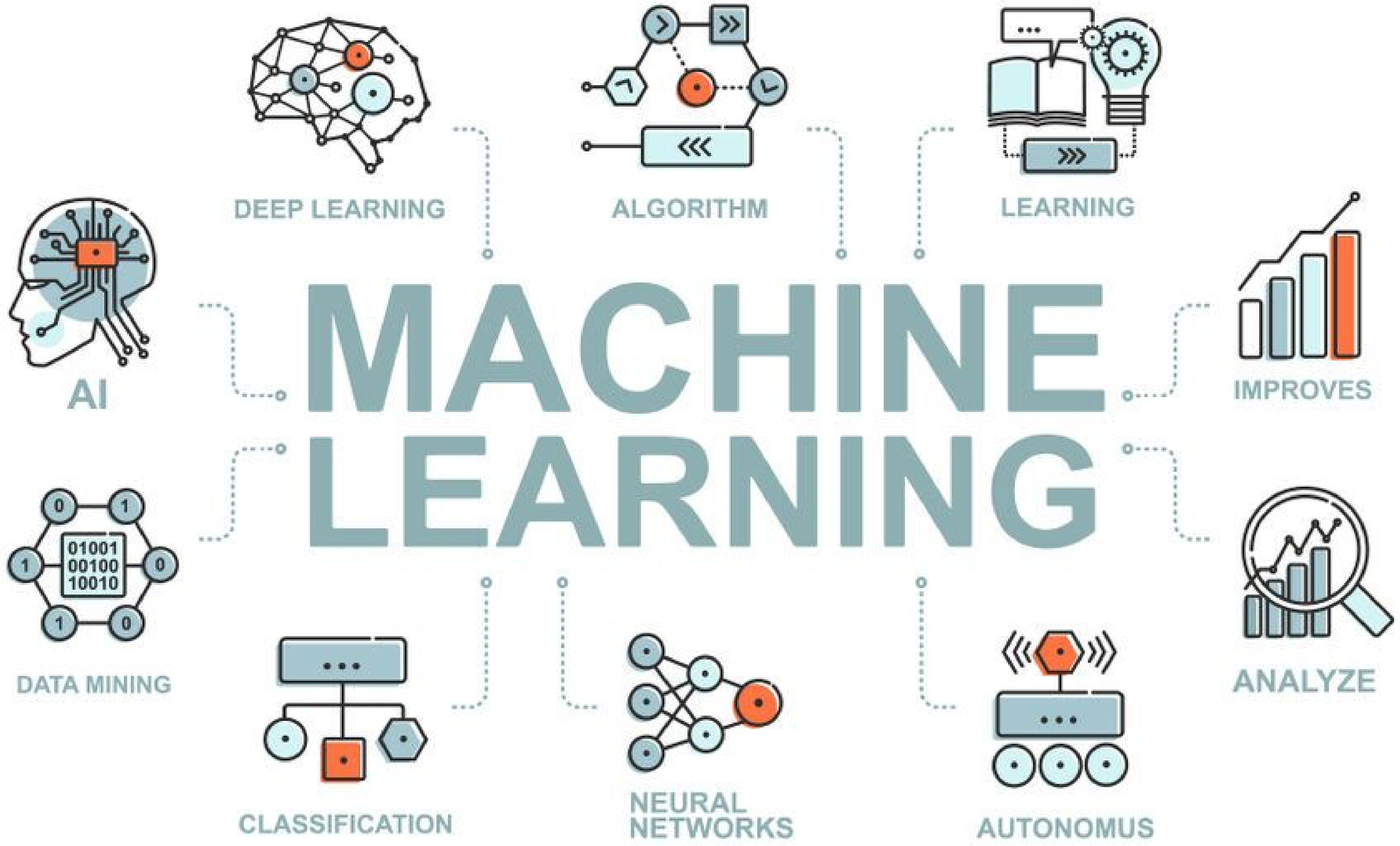
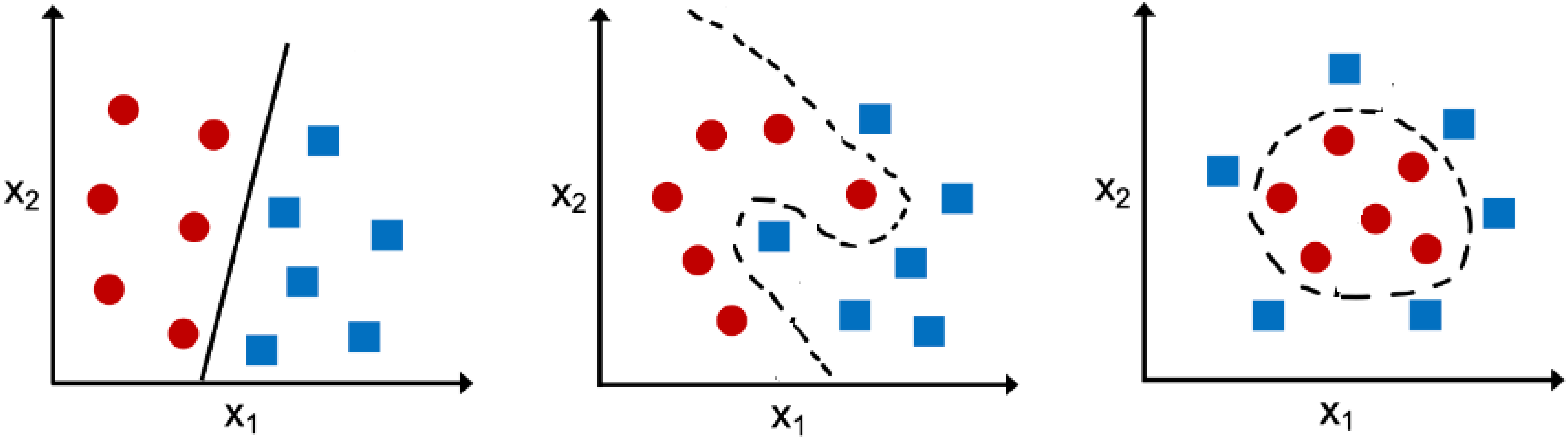


ALAN BADILLO SALAS

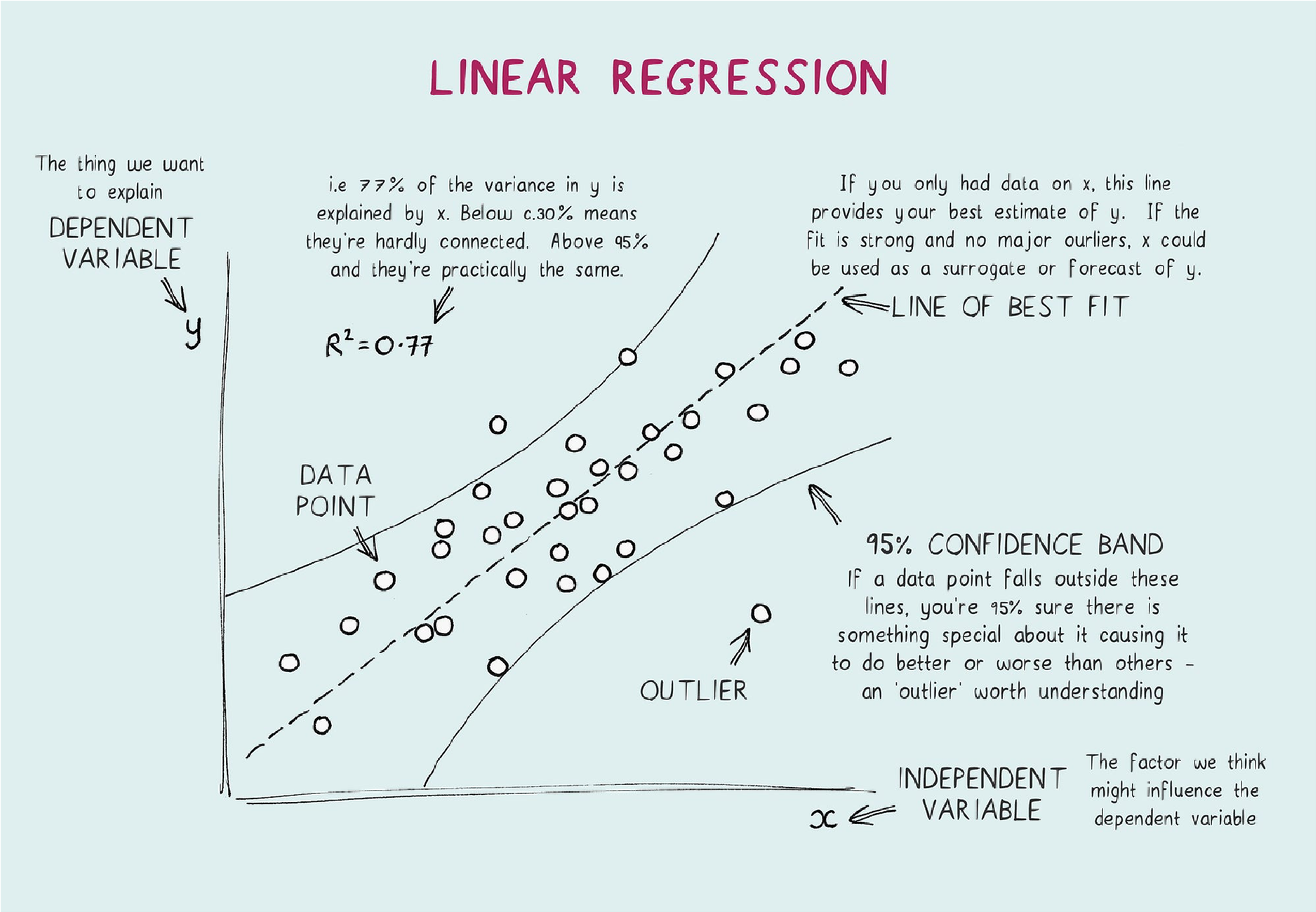
MÓDULO II

OCTUBRE 2023

CLASIFICACIÓN



REGRESIÓN



The thing we want to explain

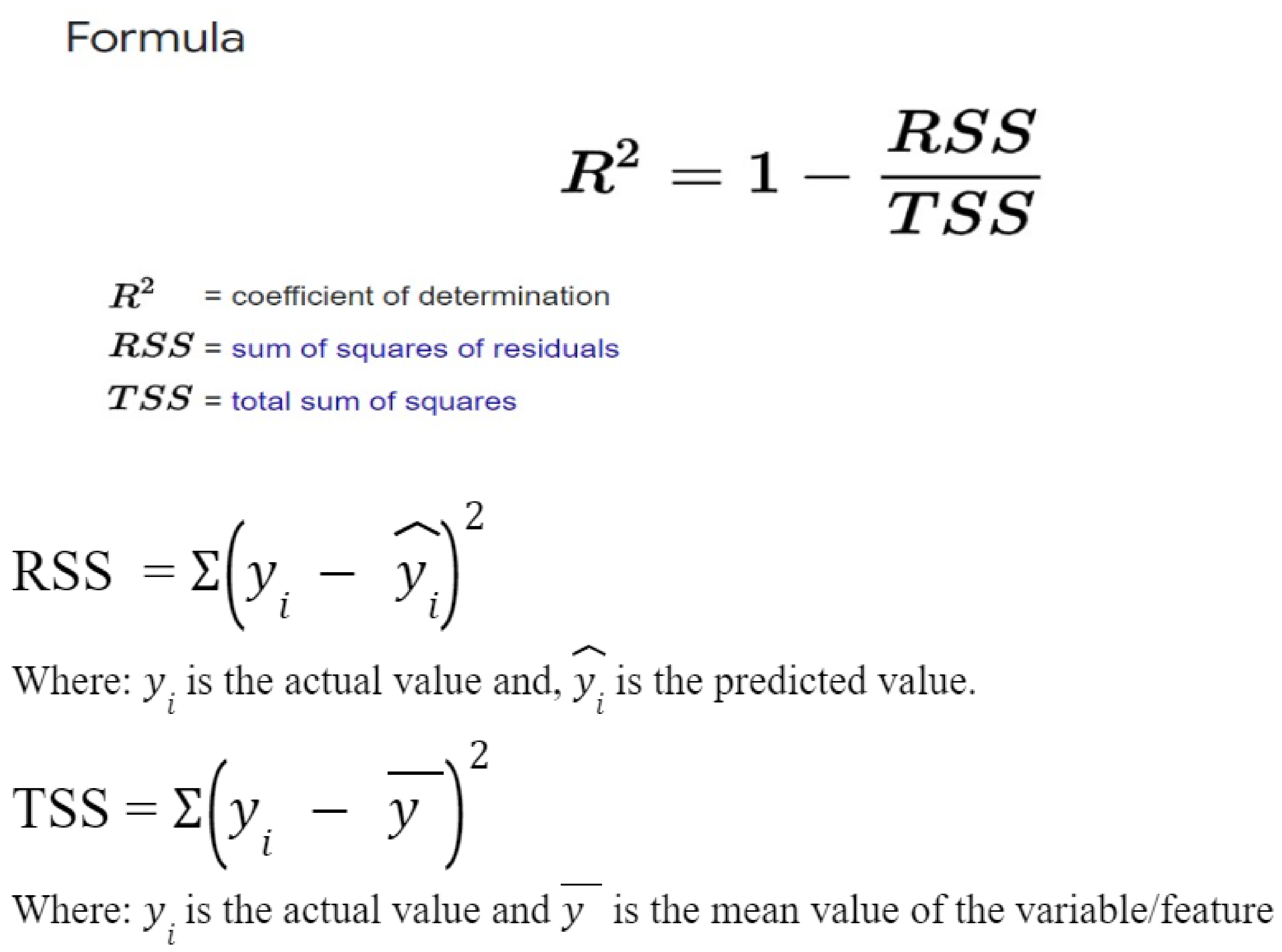
i.e 77% of the variance in y is explained by x. Below c.30% means they're hardly connected. Above 95% and they're practically the same.

If you only had data on x, this line provides your best estimate of y. If the fit is strong and no major ourliers, x could be used as a surrogate or forecast of y.

If a data point falls outside these lines, you're 95% sure there is something special about it causing it to do better or worse than others – an 'outlier' worth understanding

The factor we think might influence the dependent variable

REGRESIÓN



Formula

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coefficient of determination

RSS = sum of squares of residuals

TSS = total sum of squares

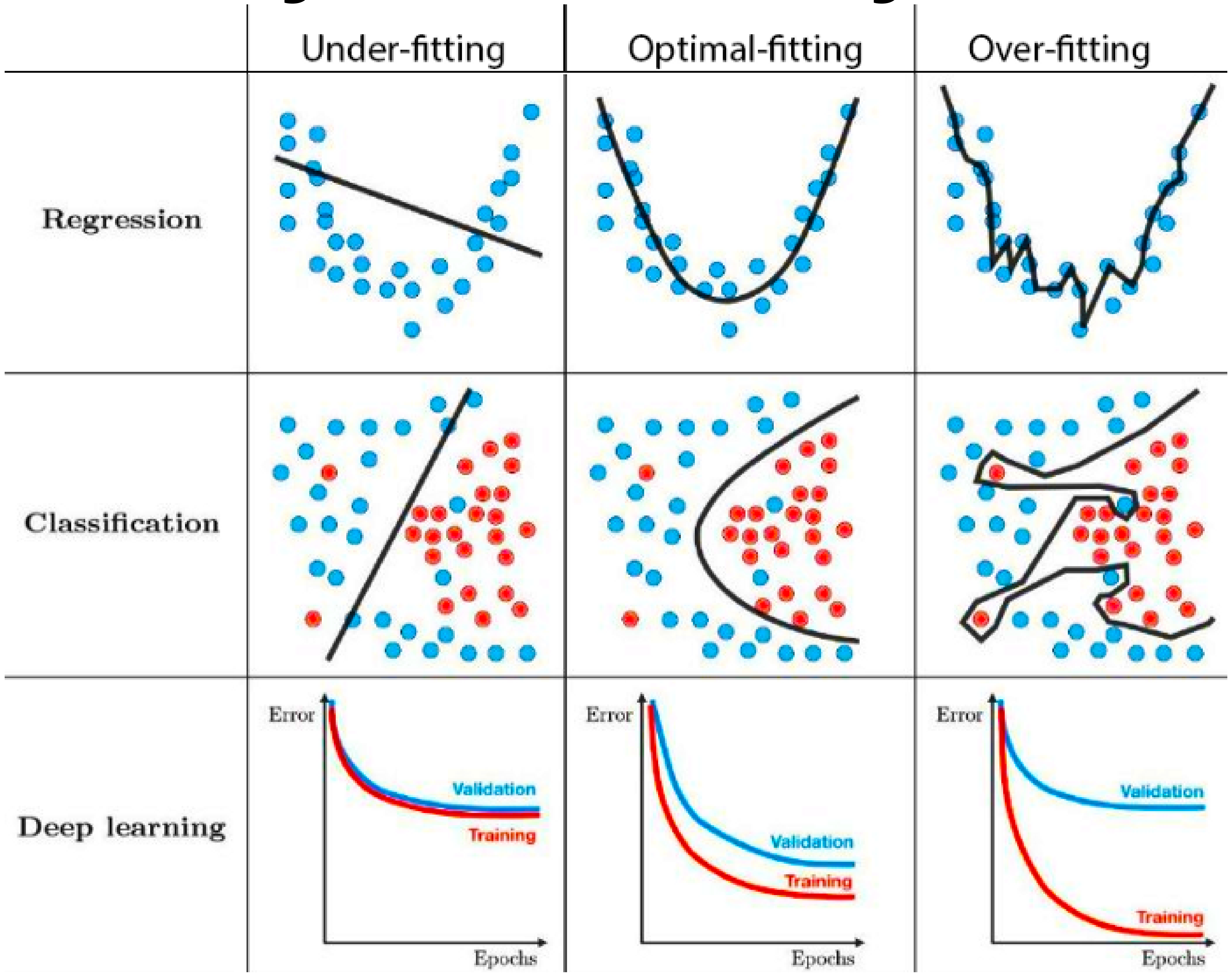
$$\mathrm{RSS} = \Sigma\left(y_i - \widehat{y}_i\right)^2$$

Where: y_i is the actual value and, $\widehat{y}_i$ is the predicted value.

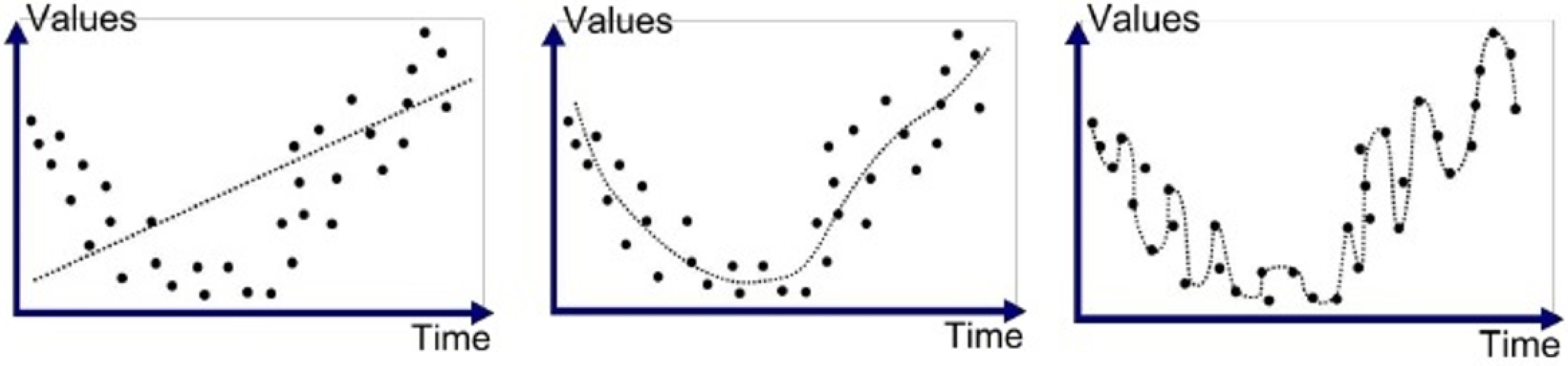
$$\mathrm{TSS} = \Sigma\left(y_i - \overline{y}\right)^2$$

Where: y_i is the actual value and $\overline{y}$ is the mean value of the variable/feature

GENERALIZACIÓN, SOBREAJUSTE Y SUBAJUSTE



GENERALIZACIÓN, SOBREAJUSTE Y SUBAJUSTE

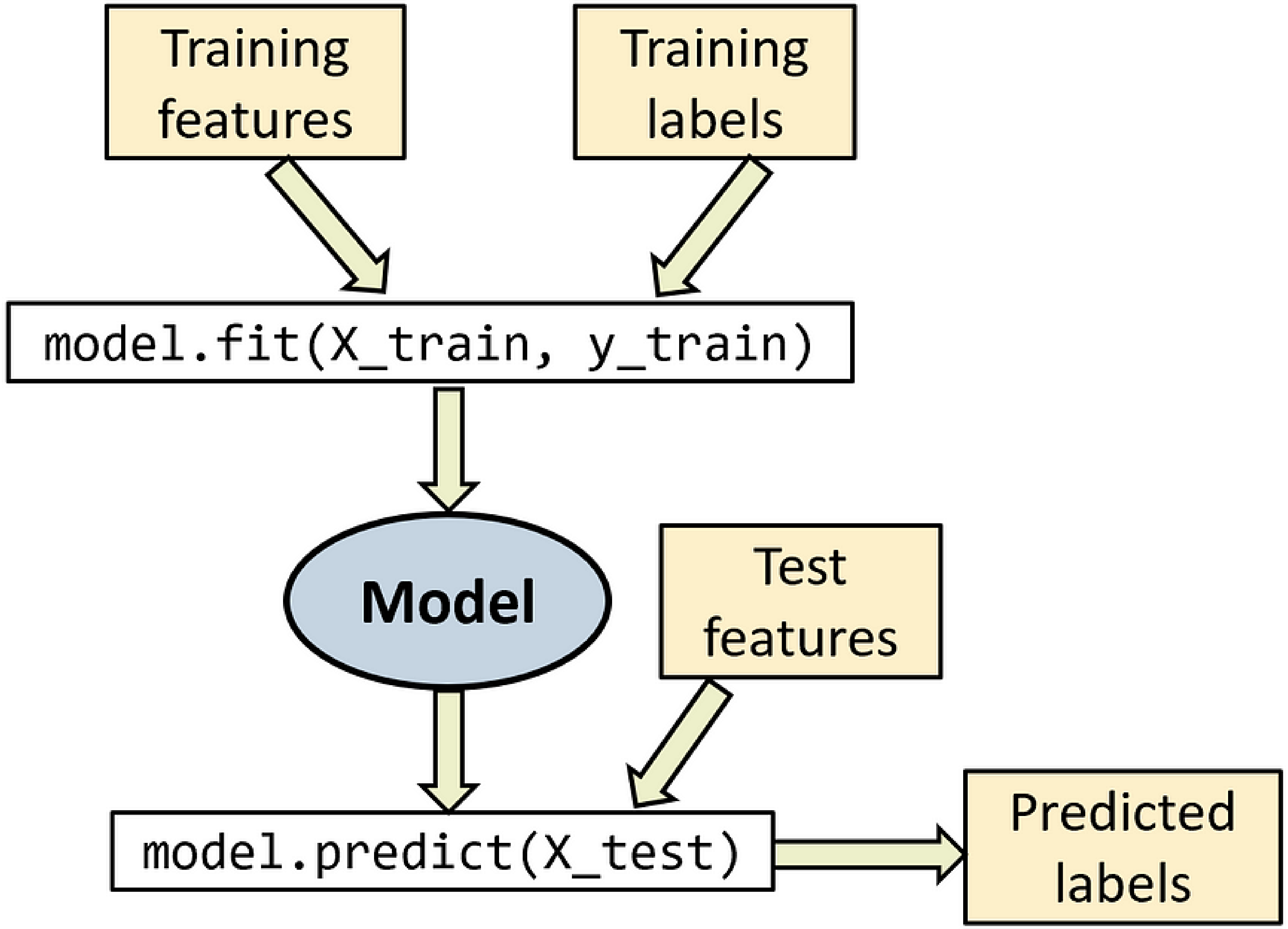


Underfitted

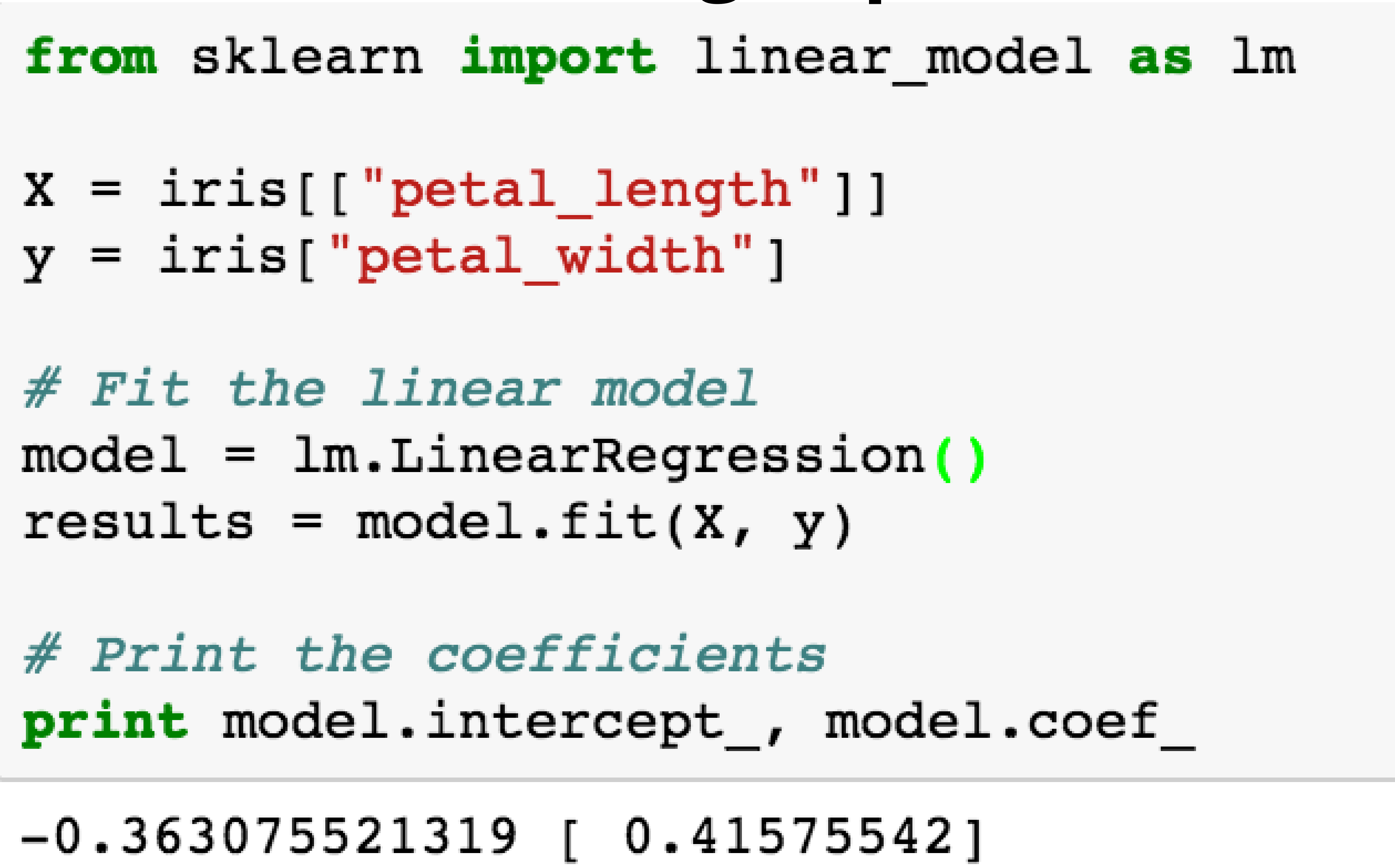
Good Fit/Robust

Overfitted

Algoritmo de Machine Learning Supervisado



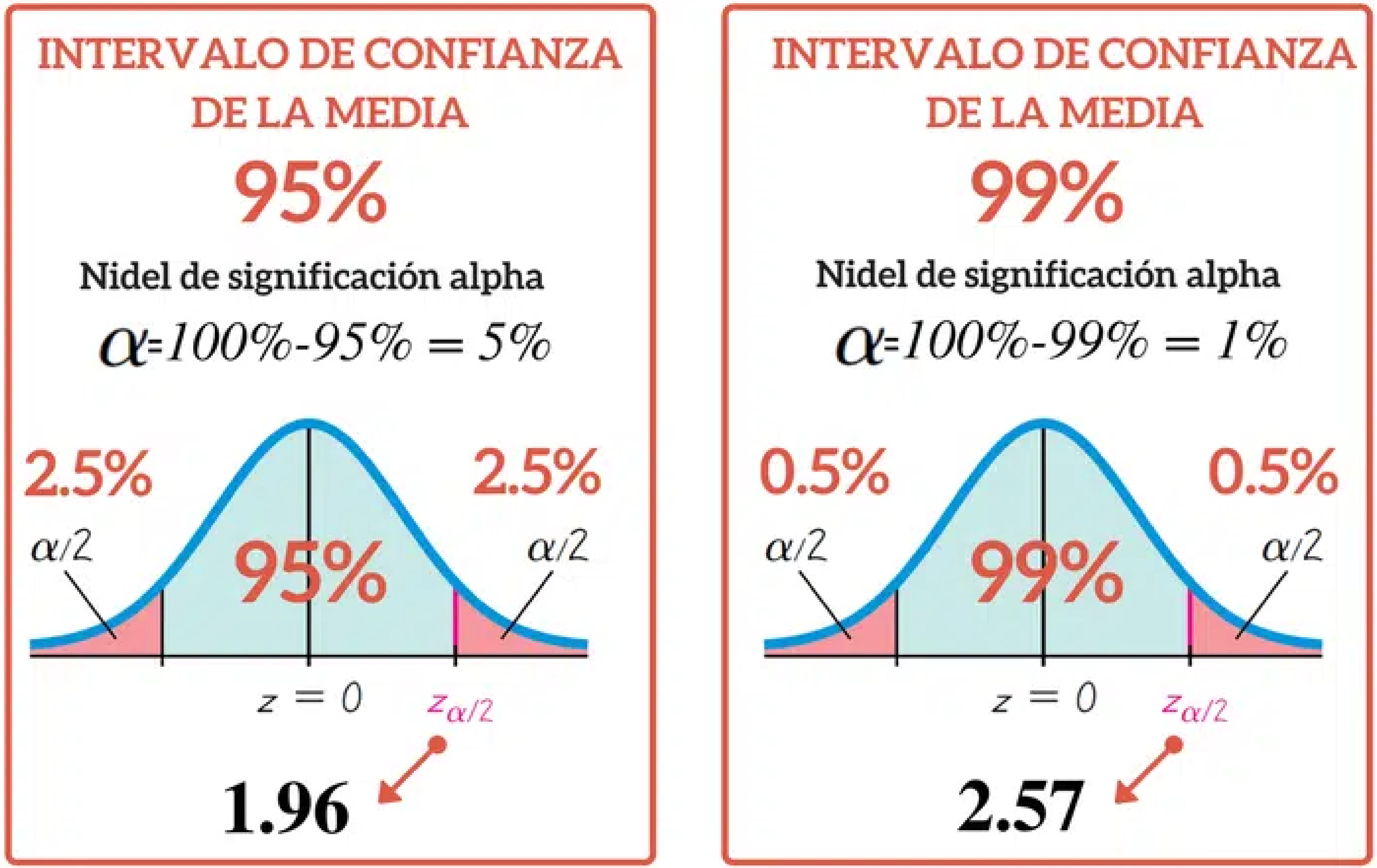
Algoritmo de Machine Learning Supervisado



```python
from sklearn import linear_model as lm

X = iris[["petal_length"]]
y = iris["petal_width"]

# Fit the linear model
model = lm.LinearRegression()
results = model.fit(X, y)

# Print the coefficients
print model.intercept_, model.coef_
```

```
-0.363075521319 [ 0.41575542]
```

Estimaciones de Incertidumbre


INTERVALO DE CONFIANZA DE LA MEDIA
95%
Nidel de significación alpha
α=100%-95% = 5%
2.5%
2.5%
α/2
95%
α/2
z = 0
$z_{\alpha/2}$
1.96
INTERVALO DE CONFIANZA DE LA MEDIA
99%
Nidel de significación alpha
α=100%-99% = 1%
0.5%
0.5%
α/2
99%
α/2
z = 0
$z_{\alpha/2}$
2.57