Classification

Cut off → 0.5 / 0.8

— target Variable is categorical

— Logistic Regression

└ linear Model

$y = f(x)$ $(ax_1 + bx_2 + cx_3)$

└ linear Eqn

$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Regression

BP	Chol	Sugar	HA (y)	z	ŷ
180	100	300	Y ✓	0.92	Y
80	20	90	N ✗	0.30	N — ∞
N			N	0.21	N 10
Y			Y	0.75	Y

$z = \beta_1 BP + \beta_2 \times Chol + \beta_3 \times sugar$ ←

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

$= \frac{1}{1+\frac{1}{e^z}}$

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

$z \to -\infty$

$z \to +\infty$

$-\infty$ $+\infty$ 473 $+\infty$

$z \to g(z) \to$ 0 / 1 → Yes / No

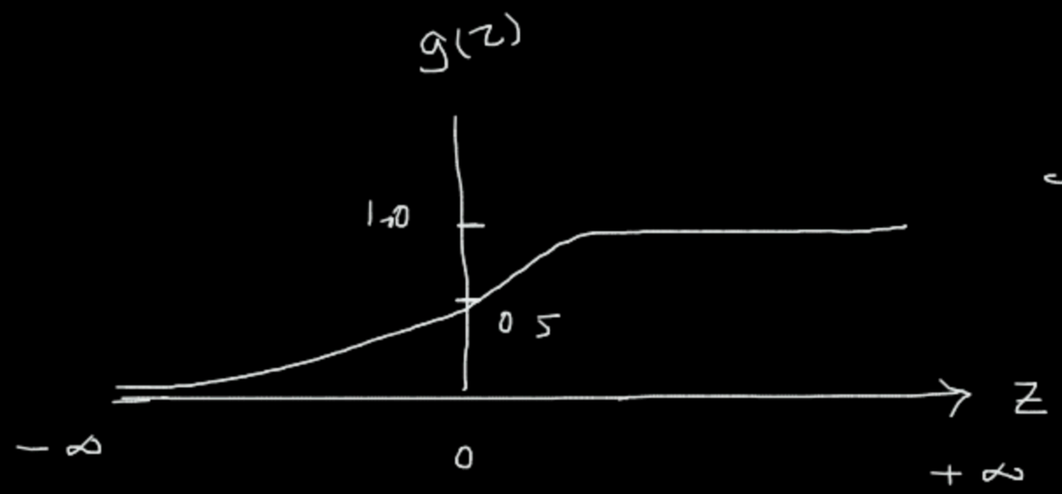
$g(z) = \frac{e^{-\infty}}{1+e^{-\infty}} \Rightarrow 0$

$g(z) = \frac{e^{\infty}}{1+e^{\infty}} \Rightarrow 1$

$e \to \infty$

$e^{-\infty} \to \frac{1}{\infty} \approx 0$

$z \to 0$ $\frac{e^0}{1+e^0} \to \frac{1}{1+1} = 0.5$



Sigmoid

$g(z)$ → Sigmoid Fn

Accuracy

↳ The percentage of Correctly predicted Values

	x_1 x_2 x_3	y	$\hat{y}$
1		y	y →
2		N	N
3		N	y
4		y	N
		y	y
10			

$\left(\frac{3}{5}\right)$

Imbalanced Data

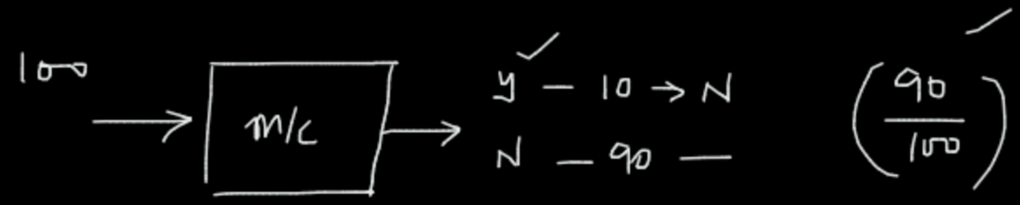
100 → 60 N — 50
→ 20 y — 50 →
M — 50

80 → Tr → 64 N / 16 y → Balanced { 45 +ve, 48 -Ve, 7 Neu →

20 → Ts 15 N
5 y

Resampling Techniques →

Training Data



F1-score pr.

Confusion Matrix

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Predict	Ground Truth: True (P)	Ground Truth: False (N)
True (P)	TP ✓	FP
False (N)	FN	TN ✓

Actual y	Predicted $\hat{y}$
Y	Y — TP
N	Y — FP
N	N — TN
Y	Y — TP
Y	N — FN
N	N — TN
Y	Y → TP
N	N — TN
Y	Y — TP
N	N — TN

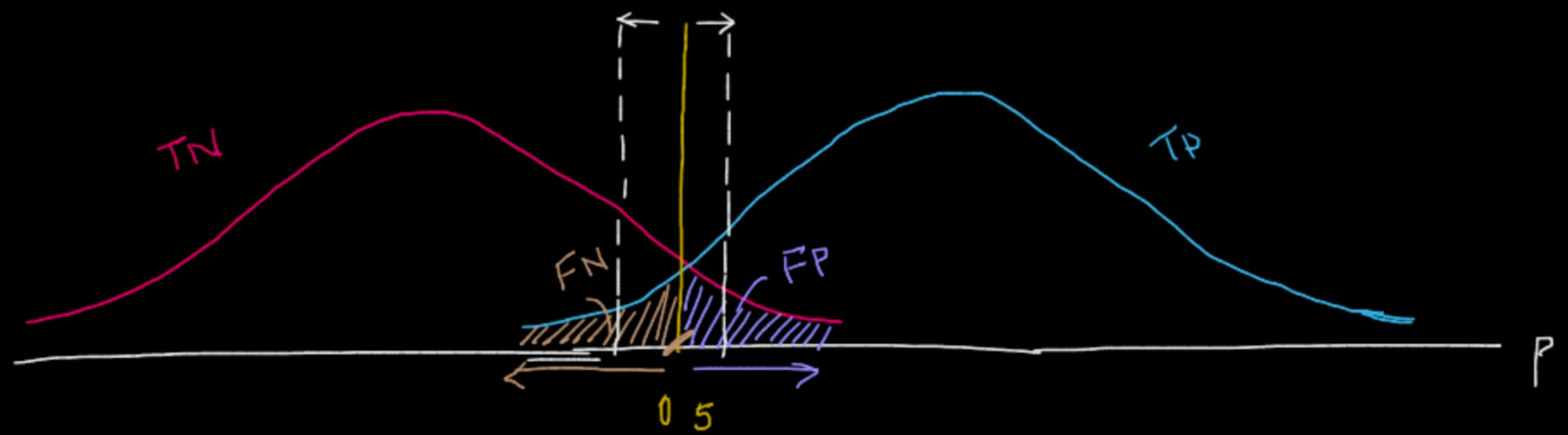
Predict \ Ground Truth	True (P)	False (N)	
True (P)	TP (4)	FP (1)	→ No of cases predicted as tve (5)
False (N)	FN (1)	TN (4)	
	↓ No of cases that are actually +ve (5) ✓		

$$\text{Accuracy} = \frac{8}{10} \rightarrow \frac{TP+TN}{TP+TN+FP+FN} \rightarrow \text{Percentage of correctly predicted cases}$$

$$\text{Precision} - \frac{TP}{TP+FP} = \left(\frac{4}{5}\right) \rightarrow \text{out of the cases predicted as +ve How many are actually +ve} \quad ✓$$

$$\text{Recall} - \frac{TP}{TP+FN} = \left(\frac{4}{5}\right) \rightarrow \text{out of the cases that are actually +ve how many were predicted as +ve} \quad ✓$$

$$\text{F1\_score} - \frac{2\,Pr \times Rc}{Pr + Rc}$$



FN↑ FP↓

FN↓ FP↑

↑ Pr — $\frac{TP}{TP+FP}$ ↓ → ↓Rc → $\frac{TP}{TP+FN}$ ↑ →

1 Doctor treating Covid cases
Predict whether patient has Covid or Not

High Recall

Churn predicted

2 Security software which predicts whether a pg is safe or not

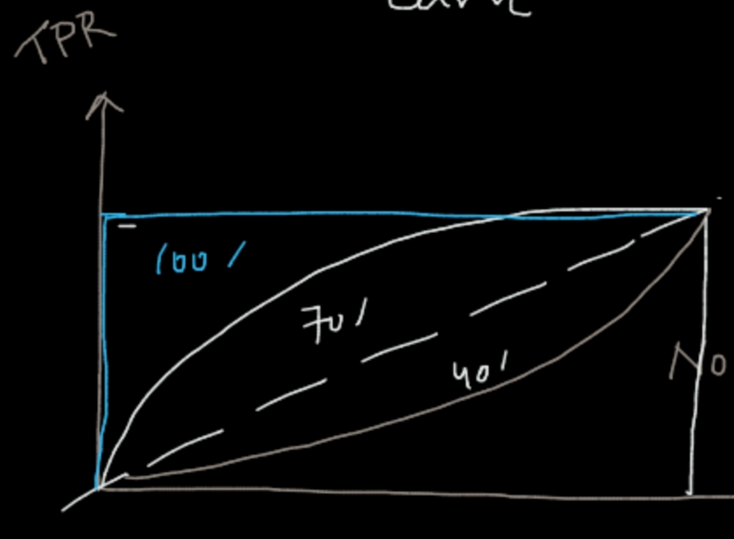
High Precision

100 → (60 → 78)

AUC – ROC

↓ Area Under the Curve

↳ Receiver operating characteristic curve



(a) Accuracy

(b) Precision

(c) Recall

(d) F1-score

(e) AUC – ROC