Bias/ Fairness Evaluatie

Het evalueren van een model kan op verschillende manieren. Denk aan accuracy, precision of recall score. Toch is dit niet altijd betrouwbaar.

Stel je voor dat je een dataset hebt van 1000 patiënten (500 man en 500 vrouw). We willen voorspellen of een patiënt kankercellen in het lichaam heeft of niet. We krijgen de volgende confusion matrix:

True Positives (TPs): 16	False Positives (FPs): 4
False Negatives (FNs): 6	True Negatives (TNs): 974

Precision: 16 / (16 + 4) = 0.8
Recall : 16 /(16 + 6) = 0.73

Een precision score van 0.8! De scores zien er prima uit toch? Wat als we even de mannen en vrouwen apart bekijken.

Female Patient Results

True Positives (TPs): 10	False Positives (FPs): 1
False Negatives (FNs): 1	True Negatives (TNs): 488

Precision: 10 / (10 + 1) = 0.91
Recall : 10 /(10 + 1) = 0.91

Male Patient Results

True Positives (TPs): 6	False Positives (FPs): 3
False Negatives (FNs): 5	True Negatives (TNs): 486

Precision: 6 / (6 + 3) = 0.667
Recall : 6 / (6 + 5) = 0.545

Conclusie

De scores voor vrouwen komen véél hoger uit dan die van de mannen. De 0.8 score zegt dus niet heel veel. Het gecombineerde model doet het dus beter voor vrouwen dan voor mannen. Het is dus biased en dus niet al te betrouwbaar.