

**Justification Document**

Jinne Haan en Daan Derks



4 juni 2025

fontys university of applied sciences

Inhoud

[Gemaakte keuzes 2](#_Toc199880807)

[Data cleaning 2](#_Toc199880808)

[Parameter resultaten 3](#_Toc199880809)

[AI gebruik 4](#_Toc199880810)

[Wie heeft wat gedaan? 4](#_Toc199880811)

# Gemaakte keuzes

Aangezien de opdracht eigenlijk 6 kleine programeer opdracht waren, hebben we de opdracht in 1 bestand gemaakt. Op deze manier kunnen we alles in 1 keer runnen om dingen makkelijk te testen.

Ook hebben we ervoor gekozen om in een python bestand te werken en niet in een notebook bestand. De reden hiervoor is omdat wij het persoonlijk fijner vinden werken, daarnaast zie je in de praktijk bij bedrijven dat python bestanden meer voorkomen dan notebook bestanden. Om ons een zo goed mogelijke voorbereiding te geven op het bedrijfsleven maken wij de python file al ons eigen.

We hebben gekozen om 1 test-scores te berekenen, zodat we bij alle modellen dezelfde score kunnen vergelijken:

* R2\_score: gekozen omdat dit de variantie uitlegt bij een goed model. Word ook vaker gebruikt en is betrouwbaar.

We hebben de data gesplit in een man dataset en vrouw dataset. Zodat we voor beide kunnen voorspellen. Vervolgens hebben we beide datasets gesplit in een 80/20 verhouding.

# Data cleaning

Tijdens de data cleaning zijn we erachter gekomen dat er geen gedupliceerde data rijen zijn. Vervolgens hebben we de status en country kolommen veranderd in getallen, zodat er analyses op uitgevoerd kunnen worden.

Vervolgens hebben we de nan waardes vervangen door de gemiddelde waarde in het land waar de data vandaan komt. De nan waardes die voorkwamen in de thinness 1-19 en 5-9 years kolommen hebben we verwijderd omdat deze rijen teveel informatie bleken te missen. Hetzelfde geldt bij de Income composition of resources en schooling kolommen.

Als laatste hebben we de kolommen ['Under-five deaths', 'GDP', 'thinness 5-9 years', 'Schooling'] verwijderd om Multicollineariteit te voorkomen en hebben we de ‘Population’ kolom verwijderd omdat die heel veel nan waardes heeft en weinig invloed heeft op de y waardes.

# Parameter resultaten

Alle modellen die we probeerden waren erg slecht. Bij geen enkel model kwam er een R2 score uit van 0,01 of hoger, dit betekent dat je beter het gemiddelde kan pakken van de target.

De beste r2 score van de verschillende modellen:

* Multivariate linear regression models: -0.009876873717058254
* Decision tree regression: -0.0014732459682358368
* Bagging with regression trees: -0.030741886333720325
* AdaBoosting with regression trees: -0.0044757761315052935
* RandomForest regression: -0.024266085916314406

# AI gebruik

We hebben voor verschillende onderdelen AI gebruikt.

Syntax: Het programmeren doen we zelf, we bedenken wat we willen doen en hoe de flow ongeveer zal gaan. Wanneer we bezig zijn met programmeren en de syntax lukt niet helemaal, dan vragen we ai hoe dat moet en om voorbeelden. Een voorbeeld: we hebben gevraagd om voor te doen hoe je vanuit de fetch\_ucirepo datasets kunt inladen in python en transformeren in een Pandas dataframe. We wisten dat dit kon, maar niet precies hoe, want het is de eerste keer dat we met die package werken.

Test-scores: nadat we hebben gekeken naar welke test-score we gaan gebruiken hebben we nog even aan ai nagevraagd of dat die het eens is met de keuze of, of dat hij toch een andere test-score zou aanraden. Hieruit bleek dat we goed bezig waren.

Errors: soms krijgen we errors en is het moeilijk te begrijpen waar het mis gaat. Dan vragen we ai om inzicht te bieden in wat er mis gaan en een mogelijk oplossing suggestie te krijgen.

# Wie heeft wat gedaan?

We hebben samen aan alle code gewerkt. Dit hebben we gedaan door naast elkaar te zitten en te programmeren, meestal hebben we dan allebei kleine taken die de hele tijd samen vallen. We hebben naast elkaar samengewerkt via de ‘live share’ extensie van visual studio code gemaakt door Microsoft. Op deze manier kunnen we een beetje als een online word document tegelijk in een document typen.

Alle keuzes die gemaakt zijn tijdens deze opdracht zijn dan ook samen genomen en overlegd.

Het justification document is gemaakt door:

Keuze van de datasets: Daan

Gemaakte keuzes: beide

AI gebruik: Jinne

Wie heeft wat gedaan?: beide

Sources:

OpenAI. (2023). *ChatGPT* (Mar 14 version) [Large language model]. <https://chat.openai.com/chat>

*Course modules: FEB25: - Data Science 8*. (n.d.). <https://canvas.fontys.nl/courses/26895/modules>