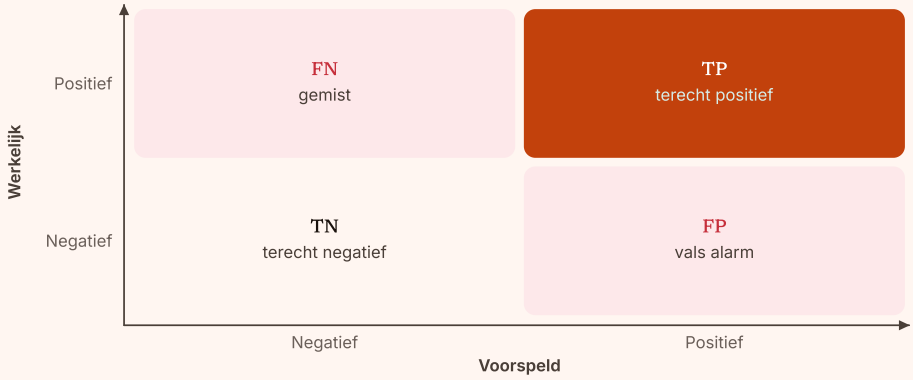
Logistische regressie, overfitting en validatie

CHEATSHEET

Tentamen ma 7 dec 2026

TENTAMENKANS Logistische regressie (logitmodel) **98%** · Confusion matrix, accuracy en sensitivity berekenen **97%** · Accuracy is misleidend bij scheve klassen **95%** · Logitcoëfficiënt is een verandering in log-odds, geen kans **94%** · Odds en odds ratio (e^b) **93%** · Overfitting en generalisatie **92%**

Confusion matrix



Overfitting voorkomen met validatie

1. Splits in train en test
2. Fit op de train-set
3. Voorspel de test-set
4. Vergelijk train- en testprestatie

M 01 Logit (formuleblad)

- odds = p/(1 – p); log-odds z = ln(p/(1 – p)) = b0 + b1 ×1 + …
- p = 1/(1 + exp(-z)); odds ratio = exp(b)
- b = verandering in log-odds, **niet** in procentpunten
- Meer eenheden: exp(n b), niet n maal de odds ratio

M 02 R

- **glm(y ~ x1 + x2, family = binomial, data = d)**
- **predict(fit, newdata, type = "response")** geeft kansen; zonder type log-odds
- **exp(coef(fit))**, **exp(confint.default(fit))** voor odds ratios met Wald-interval (zoals in de app); **confint(fit)** geeft het profile-interval, iets anders
- **table(Voorspeld, Waargenomen)** voor de confusion matrix
- **anova(m1, m2, test = "Chisq")** voor geneste modellen
- ROCR: **prediction()**, **performance()** voor ROC-curve en AUC

M 03 Output lezen

- **z value** en **Pr(>|z|)**: kleiner dan 0,05 is significant (teken en significantie)
- Null deviance (alleen intercept) en residual deviance (model); df = n – k
- AIC = 2k – 2 ln(L): lager is beter
- McFadden pseudo-R^2 = 1 – ln L(model)/ln L(null) (geen verklaarde variantie)

M 04 Confusion matrix (threshold τ)

	Werkelijk 0	Werkelijk 1
Voorspeld 0	TN	FN
Voorspeld 1	FP	TP

Let op de volgorde. R (**table**) zet 0 eerst: TN linksboven. Het oefententamen 7.1 zet 1 eerst: TP linksboven, FN rechtsboven, FP linksonder, TN rechtsonder. Lees eerst de kop van rijen en kolommen voordat je TP, FN, FP en TN aanwijst.

- accuracy = (TP + TN)/n
- sensitivity = TP/(TP + FN); specificity = TN/(TN + FP)
- precision = TP/(TP + FP)
- Lagere threshold: sensitivity omhoog, specificity omlaag
- AUC: 0,5 is gokken, 1 is perfect

M 05 Beslisregels

- Scheve klassen: vergelijk accuracy met de baseline (altijd meerderheidsklasse); kijk naar sensitivity en precision
- Hoge kosten voor een gemist geval: lagere threshold
- Overfitting: train veel beter dan test; kies het eenvoudiger model
- Beoordeel op een test set of met cross-validation, nooit op de trainingsdata alleen
- Flexibeler model: minder bias, meer variantie

M 06 Vraag 7.1 (9 p, 200 woorden)

1. Coëfficiënten: teken, significantie, odds ratio, ceteris paribus
2. Berekening met formule en getallen
3. Waarom accuracy misleidt: scheve verdeling, lage sensitivity, gevolg voor het bedrijf (threshold of overfitting alleen als de casus erom vraagt)

! Valkuilen

- p = 0,06 is geen 6% kans op H0
- "2,75 keer zoveel kans" is fout: het zijn odds
- 0,5 is een gewoonte, geen regel
- p-value zegt niets over de grootte van het effect
- MLE-eigenschappen gelden voor grote n