

# Eerlijke machine learning met studiedata

Een statistisch onderzoek naar kansengelijkheid voor studenten van De Haagse Hogeschool

Theo Bakker, lector Learning Technology & Analytics

SURF Onderwijsdagen 2023

15 november 2023



**DE HAAGSE**  
HOGESCHOOL



Fotografie: David Meulenbeld

# Contact

Dr. Theo C. Bakker

De Haagse Hogeschool

[t.c.bakker@hhs.nl](mailto:t.c.bakker@hhs.nl) | 06-25637172

Website Lectoraat Learning Technology & Analytics

<https://www.dehaagsehogeschool.nl/onderzoek/lectoraten/learning-technology-analytics>

Persoonlijke website

<https://hapax-analytics.nl>

**Lector Learning Technology & Analytics, De Haagse Hogeschool**  
onderzoek naar studiedata en gelijke kansen

**Universitair Docent, Vrije Universiteit**  
onderzoek naar neurodiversiteit in het hoger onderwijs



sinds 2012  
10.000 studenten beschuldigd van fraude  
6.000 bezwaar  
1.500 rechtszaken, 25% studenten wint  
N = 367, bijna allen een migratie-achtergrond

NOS Helen Kret

NOS op3



Woensdag 21 juni, 06:14

## Studenten met migratieachtergrond opvallend vaak beschuldigd van fraude, minister wil systeem grondig nagaan

# eerlijke machine learning met studiedata

## eerlijkheid en data

wat is eerlijk  
wat zijn studiedata  
wat zijn sensitieve data

1

## bias

wat is bias  
hoe bereken je bias

2

## detectie en mitigatie

hoe ontdek je bias  
hoe visualiseer je bias  
keuzes om bias tegen te gaan

3

## algemene aandachtspunten



# De Staat van De Haagse

Een statistisch onderzoek naar kansengelijkheid aan De Haagse Hogeschool

## Drie deelonderzoeken

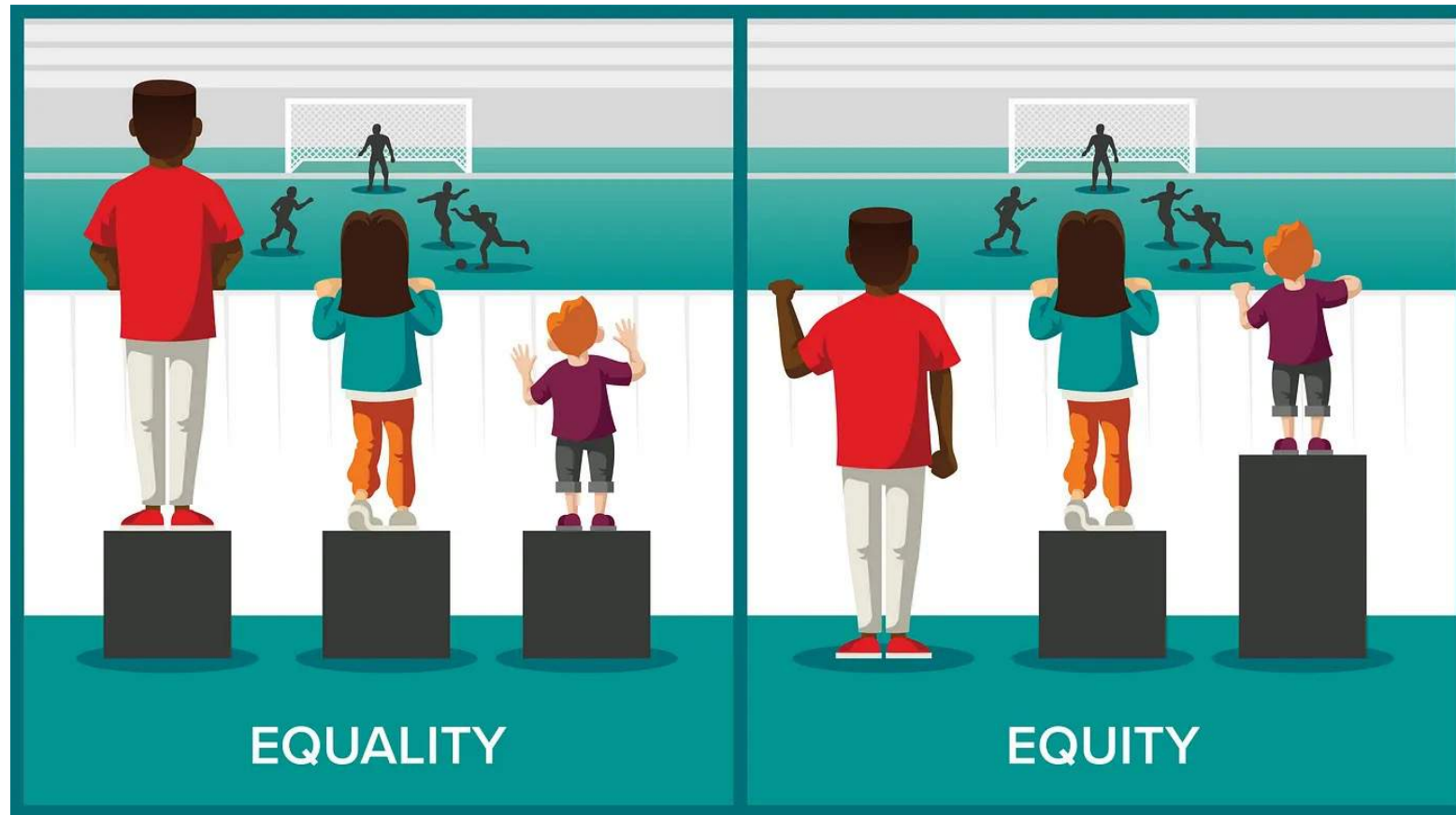
1. Gelijke kansen in **instroom**
2. Gelijke kansen in **studievoortgang en studiesucces** (doorstroom en diplomering)
3. Gelijke kansen op de **arbeidsmarkt of in een vervolgstudie**

Lectoraat Learning Technology & Analytics,  
Dr. Theo Bakker



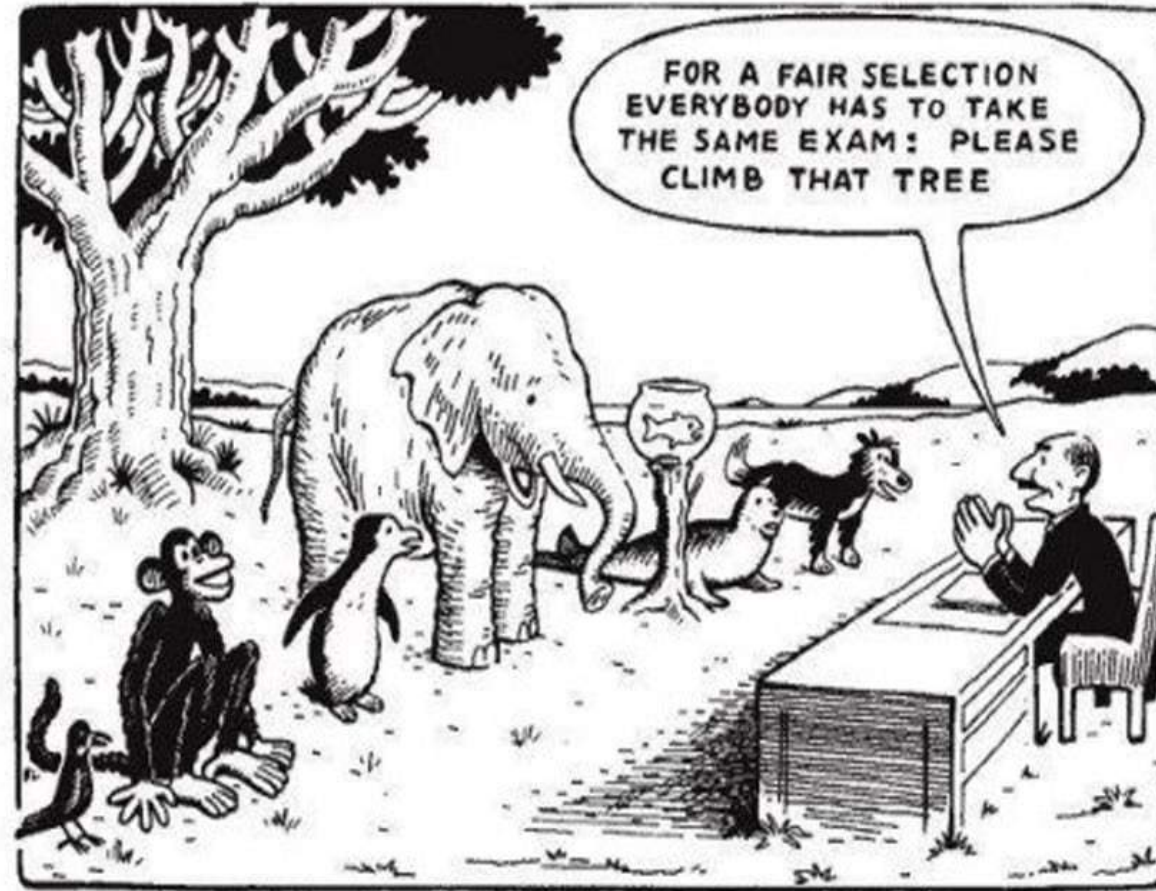
# eerlijke machine learning met studiedata

# eerlijke machine learning met studiedata



Illustratie: Medium.com

# eerlijke machine learning met studiedata



Illustratie: Hans Traxler, 1976



# eerlijke machine learning met studiedata



Illustratie: Trouw, 10 april 2019, studio vonq

# eerlijke machine learning met studiedata

inclusie + diversiteit + gelijke kansen

op onderwijs (instroom) + een diploma (doorstroom) + een passende baan

# eerlijke machine learning met **studiedata**

**inschrijvingen** + **resultaten** + online gedrag  
**studiesucces** + opleidingen + curricula + onderwijskwaliteit  
studentsucces + studentwelzijn  
achtergrondkenmerken + leeftijd + **geslacht** + vooropleiding  
eerdere resultaten + herkomst + sociaal economische status

# voorspelmodellen uitleggen en bias detecteren met sensitieve data

# voorspelmodellen uitleggen en bias detecteren met **sensitieve data**

## bijzondere persoonsgegevens

AVG

- ras of etnische afkomst
- politieke opvattingen
- religieuze of levensbeschouwelijke overtuigingen
- lidmaatschap van een vakbond
- gezondheid
- seksueel gedrag of seksuele gerichtheid
- **genetische gegevens**
- **biometrische gegevens (bedoeld voor de unieke identificatie van een persoon)**

## sensitieve data

Handboek over het Europese  
non-discriminatierecht

- **geslacht**
- **genderidentiteit**
- seksuele geaardheid
- handicap
- **leeftijd**
- ras, etniciteit, kleur en het behoren tot een nationale minderheid
- geloof of geloofsovertuiging
- **sociale afkomst, geboorte en eigendom**
- taal
- politieke over andere overtuigingen
- status anders (verzameling aan **kenmerken**)



# voorspelmodellen uitleggen en bias detecteren met **sensitieve data**

## bijzondere persoonsgegevens

AVG

- ras of etnische afkomst
- politieke opvattingen
- religieuze of levensbeschouwelijke overtuigingen
- lidmaatschap van een vakbond
- gezondheid
- seksueel gedrag of seksuele gerichtheid
- genetische gegevens
- biometrische gegevens (bedoeld voor de unieke identificatie van een persoon)

**niet identificeren**

## sensitieve data

Handboek over het Europese  
non-discriminatierecht

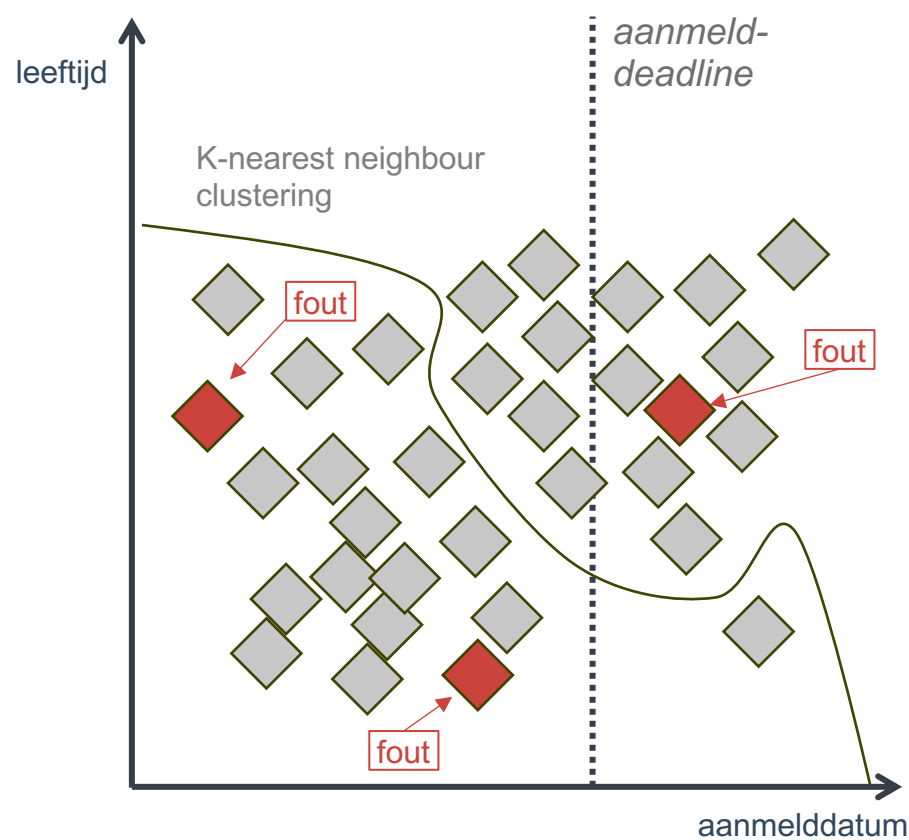
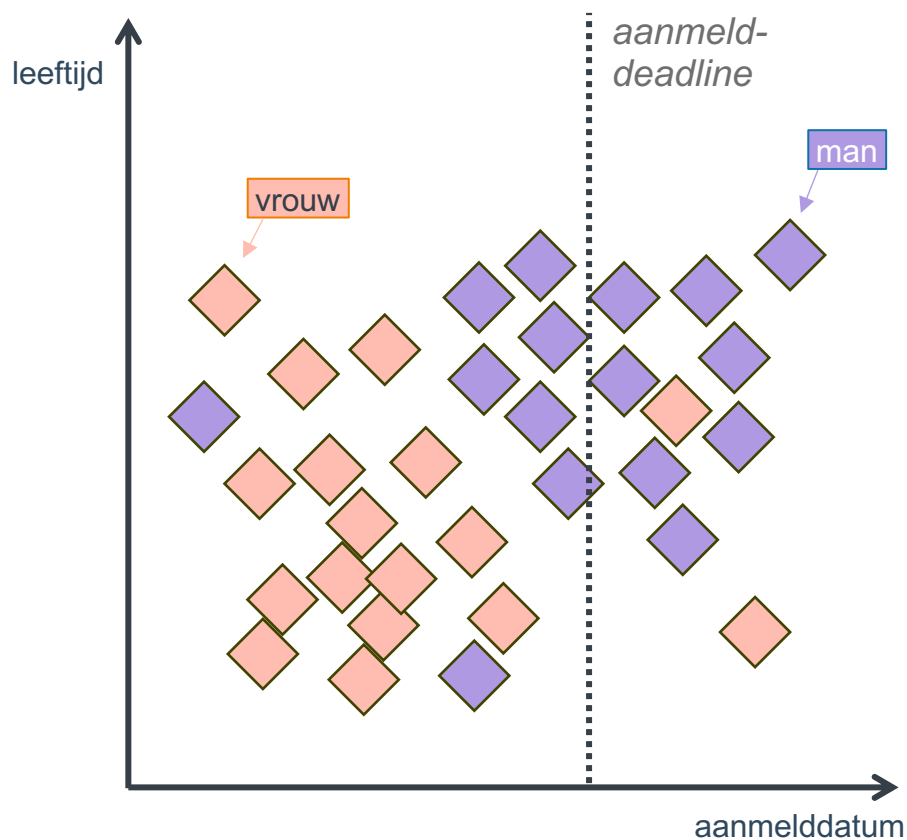
- geslacht
- genderidentiteit
- seksuele geaardheid
- handicap
- leeftijd
- ras, etniciteit, kleur en het behoren tot een nationale minderheid
- geloof of geloofsovertuiging
- sociale afkomst, geboorte en eigenaarschap
- taal
- politieke over andere overtuigingen
- status anders (verzameling aan kenmerken)

**niet discrimineren**





# voorspelmodellen uitleggen en bias detecteren met **sensitieve data**



**“no fairness without awareness”**

# criteria om **bias te detecteren** is deze selectieprocedure fair?

**aanmelding**

120 vrouw  
50 man



**selectie**

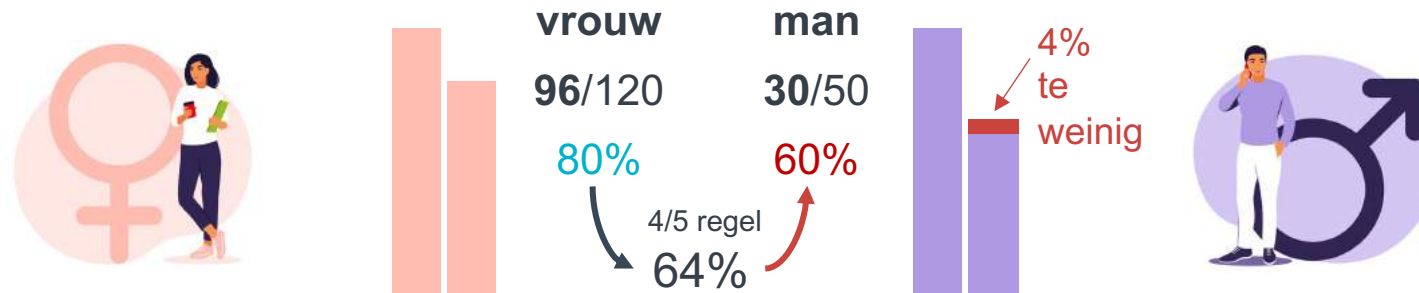
96 vrouw  
30 man

**selectieve zorg-opleiding**  
**max. 126 studenten**

eerdere cijfers, ervaring in de zorg,  
motivatiebrief en gesprek

# criteria om **bias te detecteren**

is deze selectieprocedure fair – 4/5 criterium\*



**aanmelding**

120 vrouw  
50 man



**selectie**

96 vrouw  
30 man

\* Code of Federal Regulations. Section 4d, uniform guidelines on employee selection procedures (1978)  
= *statistical parity*

# criteria om **bias te detecteren**

## confusion matrix

gaat een student  
wel of niet **het eerste**  
**studiejaar halen?**

**werkelijke** uitkomst

| voorspelde uitkomst |          |                     |                     |
|---------------------|----------|---------------------|---------------------|
|                     |          | Positief            | Negatief            |
|                     | Positief | True Positive (TP)  | False Negative (FN) |
|                     | Negatief | False Positive (FP) | True Negative (TN)  |

fout en **wel een probleem**  
voor de student >  
ten onrechte een negatief BSA

Type II Fout

Type I Fout

weliswaar fout, maar **geen probleem**  
voor de student >  
boft met een positief BSA

let op! stel je de vraag  
omgekeerd (op uitval),  
dan is de matrix precies  
andersom...



# criteria om **bias te detecteren**

## confusion matrix

gaat een student  
wel of niet **het eerste  
studiejaar halen?**

|                     |          | voorspelde uitkomst |          |                  |
|---------------------|----------|---------------------|----------|------------------|
|                     |          | Positief            | Negatief |                  |
| werkelijke uitkomst | Positief | 50                  | 50       | 100              |
|                     | Negatief | 50                  | 50       | 100              |
|                     |          | 100                 | 100      | 50% accuraatheid |

**een slechte prognose:** 50% is correct voorspeld

# criteria om **bias te detecteren**

## confusion matrix

gaat een student  
wel of niet **het eerste  
studiejaar halen?**

|                     |          | voorspelde uitkomst |          |                  |
|---------------------|----------|---------------------|----------|------------------|
|                     |          | Positief            | Negatief |                  |
| werkelijke uitkomst | Positief | 80                  | 20       | 100              |
|                     | Negatief | 20                  | 80       | 100              |
|                     |          | 100                 | 100      | 80% accuraatheid |

**een goede prognose:** 80% is correct voorspeld  
maar was het ook een **eerlijke prognose?**

# criteria om **bias te detecteren**

## confusion matrix en geslacht

gaat een student  
wel of niet **het eerste**  
**studiejaar halen?**

|                     |          | voorspelde uitkomst |          |     |
|---------------------|----------|---------------------|----------|-----|
|                     |          | Positief            | Negatief |     |
| werkelijke uitkomst | Positief | 35                  | 15       | 100 |
|                     | Negatief | 15                  | 35       | 100 |
|                     |          | 100                 | 100      | 70% |

|                     |          | voorspelde uitkomst |          |     |
|---------------------|----------|---------------------|----------|-----|
|                     |          | Positief            | Negatief |     |
| werkelijke uitkomst | Positief | 45                  | 5        | 100 |
|                     | Negatief | 5                   | 45       | 100 |
|                     |          | 100                 | 100      | 90% |



een goede prognose, maar **geen eerlijke prognose**  
mannen worden benadeeld ten opzichte van vrouwen



# criteria om bias te detecteren

## precisie en gevoeligheid

gaat een student  
wel of niet **het eerste  
studiejaar halen?**

**werkelijke** uitkomst

**voorspelde** uitkomst

|                     |          | voorspelde uitkomst |          |
|---------------------|----------|---------------------|----------|
|                     |          | Positief            | Negatief |
| werkelijke uitkomst | Positief | 30                  | 10       |
|                     | Negatief | 20                  | 40       |
|                     |          | 50                  | 50       |

zo min mogelijk  
studenten die  
onterecht afhaken

$$\frac{30}{40} = 75\% \text{ gevoeligheid}$$

70% accuraatheid

|    |    |
|----|----|
| TP | FN |
| FP | TN |

zo min mogelijk  
studenten die onterecht  
doorgaan

$$\frac{30}{50} = 60\% \text{ precisie}$$

# criteria om **bias te detecteren**

## criteria in group fairness & classificatie

### onafhankelijkheid

*independence*

**een uitkomst is onafhankelijk van een sensitief kenmerk**

na het eerste studiejaar

zijn er **evenveel positieve als negatieve resultaten** per groep

gelijke uitkomsten

maatschappelijk belang

### scheiding

*separation – equality of errors*

**elke groep heeft evenveel kans op zo min mogelijk fouten**

als een algoritme voorspelt 'deze student zal waarschijnlijk het eerste jaar halen'

zijn er **evenveel fouten** per groep

zo min mogelijk studenten die onterecht doorgaan

belang van de onderwijsinstelling

### toereikendheid

*sufficiency - calibration*

**elke groep heeft evenveel kans om gevonden te worden**

als een algoritme voorspelt 'deze student zal waarschijnlijk het eerste jaar halen'

is dit **even vaak juist** per groep

zo min mogelijk studenten die onterecht afhaken

belang van de individuele student



# criteria om bias te detecteren

## criteria in group fairness & classificatie

onafhankelijkheid

*independence*

scheiding

*separation – equality of errors*

toereikendheid

*sufficiency - calibration*

het is niet mogelijk om aan alle drie de criteria tegelijkertijd te voldoen  
bias balanceren is cruciaal = pariteit

# voorspelmodellen uitleggen en bias detecteren met sensitieve data

algoritmes / voorspelmodellen + predict / explain  
lineaire regressie + random forest + gbm  
(R)MSE + confusion matrix + ROC + AUC

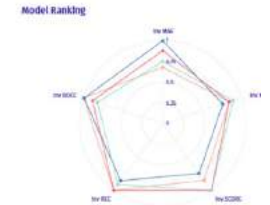
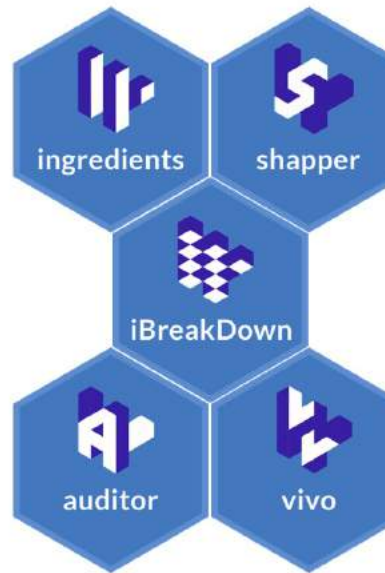
# voorspelmodellen uitleggen en bias detecteren met sensitieve data



model

explainer

uitleg

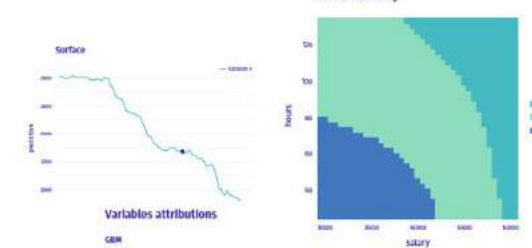


Factor Merger

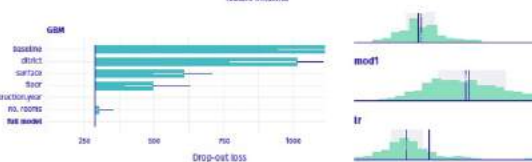
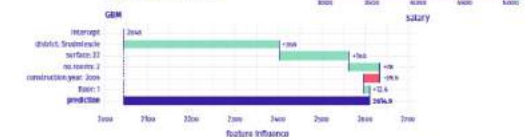
Random Forest



Hours vs salary

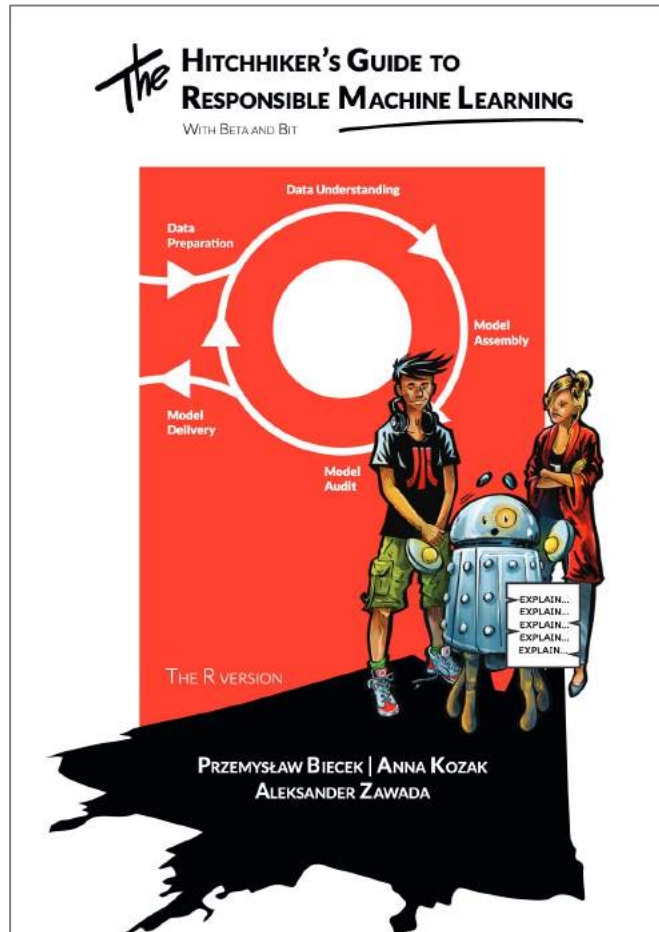


Variables attributions



Zie <https://drwhy.ai>.

# voorspelmodellen uitleggen en bias detecteren met sensitieve data



Model Audit

Model Performance

Depending on the type of predictive problem and what we assume about the distribution of the outcome, various measures of model performance can be used. Here is a short summary; find a more detailed description in the EMA book.

For regression problems, when we predict a quantitative variable, especially when we assume Gaussian noise, commonly used performance measures are Mean Squared Error<sup>15</sup> and Rooted Mean Squared Error<sup>16</sup>.

For binary classification problems, the outcome is commonly summarized with a  $2 \times 2$  contingency table with possible results coded as True Positive, True Negative, False Positive, and False Negative. Positive means that the test suggests a pregnancy, while Negative means no pregnancy. True and False describe whether the test result is correct or not. Below is an example of such a table for a simple „morning sickness” test for the pregnancy.

| Morning sickness / pregnancy | Pregnant                   | Not pregnant      |                   |
|------------------------------|----------------------------|-------------------|-------------------|
| Has sickness                 | TP = 39                    | FP = 150          | PPV = Prc = 20.6% |
| Has not                      | FN = 61                    | TN = 850          | NPV = 93.3%       |
|                              | Sensitivity = Recall = 39% | Specificity = 85% | F1 = 33.8%        |

Based on such a contingency table, the most commonly used measures of performance are Accuracy<sup>17</sup>, Sensitivity<sup>18</sup>, Specificity<sup>19</sup>, Precision<sup>20</sup>, Recall<sup>21</sup>, F1 score<sup>22</sup>, Positive Predicted Value<sup>23</sup> and Negative Predicted Value<sup>24</sup>.

Note that in the covid-mortality-risk-assessment problem, we are not interested in the binary prediction survived / dead, but rather in the validity of the ranking of risk scores. For such types of problems, instead of a contingency table, one looks at Receiver Operating Characteristic (ROC) curve, and the commonly used measure of performance is the Area Under the ROC Curve (AUC). Figure 6 shows how this measure is constructed.

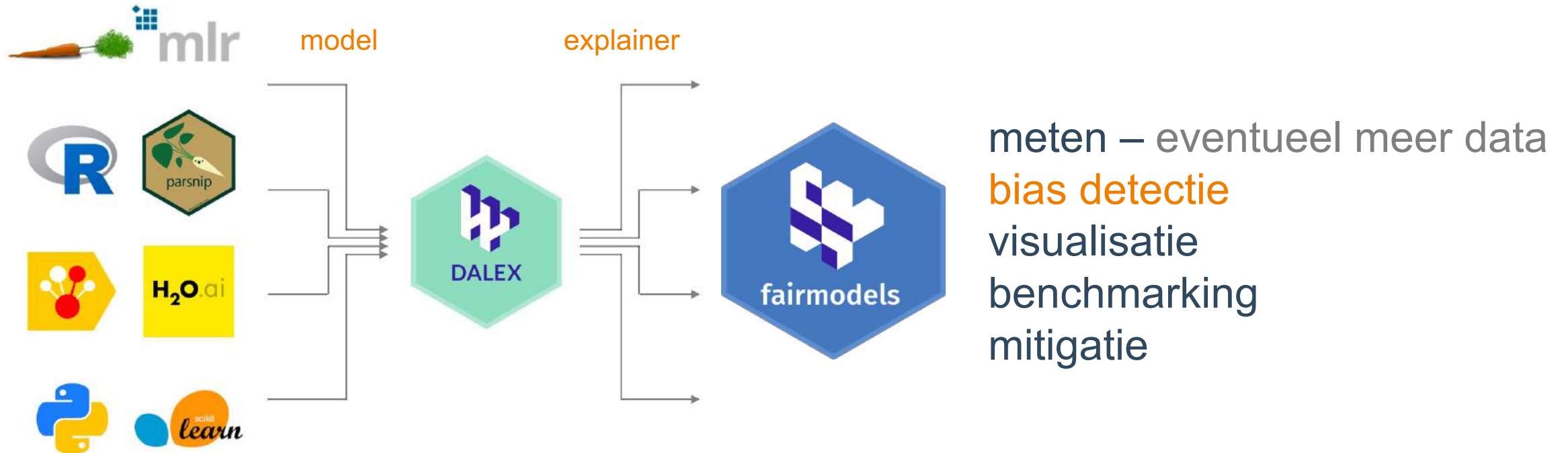
Figure 6: Panel A shows the distribution of scores obtained from the CDC model for the test data divided by the survival status. By taking different cutoffs, one can turn such numerical scores into binary decisions. For each such a split, the Sensitivity and specificity can be calculated and drawn on a plot. The CDC model returns only nine different values, making it reasonable to consider ten different cutoffs. Panel B shows these 10 points corresponding to different cutoffs. The ROC curve is the piecewise line connecting these points and the AUC is the area under this curve. The AUC takes values from 0 to 1, where 1 is the perfect ranking and a purely random ranking leads to the AUC of 0.5.

A) Distribution of scores among those who died

B) Receiver Operator Characteristic

Zie <https://betaandbit.github.io/RML/>

# voorspelmodellen uitleggen en bias detecteren met sensitieve data



# voorspelmodellen uitleggen en bias detecteren met sensitieve data



model: bias bij een negatief BSA  
lineaire regressie + random forest + gbm  
(R)MSE + confusion matrix + ROC + AUC  
op basis van de simulatie-dataset van het versnellingsplan

## 1. inlezen

- dataset: eerstejaars, hoofdinschrijving, voltijd, positief/negatief bsa
- variabelen: bsa, geslacht, leeftijd, buitenlands diploma, aansluiting, aanmeldingsdatum
- missende waarden verwijderd

## 2. explainer voor random forest

- geslacht (vrouw), buitenlands diploma (ja), aansluiting (direct na diploma)

## 3. detectie bias



# voorspelmodellen uitleggen en bias detecteren met sensitieve data



Preparation of a new explainer is initiated

```
-> model label      : ranger ( default )
-> data             : 24744 rows 5 cols
-> target variable  : 24744 values
-> predict function : yhat.ranger will be used ( default )
-> predicted values : No value for predict function target column. ( default )
-> model_info       : package ranger , ver. 0.15.1 , task classification ( default )
-> predicted values : numerical, min = 0.1488081 , mean = 0.8135431 , max = 0.9853193
-> residual function: difference between y and yhat ( default )
-> residuals        : numerical, min = -0.9675814 , mean = 0.0001491409 , max = 0.7531143
```

A new explainer has been created!

##

## 2.1.3 Maak een fairness object: Geslacht ####

```
fobject <- fairness_check(rf_explainer,
                          protected = dfSim$DEM_Geslacht,
                          privileged = "Vrouw",
                          cutoff = 0.5,
                          colorize = FALSE)
```

```
print(fobject)
```

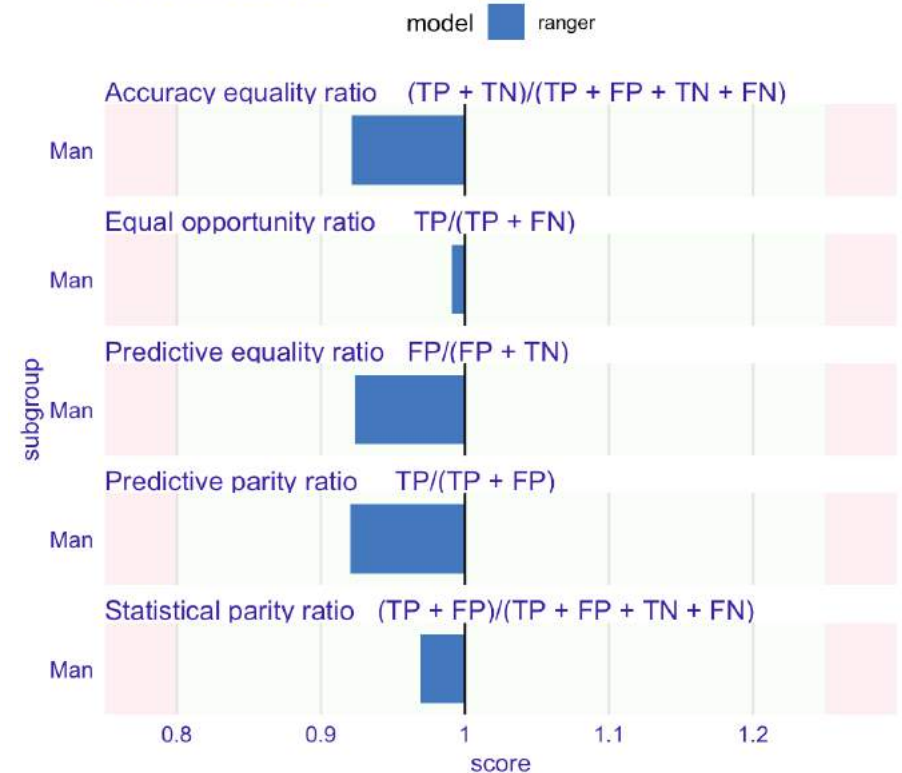
Fairness check for models: ranger

ranger passes 5/5 metrics

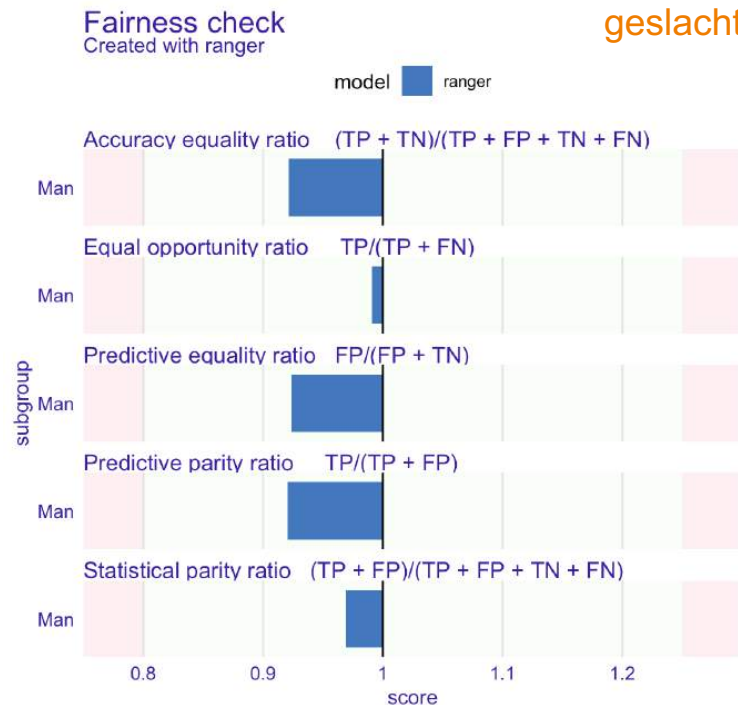
Total loss : 0.2747894

Fairness check  
Created with ranger

geslacht

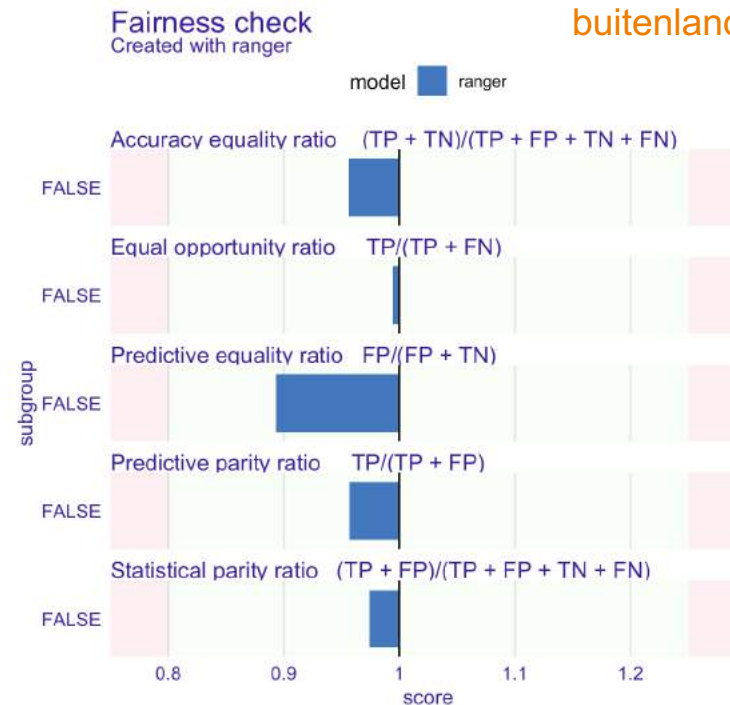


# voorspelmodellen uitleggen en bias detecteren met sensitieve data



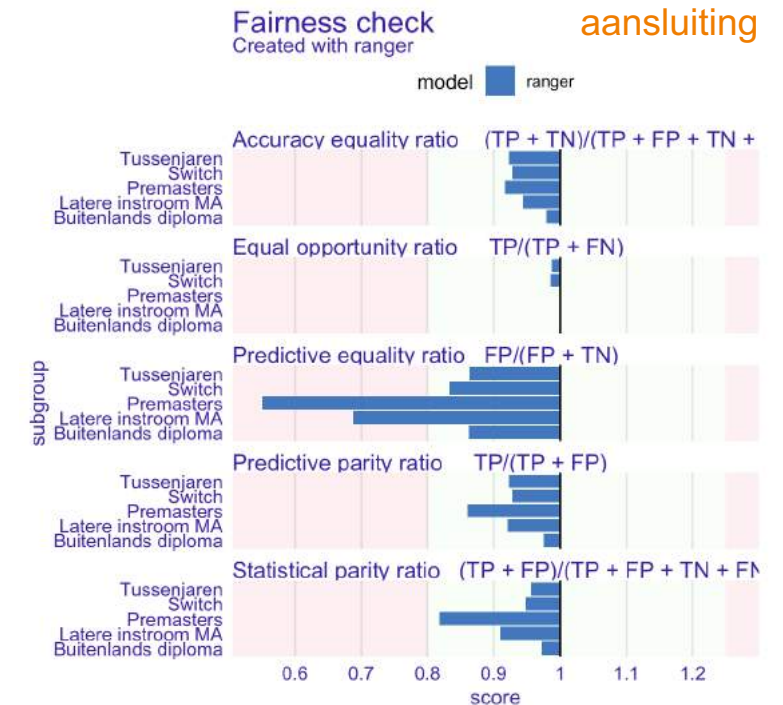
Fairness check for models: ranger

ranger passes 5/5 metrics  
Total loss : 0.2747894



Fairness check for models: ranger

ranger passes 5/5 metrics  
Total loss : 0.2254708



Fairness check for models: ranger

ranger passes 4/5 metrics  
Total loss : 2.329799



# voorspelmodellen uitleggen en bias detecteren met sensitieve data



model: bias bij een negatief bsa

lineaire regressie + classification & regression trees + random forest

(R)MSE + confusion matrix + ROC + AUC

op basis van de simulatie-dataset van het versnellingsplan

## 4. meerdere modellen toepassen

- voeg modellen in aangepaste vorm (minder variabelen)
- ander soorten modellen (lineaire regressie, general boosted model)

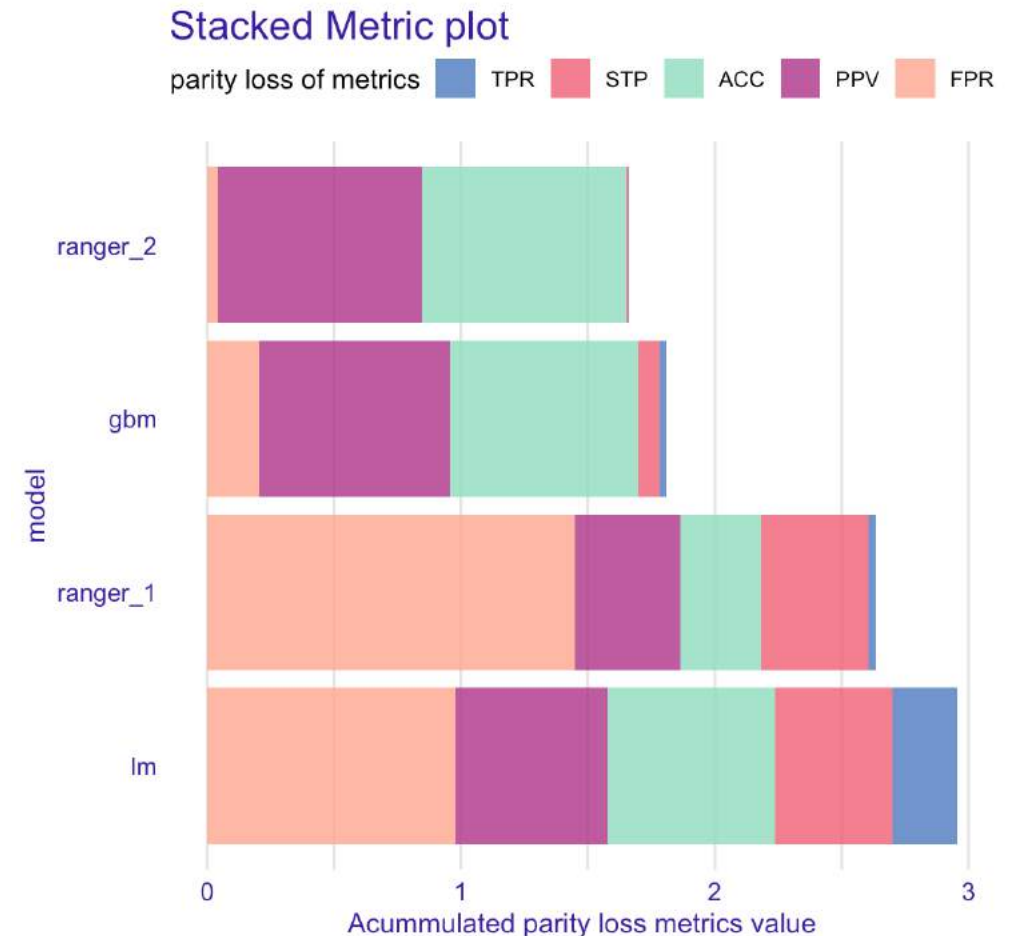
## 5. maak een keuze voor het beste model

# voorspelmodellen uitleggen en bias detecteren met sensitieve data



```
## -----  
## 2.2.3 Maak een fairness object #####  
  
fobject <- fairness_check(explainer_1, explainer_2,  
                          explainer_3, explainer_4,  
                          protected = dfSim$INS_Aansluiting_cat,  
                          privileged = "Direct na diploma",  
                          verbose = TRUE)  
  
## ~~~~~  
## 2.3 Kies het beste model #####  
  
## -----  
## 2.3.1 Toon de metric scores (stack) #####  
  
sm <- stack_metrics(fobject)  
plot(sm)
```

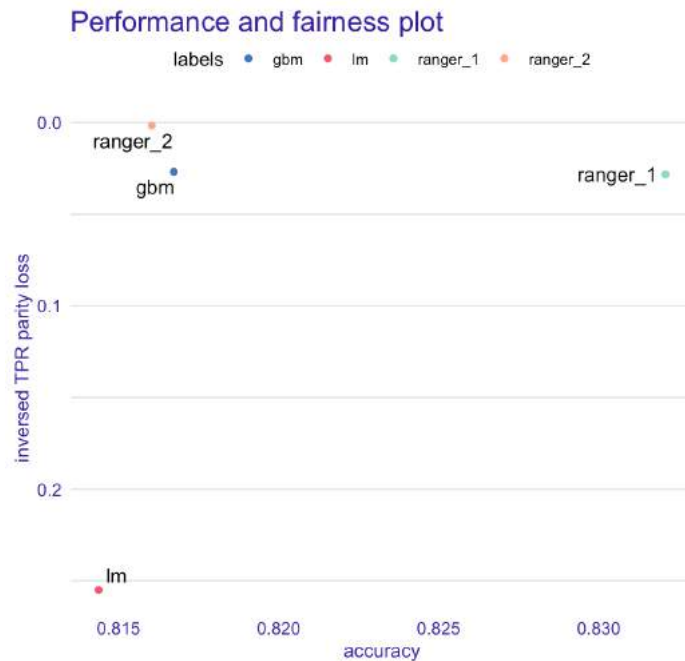
het ranger 2 model heeft de beste kaarten  
met name op de false positive rate (FPR = precisie)



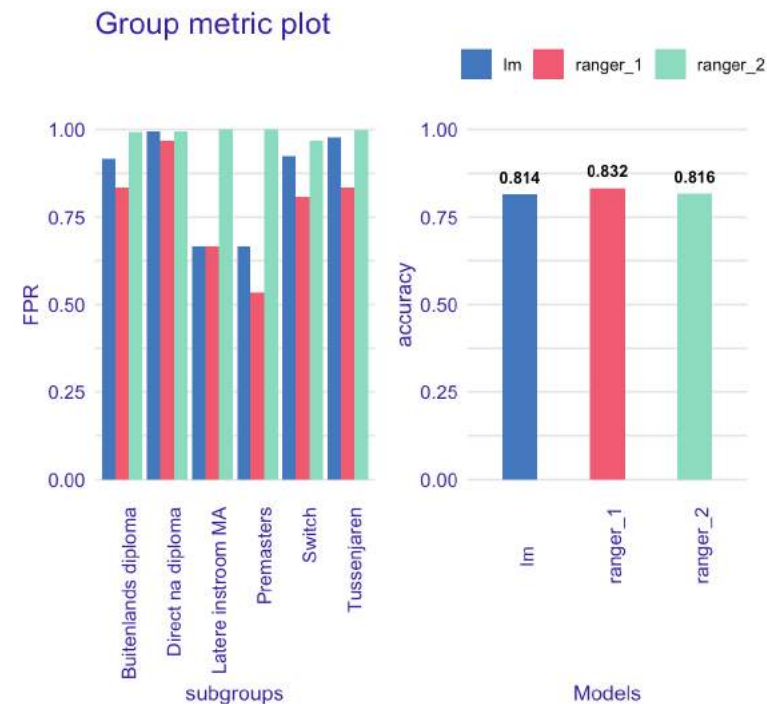
# voorspelmodellen uitleggen en bias detecteren met sensitieve data



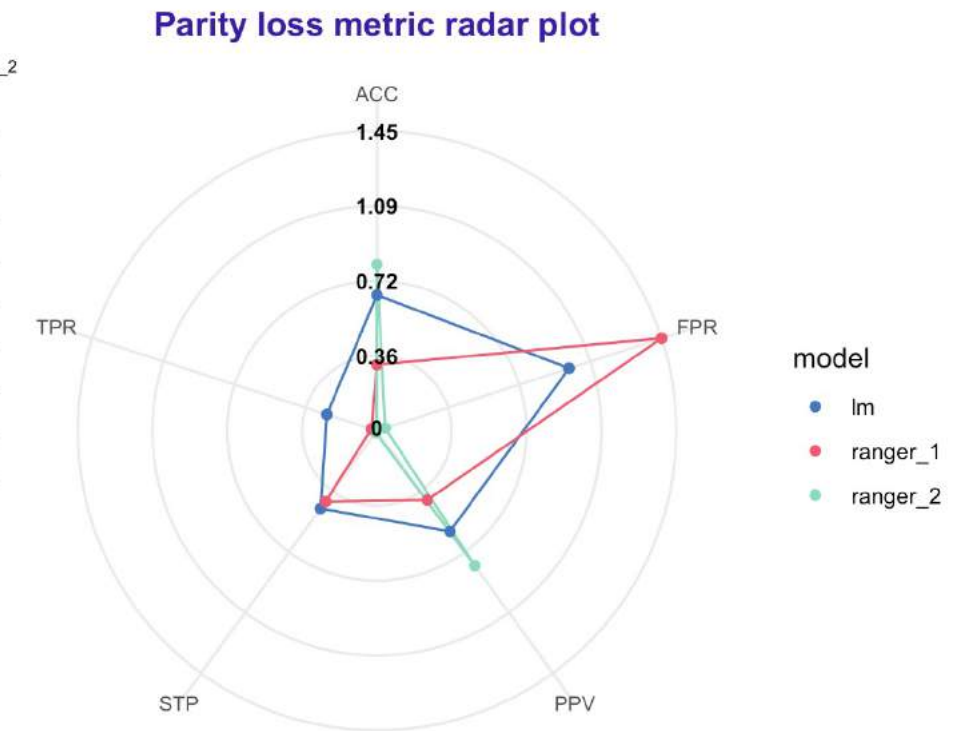
## performance per model



## performance per metriek



## benchmark per metriek



via een aantal benchmarks is dit verder te onderbouwen  
er is een iets lagere accuraatheid, maar het model is eerlijker voor premasterstudenten



## voorspelmodellen uitleggen en bias detecteren met sensitieve data

### valkuilen & hersenbrekers

na het balanceren voor 1 attribuut, moet je overige attributen **opnieuw testen**  
modellen moeten eerst voldoen aan specifieke **assumptions**  
wat een juiste balans is, is soms een **politieke** keuze  
balans in de praktijk brengen, kan **ongelijke behandeling** betekenen  
**om tot gelijke uitkomsten te komen**

# Referenties

*Code of Federal Regulations. Section 4d, uniform guidelines on employee selection procedures* (1978)

Wiśniewski, J., & Biecek, P. (2021). fairmodels: A Flexible Tool For Bias Detection, Visualization, And Mitigation. *arXiv*. doi: 10.48550/arxiv.2104.00507

# Creative commons licentie

## Naamsvermelding-NietCommercieel-GelijkDelen 4.0 Internationaal (CC BY-NC-SA 4.0)

### Je bent vrij om:

- **het werk te delen** — te kopiëren, te verspreiden en door te geven via elk medium of bestandsformaat
- **het werk te bewerken** — te remixen, te veranderen en afgeleide werken te maken

De licentiegever kan deze toestemming niet intrekken zolang aan de licentievoorwaarden voldaan wordt.



### Onder de volgende voorwaarden:

- **Naamsvermelding** — De gebruiker dient de maker van het werk te [vermelden](#), een link naar de licentie te plaatsen en [aan te geven of het werk veranderd is](#). Je mag dat op redelijke wijze doen, maar niet zodanig dat de indruk gewekt wordt dat de licentiegever instemt met je werk of je gebruik van het werk.
- **NietCommercieel** — Je mag het werk niet gebruiken voor [commerciële doeleinden](#).
- **GelijkDelen** — Als je het werk hebt geremixt, veranderd, of op het werk hebt voortgebouwd, moet je het veranderde materiaal verspreiden onder [dezelfde licentie](#) als het originele werk.
- **Geen aanvullende restricties** — Je mag geen juridische voorwaarden of [technologische voorzieningen](#) toepassen die anderen er juridisch in beperken om iets te doen wat de licentie toestaat.

**let's change**