep geen diensten kunnen gebruiken.

## [SLIDE 19-20]

In het laatste deel van ons project hebben we data gebruikt om een model te maken dat voorspellingen kan doen over testgegevens. We hebben hiervoor verschillende classifiers geprobeerd en uiteindelijk gebruik gemaakt van de Random Forest-classifier omdat deze ons de hoogste nauwkeurigheid gaf. Random forest werkt door het combineren van vele beslissingsbomen, waarbij elke boom een beslissing neemt op basis van een subset van de beschikbare gegevens.

Na het trainen van de gegevens hebben we het model gebruikt om te voorspellen of passagiers geteleporteerd werden of op het schip bleven en deze resultaten ingediend op Kaggle. Onze inzending scoorde 0.79728.

## [SLIDE 21]

Er zijn een aantal dingen die we geleerd hebben uit dit project. Over het project zelf hebben we kunnen concluderen dat cryosleep de grootste invloed had op de overlevingskans van een passagier op spaceship titanic.

We zullen ook onthouden dat PowerBi een krachtige tool is om een grote dataset te visualiseren en waardoor ze ook beter te verstaan is.

We hebben ook doorgehad dat negatieve of foute waarden in een dataset de precisie van je model sterk kunnen verlagen en je verschillende manieren hebt om met deze waarden om te gaan en op te lossen.