



# BIG DATA AND BUSINESS DEVELOPMENT

Predicting Olympic Games

Howest, 2024-2025

Dries Deboosere, Jasper François, Lesy Maxim

## Inhoudsopgave

Legende .....	2
Onderzoeksraag .....	3
Voorbereiding en analyse van de dataset .....	3
Algemene aandachtpunten bij concrete features .....	3
Enkel topsporters .....	4
Historische data .....	5
Historische lichaamsindicatoren atleet .....	5
Historische uitslagen wedstrijden .....	5
Sportsoorten- classificaties en subtypes .....	5
Zwakke correlaties .....	6
Resultaten .....	7
Model 1: kans op winst .....	7
Model 2: best passende sport op basis van lichaamsindicatoren .....	8
Applicatie & Random Forest Classification model .....	8
Conclusie .....	9

## Legende

Het document is zo opgesteld dat het een blik geeft op het aangelegde traject tijdens het onderzoek.  
Gaandeweg werden vaststellingen en keuzes gemaakt.

De tabel hieronder geeft de betekenis van het icoongebruik in dit document aan die het aangelegde traject helpen te verduidelijken:

-  De vaststelling leidt tot een probleem ten opzichte van het oorspronkelijke doel
-  De vaststelling geeft blijkt het opzet te kunnen bereiken
-  De vaststelling bemoeilijkt het correct interpreteren van de data
-  De oplossing gaat als volgt ...

## Onderzoeksraag

De dataset bevat heel wat informatie over de individuele uitslagen van de Olympische Spelen. Daarbij hebben we zicht op de uitkomst van het spel, welke speler dit was, enzovoort.

Toch zijn er, na een uitgebreide analyse van de dataset, een aantal belangrijke aandachtspunten naar boven gekomen. Hierdoor konden we tijdens onze onderzoekstocht niet zomaar een pad van A naar B bewandelen.

Het opzet van het onderzoek was om te achterhalen of er een verband gelegd kan worden tussen het gewicht, de lengte (en daarmee indirect het BMI) en andere indicatoren van een atleet om te achterhalen **welke kans een atleet heeft op het halen van een podiumplaats (en dus een medaille) op basis van zijn lichaamsindicatoren (gewicht, lengte, BMI, geslacht)**. Tijdens ons onderzoek is echter gebleken dat we deze vraag niet zomaar kunnen beantwoorden. Waarom dat zo is wordt verderop duidelijk gemaakt.

Om een relevantere einduitwerking te kunnen afleveren, werd de onderzoeksraag uiteindelijk gewijzigd naar:

---

*Welke (sport)discipline beoefent een atleet het meest waarschijnlijk (of het best) op basis van zijn/haar lichaamsindicatoren (lengte, gewicht, BMI, geslacht)?*

---

## Voorbereiding en analyse van de dataset

### Algemene aandachtspunten bij concrete features

Naast de voor de hand liggende aandachtspunten, zoals bijvoorbeeld het correct omzetten naar het juiste datatype van een feature, zijn deze noemenswaardig:

- **Gewicht en lengte:** niet steeds meegeleverd of in een foutief formaat, maakt het onmogelijk om het model correct te trainen eveneens het berekenen van het BMI
- **Leeftijd** valt niet exact te achterhalen, maar wel een ruwe schatting. We kennen de geboortedatum van de atleet, maar we kunnen de exacte datum van de Spelen niet altijd achterhalen tot op de dag. Hierdoor kunnen we niet altijd exact berekenen hoe oud de atleet was tijdens de Spelen.
- **Medaille:** bevat heel wat ‘na’-waardes, maar deze moeten zeker opgenomen worden omdat ze wijzen op ‘geen winst’. Ander waardes in deze feature zijn ‘Gold’, ‘Silver’ en ‘Bronze’, welke allen duiden op een winst.



Met deze punten moeten zeker rekening gehouden worden, maar doorgaans kunnen we de data omvormen of eruit filteren.



*Indien we gegevens te kort komen (gewicht óf lengte mist, bijvoorbeeld), dan filteren we deze eruit. Data die niet omgezet kan worden naar een ander type (bv. een gewicht van “78-80” in plaats van “78” als concreet getal) wordt ook verwijderd.*

## Enkel topsporters

De dataset is wezenlijk een selectie van een zeer selecte groep mensen die enkel binnen de categorie ‘topsporters’ vallen en dat bovendien ook telkens binnen een heel specifieke sport.

De dataset is dus alleen representatief voor topsporters die een bepaalde sport beoefenen. We kunnen geen voorspellingen doen voor de gemiddelde persoon, omdat de dataset daar geen informatie over bevat. Als we een voorspelling willen maken, dan moet dat voor topsporters zijn.

We kunnen daarom argumenteren dat de mensen die deelnemen aan een bepaalde discipline aan bepaalde vereisten moeten voldoen. Neem als voorbeeld sumoworstelen waarin spelers met een hoog gewicht/BMI doorgaans deelnemen (en bijgevolg kunnen winnen), terwijl dit voor duurlopen eerder atleten zijn met een lager BMI.

Dit maakt het verband leggen tussen het BMI en de kans op winst moeilijk waarneembaar zonder daadwerkelijk rekening te houden met de sport (of sub sport) in kwestie.

Een concreet voorbeeld van dit aandachtspunt gaat als volgt:

- Iemand met een laag BMI zal allicht nooit ‘sumoworstelen’ winnen.
- Bovendien zou deze persoon waarschijnlijk nooit geselecteerd worden om überhaupt deel te nemen aan de Spelen in **deze** discipline, maar mogelijk wel in een **andere**.
- Bijgevolg zitten onder de sport ‘sumoworstelen’ enkel personen met een hoog BMI – dit omwille van de Olympische selectie die reeds gemaakt is.

De BMI’s en de kans op een winst kunnen dus niet zomaar rechtstreeks vergeleken worden onder aanvullend rekening te houden met de sport, maar we moeten ons nog steeds indachtig blijven dat we een proef doen op een selecte groep mensen.

Verder moeten we opmerken dat we in theorie wel juiste conclusies zouden moeten kunnen trekken indien we rekening houden met deze vaststelling (en dus de soort sport indachtig blijven), maar deze aanname wordt **ontkracht wanneer we de correlaties tussen de features in de uiteindelijke dataset analyseren**. Dit omdat atleten op dit niveau in hun eigen sport al een erg gelijkaardig fysiek profiel hebben. Er zit bijvoorbeeld weinig variatie in het BMI van de beste of slechtste zwemmer omdat we al spreken over een zeer selecte topgroep.

## Historische data

### Historische lichaamsindicatoren atleet

Een eerste belangrijke opmerking is dat de lengte en het gewicht van een atleet zijn slechts éénmaal gekend zijn, namelijk zoals deelgenomen tijdens de allerlaatste Olympische Spelen. Dit maakt dat, wanneer een atleet aan meerdere spelen deelnamen, hij volgens de dataset dezelfde lichaamsindicatoren heeft voor alle Spelen. Anders gezegd: we kennen de lichaamsindicatoren (zoals gewicht, lengte, ...) tijdens de individuele Olympische Spelen niet.



Deze vaststelling maakt het alvast onmogelijk om met historische data aan de slag te gaan en bijvoorbeeld te onderzoeken of een verandering in gewicht en/of BMI van een individuele atleet een invloed heeft op de prestaties.

### Historische uitslagen wedstrijden

In tegenstelling tot de lichaamsindicatoren zijn de uitslagen van de wedstrijden wel compleet.



Met deze info kunnen we, zoals vooropgesteld, voorspellingen maken die kaderen om een uitslag te voorspellen – en dit eventueel op basis van de lichaamsindicatoren.

## Sportsoorten- classificaties en subtypes

Hoewel de dataset veel gegevens bevat, is deze eigenlijk niet uniform. Concreet bedoelen we hiermee dat de dataset handelt over verschillende **sporten met elk hun eigen regels**.

Enkele voorbeelden:

- Bij voetbal win je als team, terwijl je bij tennis wint als enkeling.
- Bij tennis win je dan weer door een hoog aantal punten te scoren, terwijl je bij bijvoorbeeld lopen wint door als eerste te finishen



Bovenstaande hoeft niet steeds een probleem te zijn. Op het einde van de rit wordt voor elke sport tenslotte een winnaar met eerste, twee en derde plaats aangeduid. Dit vertaalt zich in de gouden, bronzen en zilveren medaille voor de winnaars en niets voor de ‘verliezers’.

Bovendien kunnen we de dataset hiervoor veel minder al een algemene dataset met coherente data beschouwen, maar is het eigenlijk een collectie van fijnmazige groepen samen. Hier moeten we dus rekening mee houden.



Willen we echter iets persoonsgebonden doen, dan komen we in de problemen voor teamsporten. Om die reden filteren we alle teamsporten uit de dataset.

## Zwakke correlaties

Onze aanname was dat het BMI een invloed zou hebben op de prestaties van een atleet. Echter, na het analyseren van de correlaties van onze dataset blijkt dat hier slechts een (zeer) zwakke correlatie tussen bestaat. Bovendien heerst er een zwakke correlatie tussen eender welke twee andere features, denk bijvoorbeeld aan een correlatie tussen BMI en de kans op een medaille/winst.

	sport	event	pos	medal	weight	height	sex	year	age	bmi
sport	1.000000	-0.114316	-0.100071	0.024046	0.114818	0.062520	0.039147	0.107384	-0.127503	0.127641
event	-0.114316	1.000000	-0.058905	-0.042034	0.115164	-0.045473	0.078364	-0.005488	0.197296	0.200092
pos	-0.100071	-0.058905	1.000000	0.354855	-0.109169	-0.047392	0.050561	0.089321	0.028669	-0.111983
medal	0.024046	-0.042034	0.354855	1.000000	-0.051628	-0.023684	-0.009739	0.087372	0.001878	-0.050593
weight	0.114818	0.115164	-0.109169	-0.051628	1.000000	0.746315	0.525910	-0.034523	0.138662	0.843092
height	0.062520	-0.045473	-0.047392	-0.023684	0.746315	1.000000	0.536039	0.013791	0.061781	0.281351
sex	0.039147	0.078364	0.050561	-0.009739	0.525910	0.536039	1.000000	-0.177432	0.135268	0.350036
year	0.107384	-0.005488	0.089321	0.087372	-0.034523	0.013791	-0.177432	1.000000	0.123650	-0.066070
age	-0.127503	0.197296	0.028669	0.001878	0.138662	0.061781	0.135268	0.123650	1.000000	0.158272
bmi	0.127641	0.200092	-0.111983	-0.050593	0.843092	0.281351	0.350036	-0.066070	0.158272	1.000000

**Deze vaststelling ondermijnt ons initieel opzet. Bovendien komen eerder erkende problemen ook hier weer naar boven. Klaarblijkelijk kan er geen verband gelegd worden tussen iemands BMI en zijn resultaat.**



**We moeten dus helaas besluiten dat de data die we uit de dataset kunnen halen, geen bruikbare data is om een correct model te bouwen en voorspellingen te maken zoals we ze vooropstelden.**



Om de correlaties zo minimaal mogelijk te houden vereenvoudigen we de ‘medailles’ en ‘posities’ uiteindelijk naar een ook binair resultaat. Zo kregen we antwoord op de vraag of iemand op het podium staat of niet (en heeft hij/zij dus iets gewonnen). Daarmee hopen we een duidelijker verband waar te kunnen nemen, maar ook op deze manier valt geen significante correlatie te bespeuren.

## Resultaten

De bovenstaande bedenkingen en problemen leiden ertoe dat we moeten vaststellen dat het niet mogelijk is om aan de hand van het BMI te achterhalen of iemand een grote kans heeft op een winst. Ook alternatieve methodes, zoals werken met de lengte en het gewicht in plaats van het BMI, leveren geen betere resultaten op.

We zijn er ons dus ten volle van bewust dat we werken met een dataset die uiteindelijk lage correlaties tussen de individuele features heeft, waardoor we verwachten dat het model niet accuraat werkt.

Niettegenstaande schreven we verschillende modellen.

### Model 1: kans op winst

Aanvankelijk werd er een K-Nearest Neighbors-model (KNN) getraind die kan voorspellen wat de kans is op winst. Deze heeft gek genoeg een accuraatheid van ongeveer 85 %.

Een observatie is dat het model doorgaans voorspelt dat er geen medaille verdient zal worden. Aangezien er telkens slechts 3 medailles gewonnen kunnen worden (Goud, Zilver, Brons) en **alle** andere spelers niets winnen – is het dus logisch dat het model aanneemt dat we niets zullen winnen: de overgrote meerderheid in de dataset wint tenslotte ook niets.

Bovendien weten we dat we starten vanaf een dataset met lage correlaties. De voorspelling, ook al ‘accuraat’, moeten we dus met een korrel zout nemen.



We concluderen dat dit KNN-model programmatorisch juist is, maar dat het functioneel niet bruikbaar is.

## Model 2: best passende sport op basis van lichaamsindicatoren

Omwille van de eerder aangehaalde problemen, wijzigden we de oorspronkelijke onderzoeksraag naar een quasi omgekeerde opzet, namelijk:

---

*Welke (sport)discipline beoefent een atleet het meest waarschijnlijk (of het best) op basis van zijn/haar lichaamsindicatoren (lengte, gewicht, leeftijd, geslacht)?*

---

We kozen voor deze vraag omdat het, van een gebruikersstandpunt, een concreter antwoord geeft.

We hebben ons model dan zodanig opgebouwd zodat het gebruikt kan worden voor bovenstaande onderzoeksraag. We stelden hiervoor een Random Forest Classification model, dit omdat de voorspelling die we willen maken een categoriek resultaat.

### Applicatie & Random Forest Classification model

Als applicatie hebben we gekozen voor een ASP.NET Blazor webapplicatie omdat we hier al ervaring mee hebben.

Om met Tensorflow te werken in een .NET omgeving zijn er verschillende libraries beschikbaar. Een van de meest bekende is Tensorflow.NET ([TensorFlow.NET \(scisharp.github.io\)](https://TensorFlow.NET (scisharp.github.io))). Maar na enkele uren proberen om ons model te laden en te gebruiken, kwamen we erachter dat dit niet haalbaar was vanwege ons gekozen model. Ons model is gebaseerd op Random Forest Classification, en dit is niet compatibel met Windows via Tensorflow en dus ook niet met Tensorflow.NET. We hadden de optie om Tensorflow te downgraden naar een lagere versie om sommige dependencies toch te laten werken, maar dit zou weer andere dependencies verstoren.

We overwogen eerst om ons framework te veranderen naar een JavaScript webapplicatie, maar omdat we al een UI hadden gemaakt en uitgewerkt, kozen we voor een tussenoplossing.

We maakten een API in Python waar we de input van de webapplicatie naar toe konden sturen. Ons model kon in Python voorspellingen maken op basis van deze input.

We hebben eerst een app gemaakt die de sport kon voorspellen, niet de discipline (of event in onze dataset). Toen we de app werkend hadden (input en voorspelling), hebben we een ander model gemaakt dat de discipline kon voorspellen. Maar dit model was veel groter: 16 GB. Onze Python API kon dit niet aan. Daarom zijn we teruggegaan naar ons eerste model dat de sport voorspelt en maar 1,5 GB is.

We hebben bepaald welk classificatie- of regressiemodel het beste werkte voor ons doel. Random Forest Classification bleek het meest geschikt.

Random Forest Classification is een populaire keuze voor voorspellende modellen, vooral in gevallen waar de relatie tussen inputvariabelen en de doelvariabele complex of niet-lineair kan zijn, zoals vaak voorkomt in sportanalyse.

## Conclusie

We hebben ons oorspronkelijke idee om een medaillevoorspelling te doen op basis van een BMI moeten opgeven omdat de data die we hadden niet toereikend waren.

We hebben ons idee bijgesteld zodat we de meest geschikte sport konden voorspellen op basis van persoonlijke kenmerken.

Maar ook hier moeten we rekening mee houden dat de brongegevens onvoldoende informatie bevatten om een betrouwbare voorspelling te kunnen doen.