SUPERVISED LEARNING - CLASSIFICATION

JENS BAETENS

GLOSSARY

- Supervised
- Unsupervised
- Reinforcement Learning
- Regression
- Overfitting
- Underfitting
- Learning Rate
- Loss Function
- Feature Engineering
- Normalisation
- Regularisation
- Trainen van een model

Categorieën → gender
→ age-class

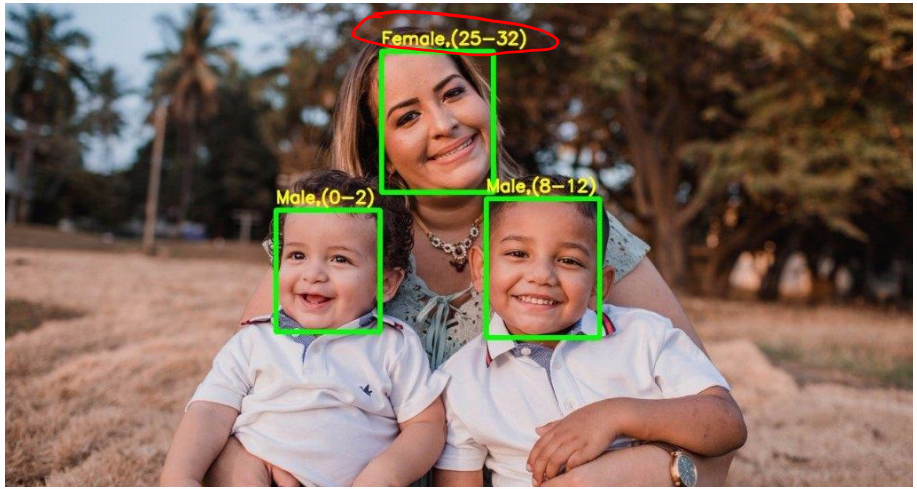
WAT IS CLASSIFICATIE?

Supervised learning

Input omzetten naar klasse

Classifier genoemd

Model



WAT IS CLASSIFICATIE?

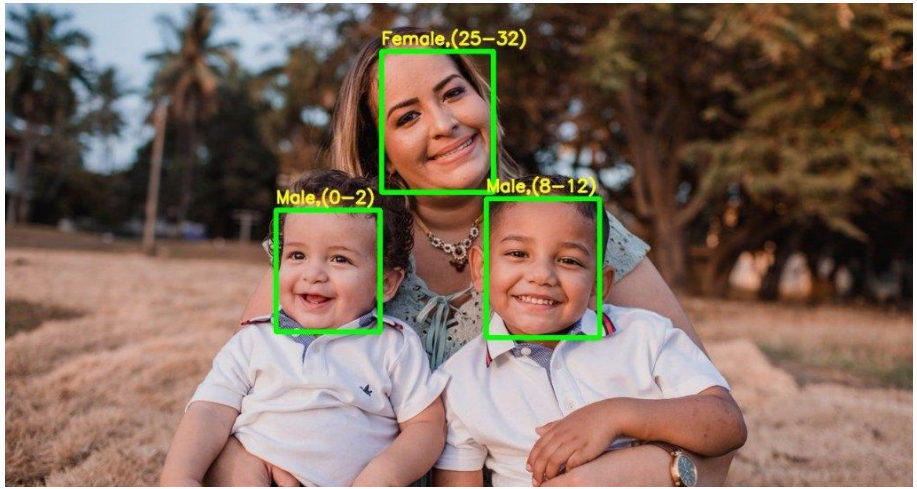
Gezichtsherkenning

Geschriftherkenning

Spam detectie

Kwaliteitscontroles

Medische diagnoses



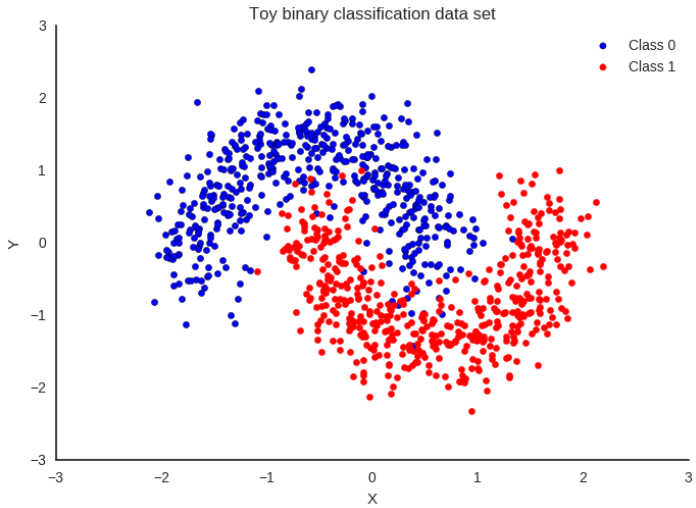
TYPES CLASSIFIERS - BINARY

True / False

Twee verschillende klassen

Voorbeeld: Goede of slechte kwaliteit, man of vrouw, Goed- of kwaadaardig

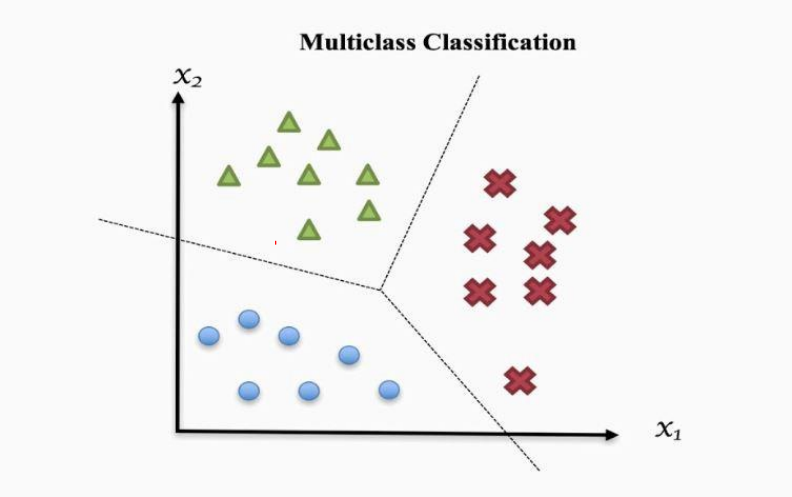
Basis voor de andere



TYPES CLASSIFIERS - MULTICLASS

N>2 verschillende klassen (maar 1 mogelijk voor elke input)

Voorbeeld: Gezichtsherkenning (1 klasse per persoon), Hondenrasherkenning, …



TYPES CLASSIFIERS - MULTILABEL

N>2 verschillende klassen maar meerdere mogelijk per input

Voorbeeld: Beeldherkenning, Meerdere genres mogelijk voor een film, …

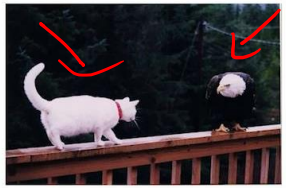
Binary Classification

- Spam
- Not spam

Multiclass Classification

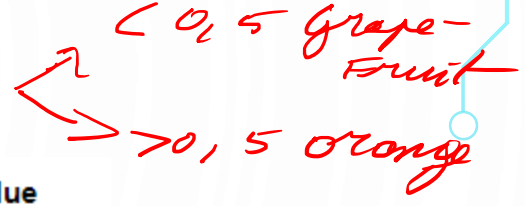
- Dog
- Cat
- Horse
- Fish
- Bird
- …

Multi-label Classification



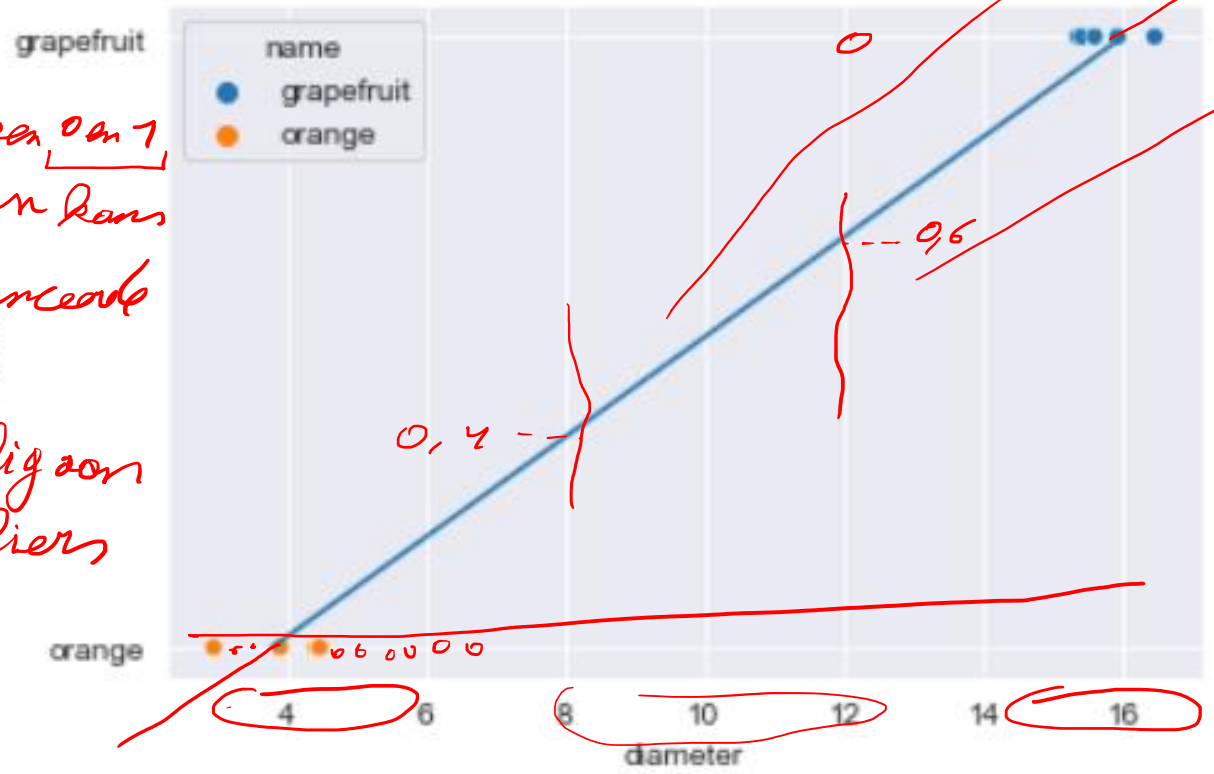
- Dog
- Cat
- Horse
- Fish
- Bird
- …

KAN HET MET LINEAIRE REGRESSIE?



		name	diameter	weight	red	green	blue
name							
grapefruit	**9995**	grapefruit	15.35	253.89	149	77	20
	9996	grapefruit	15.41	254.67	148	68	7
	9997	grapefruit	15.59	256.50	168	82	20
	9998	grapefruit	15.92	260.14	142	72	11
	9999	grapefruit	16.45	261.51	152	74	2
orange	**0**	orange	2.96	86.76	172	85	2
	1	orange	3.91	88.05	166	78	3
	2	orange	4.42	95.17	156	81	2
	3	orange	4.47	95.60	163	81	4
	4	orange	4.48	95.76	161	72	9

KAN HET MET LINEAIRE REGRESSIE?

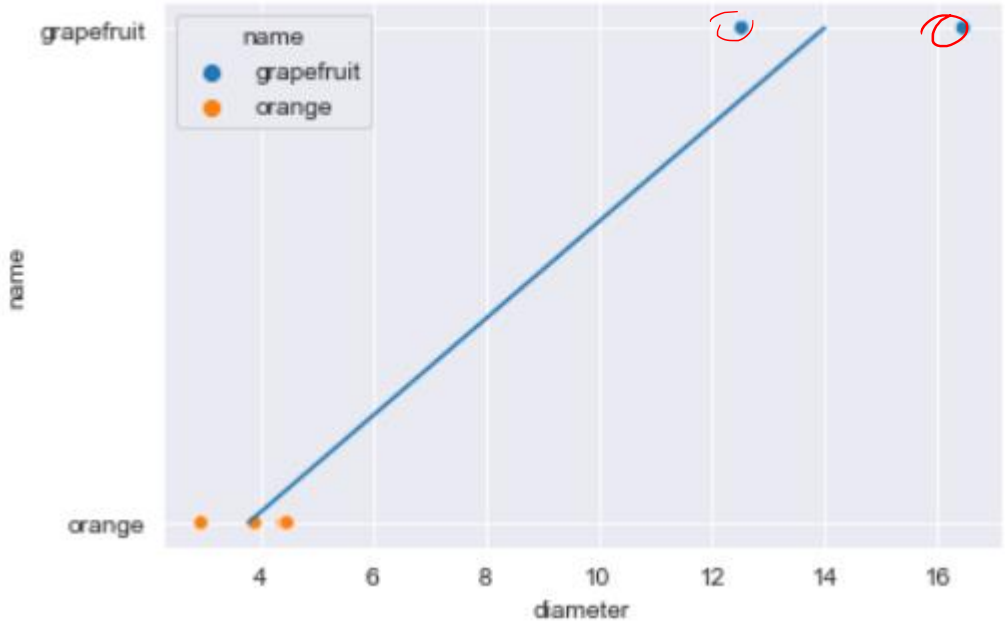


KAN HET MET LINEAIRE REGRESSIE?

Gevoelig voor outliers

Zeer breed “fuzzy” middenstuk

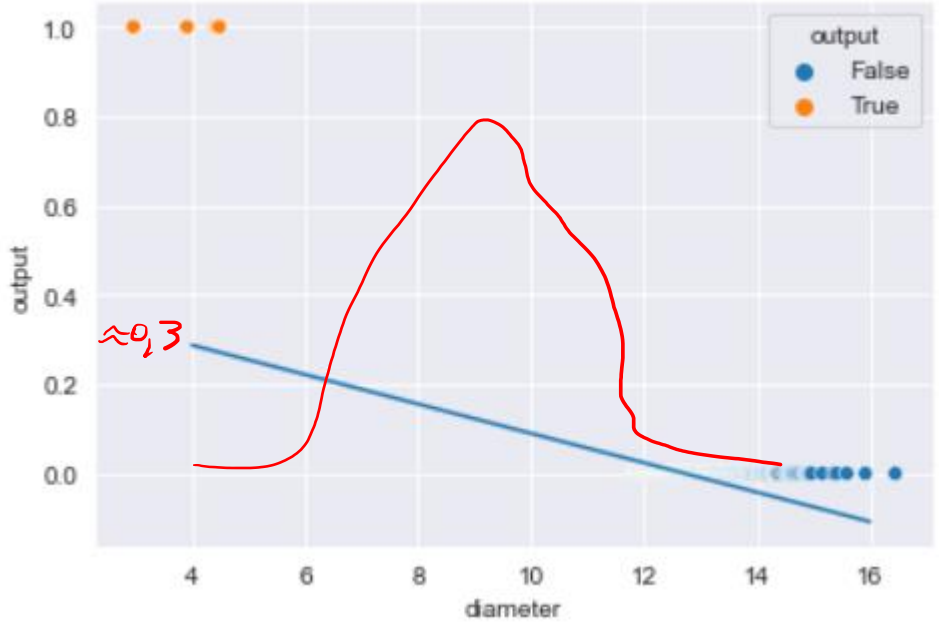
Komt niet overeen met een kans



KAN HET MET LINEAIRE REGRESSIE?

Ongebalanceerde klassen

=> Geen lineaire regressie mogelijk



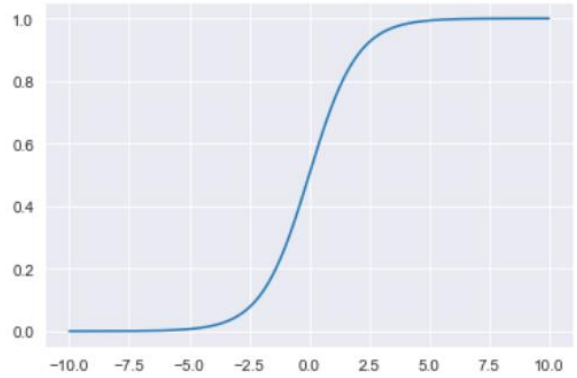
CLASSIFICATIE – LOGISTIC REGRESSION

JENS BAETENS

LOGISTIC REGRESSION

Logistische functie (sigmoid)

$$f(z) = \frac{1}{1+e^{-z}}$$



Geeft een waarde terug tussen 0 en 1

- De kans het tot de klasse hoort

$$f_{\boldsymbol{w}}(x) = \frac{1}{1+e^{-\boldsymbol{w}^T x}}$$

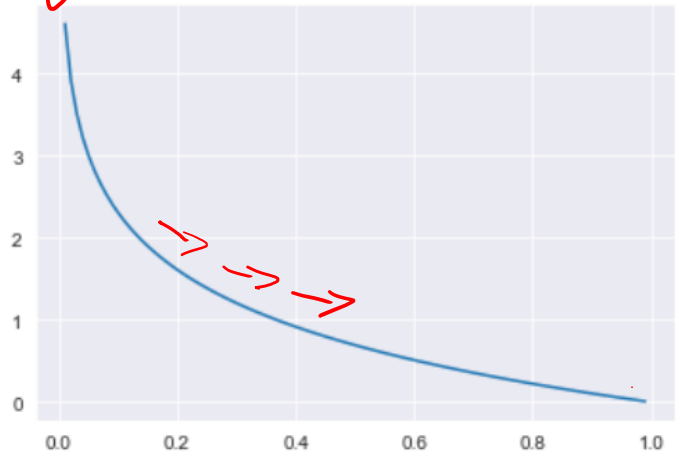
$$\boldsymbol{w}^T x = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_N x_N$$ ≈ Lineaire regressie

LOGISTIC REGRESSION

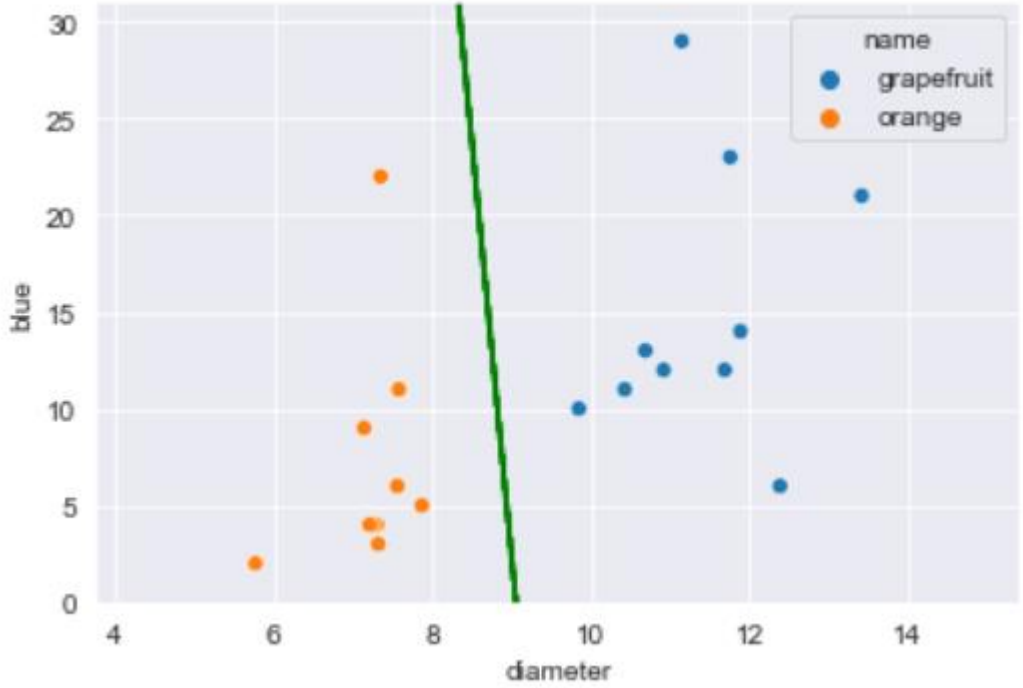
$$L(\boldsymbol{w}) = \begin{cases} -ln(f_{\boldsymbol{w}}(x)) & \text{als } y = 1 \\ -ln(1 - f_{\boldsymbol{w}}(x)) & \text{als } y = 0 \end{cases}$$

$$L(\boldsymbol{w}) = -\frac{1}{N}\left[\sum_{i=1}^{N} y_i ln(f_{\boldsymbol{w}}(x_i)) + (1 - y_i) ln(1 - f_{\boldsymbol{w}}(x_i))\right]$$

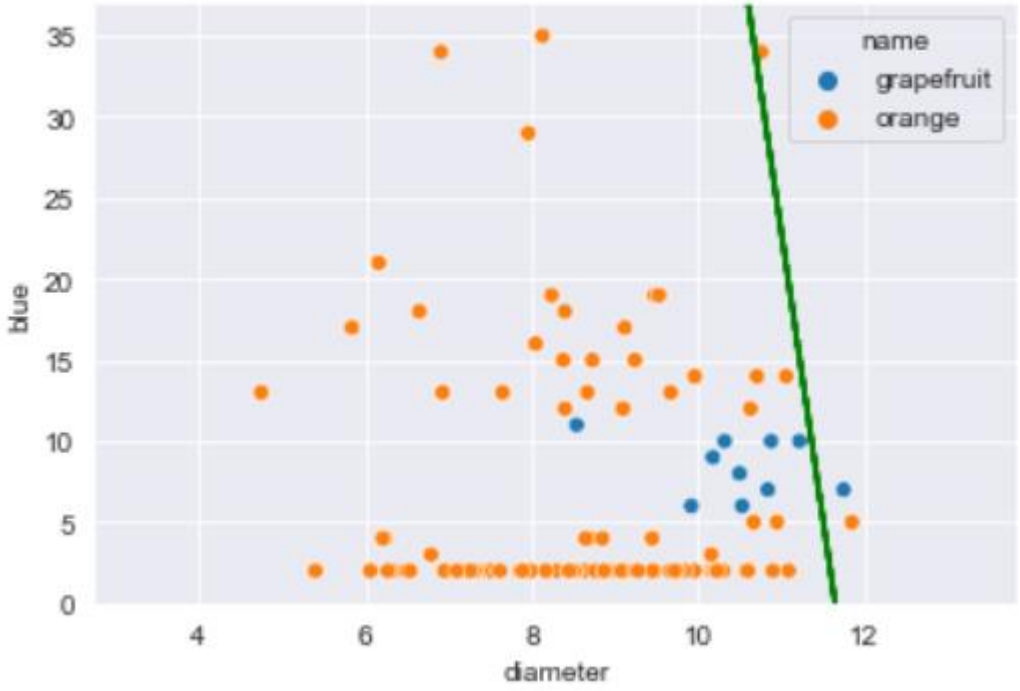
Minimalisatie dmv Gradient Descent



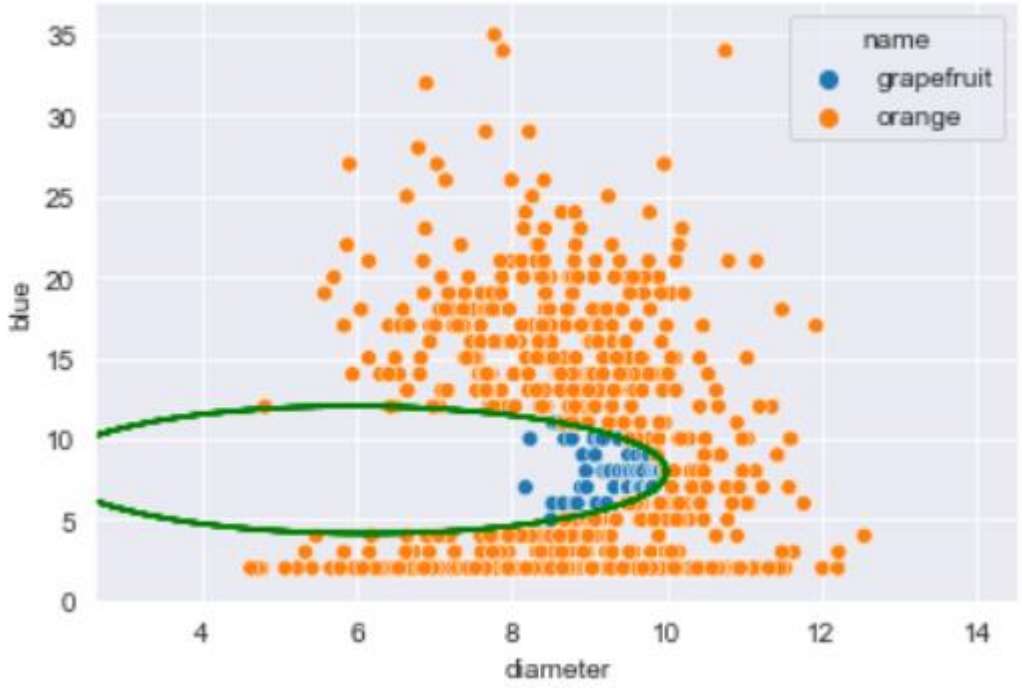
LOGISTIC REGRESSION



LOGISTIC REGRESSION – HIGHER ORDER FEATURES



LOGISTIC REGRESSION – HIGHER ORDER FEATURES

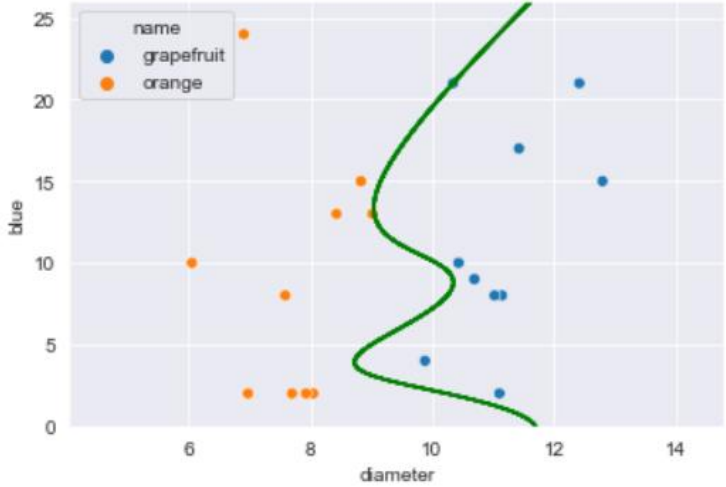


LOGISTIC REGRESSION – REGULARISATIE

Regularisatie via C-parameter

Inverse regularisatie sterkte

Hoge waarde = weinig regularisatie




```
model = LogisticRegression(C=10) # C= inverse regularisatiesterkte
model.fit(X, df_trimmed.output)
```

LOGISTIC REGRESSION – EVALUATIE

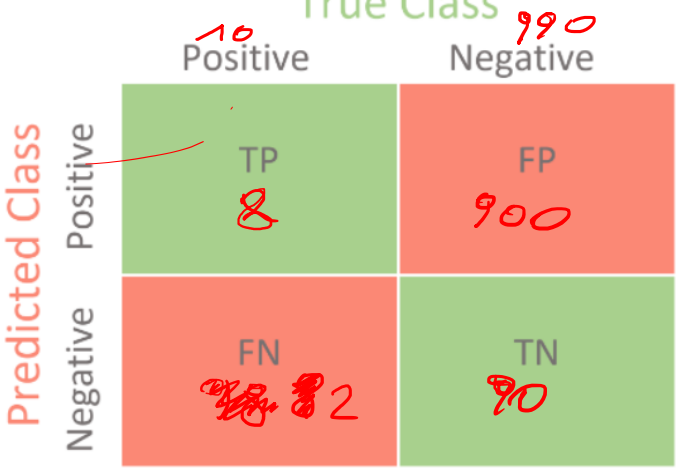
Accuraatheid $\frac{98}{100} = 98\%$

Precisie $\frac{TP}{TP+FP}$

Specificiteit $\frac{TN}{TN+FP}$

Recall 80% $\frac{TP}{TP+FN}$

F1-Score $2\frac{Precisie*Recall}{Precisie+Recall} = \frac{2TP}{2TP+FP+FN}$



https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

VOORBEELD CONFUSION MATRIX

Accuraatheid = 9/12

Sensitiviteit/Recall = 6/8

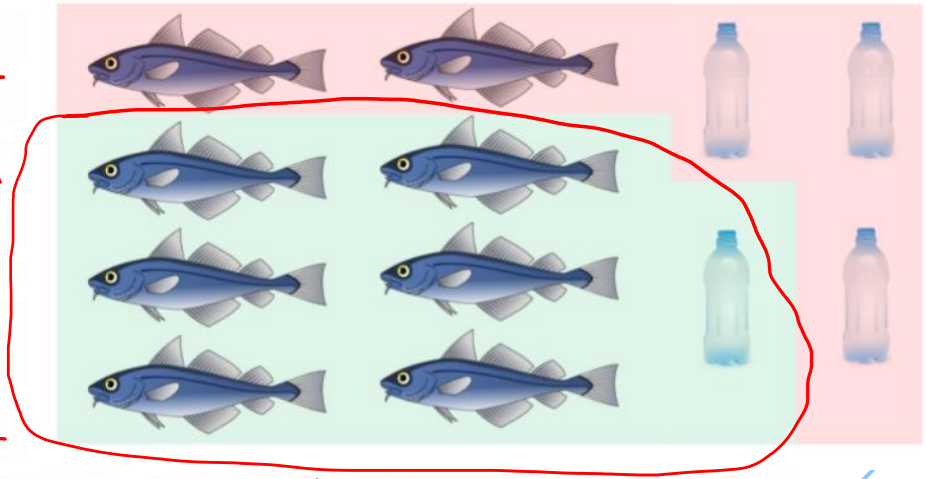
- Weinig positieve samples gemist

Specificiteit = 3/4

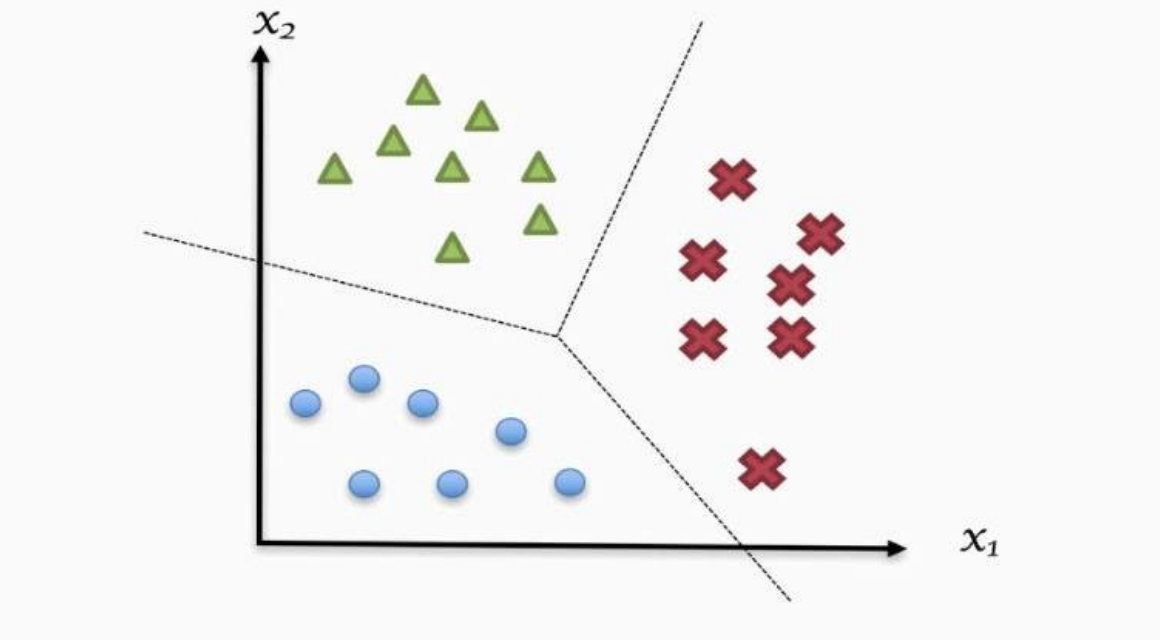
- Weinig negatieve samples gemist

Precision = 6/7

- Weinig negatieve samples als positieve geclassificeerd

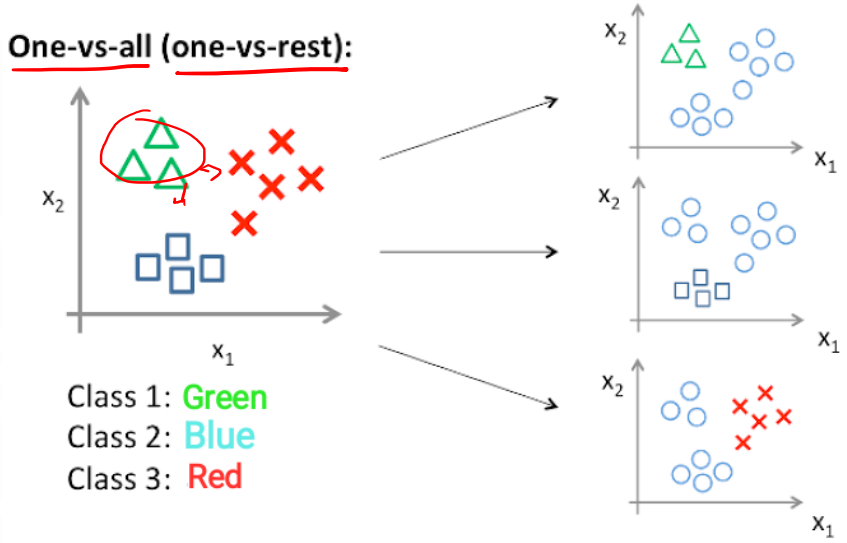


LOGISTIC REGRESSION – MULTICLASS



Multi-class classifier bestaat uit meerdere binaire classifiers

LOGISTIC REGRESSION – ONE VS ALL

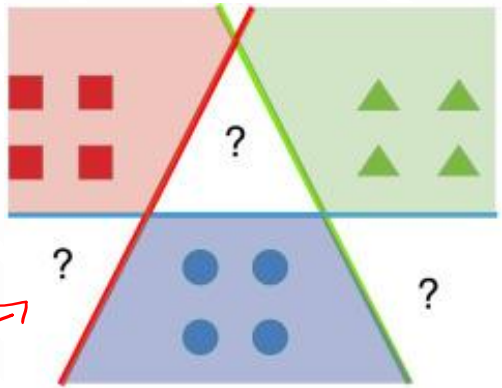


N classifiers nodig
1 / klasse

Conflicten mogelijk
→ Kies methode
hoogste kans

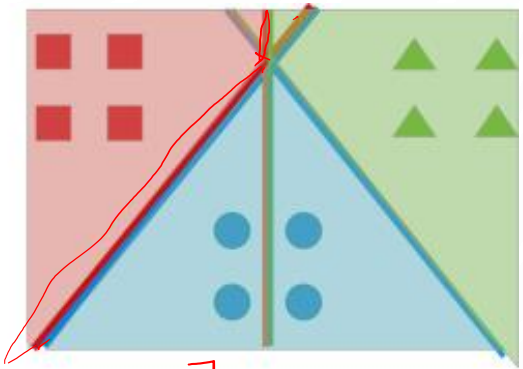
LOGISTIC REGRESSION – ONE VS ONE

One vs all → One vs One



Conflicten

Is het rood?



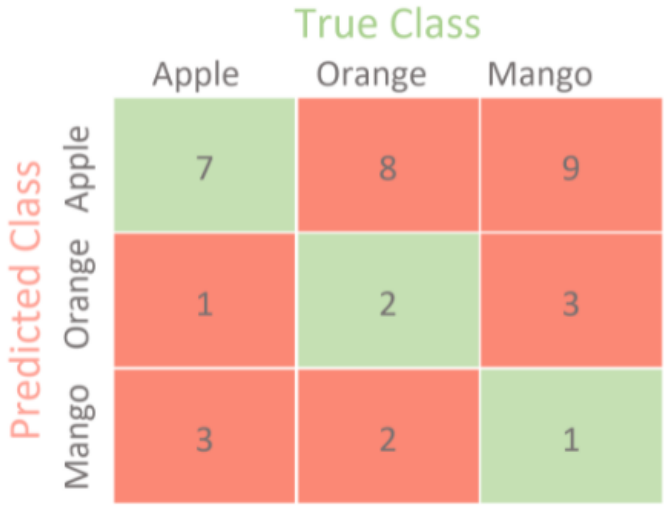
Is het niet rood?

→ R–B
→ R–G
→ G–B

N → N(N−1) classifiers

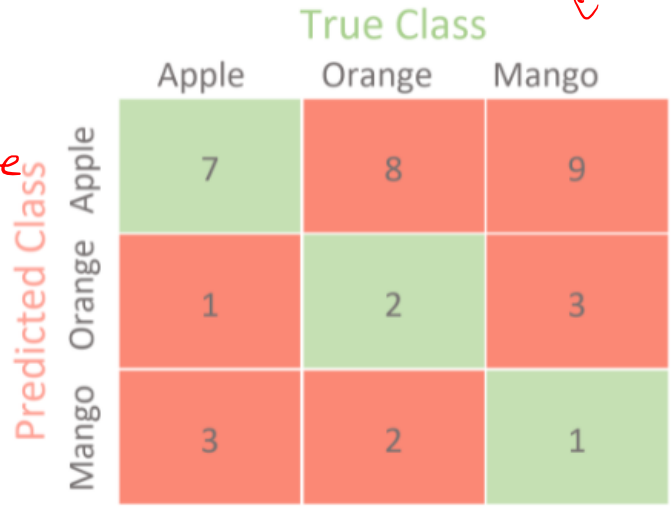
moeilijk → training tijd is langer.

LOGISTIC REGRESSION – MULTICLASS EVALUATIE



https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

LOGISTIC REGRESSION – MULTICLASS EVALUATIE



Precisie →

↓ Recall

Micro → telkle klasse op

Macro

Weighted

Class	Precision	Recall	F1-score
Apple	7/24 0.29	0.64	0.40
Orange	0.33	0.17	0.22
Mango	0.17	0.08 1/12 3	0.11

Micro 10/36 10/36 10/36

Macro $\frac{0,29+0,33+0,17}{3}$ $\frac{0,64+0,17+0,08}{3}$ ...

Weighted $\frac{0,29\cdot\frac{11}{36}+6\cdot 0,33+\frac{13}{36}\cdot 0,17}{36}$

https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

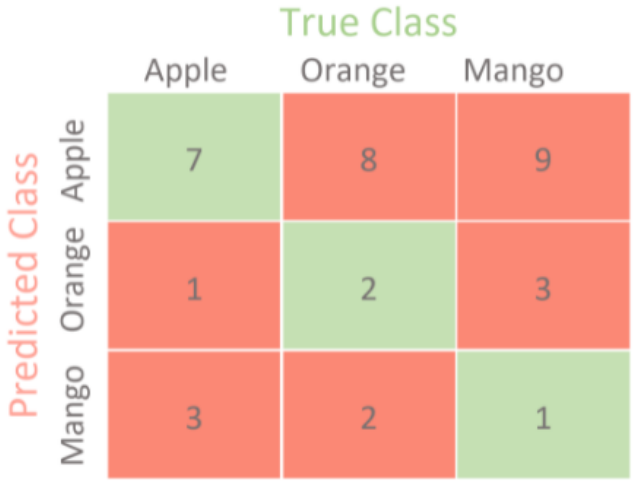
LOGISTIC REGRESSION – MULTICLASS EVALUATIE

Micro – F1: Globale waarden

Macro – F1: Gemiddelde F1 – scores

Weighted F1: Gew. Gemiddelde

- Gewichten = # samples



Class	Precision	Recall	F1-score
Apple	0.29	0.64	0.40
Orange	0.33	0.17	0.22
Mango	0.17	0.08	0.11

in de data
niet in de voorspellingen

https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

GLOSSARY

- Classificatie
- Binary classifier
- Multi-class classifier
- Multi-label classifier
- True/False Positive/Negative
- Accuraatheid / Specificiteit / …
- One-vs-All
- One-vs-One
- Confusion matrix