INTRODUCTION TO PYTHON FOR STATISTICAL LEARNING

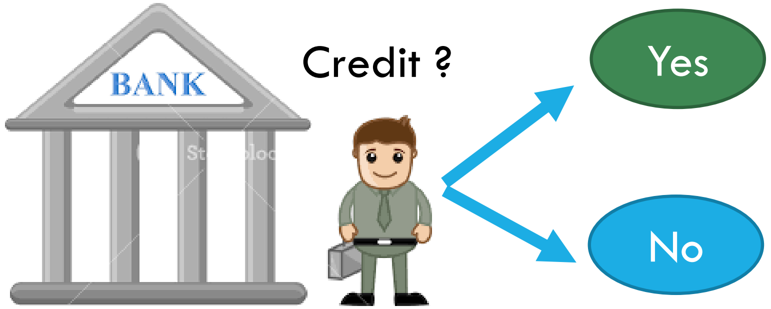
Fathi Abdul Muhyi

MACHINE LEARNING

“Sebuah Algoritma yang dapat mempelajari karakteristik data, membuat prediksi, menggali informasi berdasarkan data”

PENERAPAN MACHINE LEARNING

Classification



Credit ?

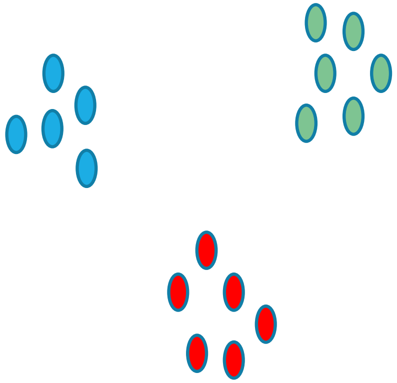
Yes

No

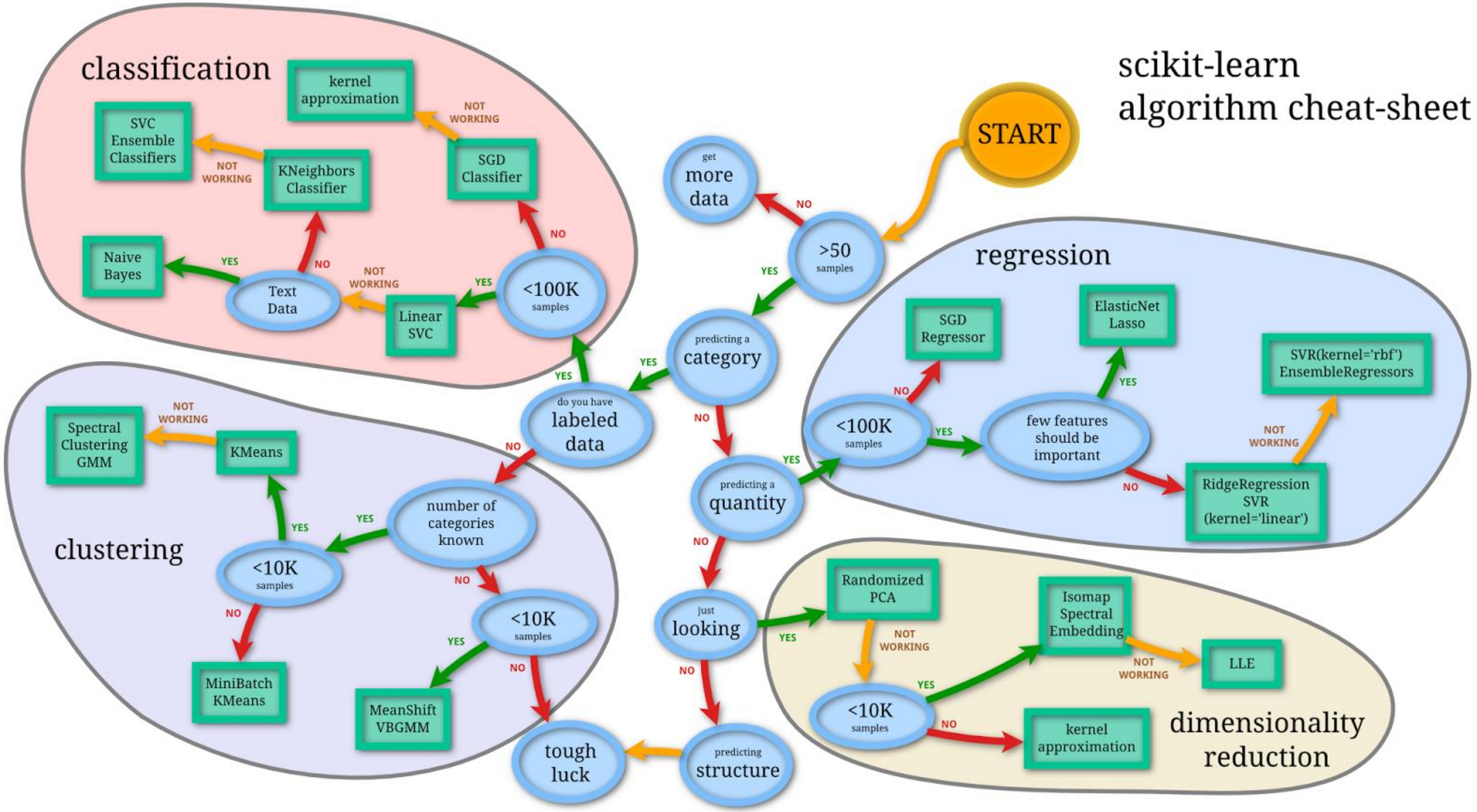
Regression

How Much ?

Clustering



ML FLOWCHART



METODE KLASIFIKASI

Classification is a method to build a model to classify the response variable by certain charcteristic (feature variable).

There are many problem in case of classification :

1. Credit scoring, calon penerima kredit mampu bayar atau tidak

2. Churn, customer yang berpotensi meninggalkan jasa/produk

3. Direct Marketting, identifikasi prospective customer

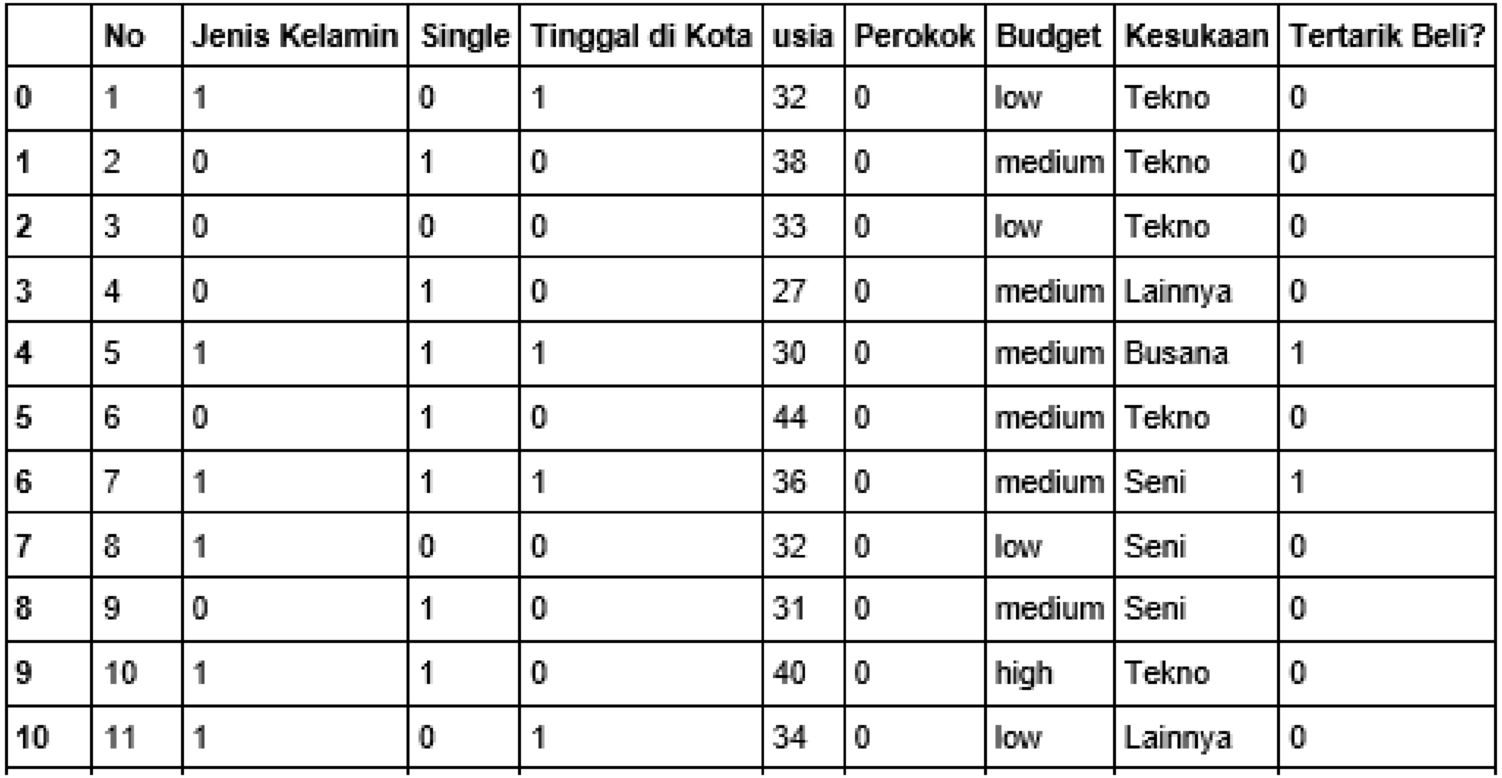
CONTOH KASUS SEDERHANA

Toko X mengumpulkan data pelanggan untuk mengetahui apakah pelanggan mereka tertarik untuk membeli produk jenis baru

Toko tersebut memiliki sejumlah pelanggan baru yang belum diketahui tertarik atau tidak

Peubah Penjelas

Peubah Respon



	No	Jenis Kelamin	Single	Tinggal di Kota	usia	Perokok	Budget	Kesukaan	Tertarik Beli?
0	1	1	0	1	32	0	low	Tekno	0
1	2	0	1	0	38	0	medium	Tekno	0
2	3	0	0	0	33	0	low	Tekno	0
3	4	0	1	0	27	0	medium	Lainnya	0
4	5	1	1	1	30	0	medium	Busana	1
5	6	0	1	0	44	0	medium	Tekno	0
6	7	1	1	1	36	0	medium	Seni	1
7	8	1	0	0	32	0	low	Seni	0
8	9	0	1	0	31	0	medium	Seni	0
9	10	1	1	0	40	0	high	Tekno	0
10	11	1	0	1	34	0	low	Lainnya	0

REGRESI LOGISTIK BINER

Memprediksi peluang ketertarikan untuk membeli

$$P(Y = 1) = \frac{\exp(\beta_0+\beta_1 x_1+ \ldots+\beta_k x_k)}{1+\exp(\beta_0+\beta_1 x_1+ \ldots+\beta_k x_k)}$$

Y : peubah kategorik biner

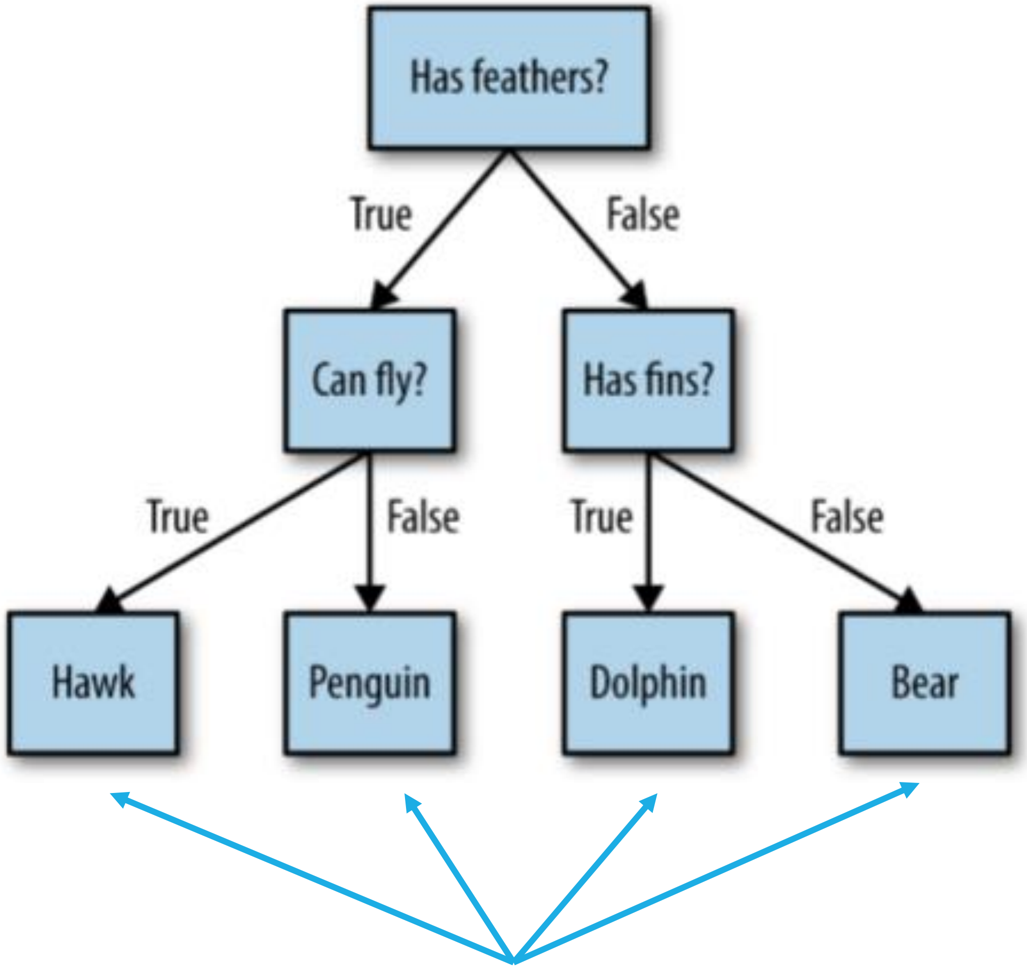
x_i : peubah penjelas

β_i : Parameter Regresi

POHON KEPUTUSAN

Hierarchi of if/else question, leading to a decision (Muller and Guido, 2016)

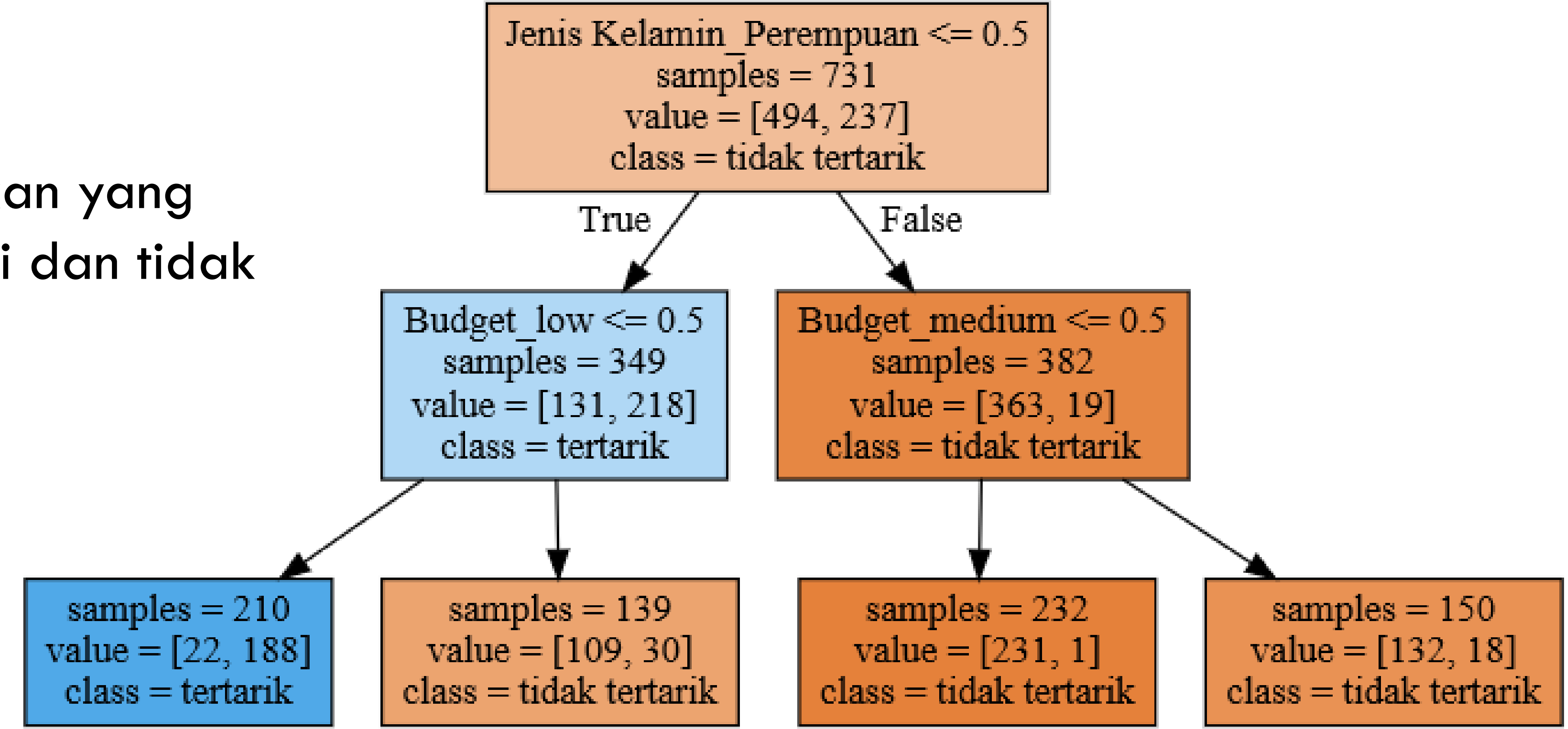
How to distinguish :
Hawk, Penguin, Dolphin, Bear ?



Terminal Node / Leaf

POHON KEPUTUSAN

Membedakan yang tertarik beli dan tidak



Jenis Kelamin_Perempuan <= 0.5
samples = 731
value = [494, 237]
class = tidak tertarik

True

False

Budget_low <= 0.5
samples = 349
value = [131, 218]
class = tertarik

Budget_medium <= 0.5
samples = 382
value = [363, 19]
class = tidak tertarik

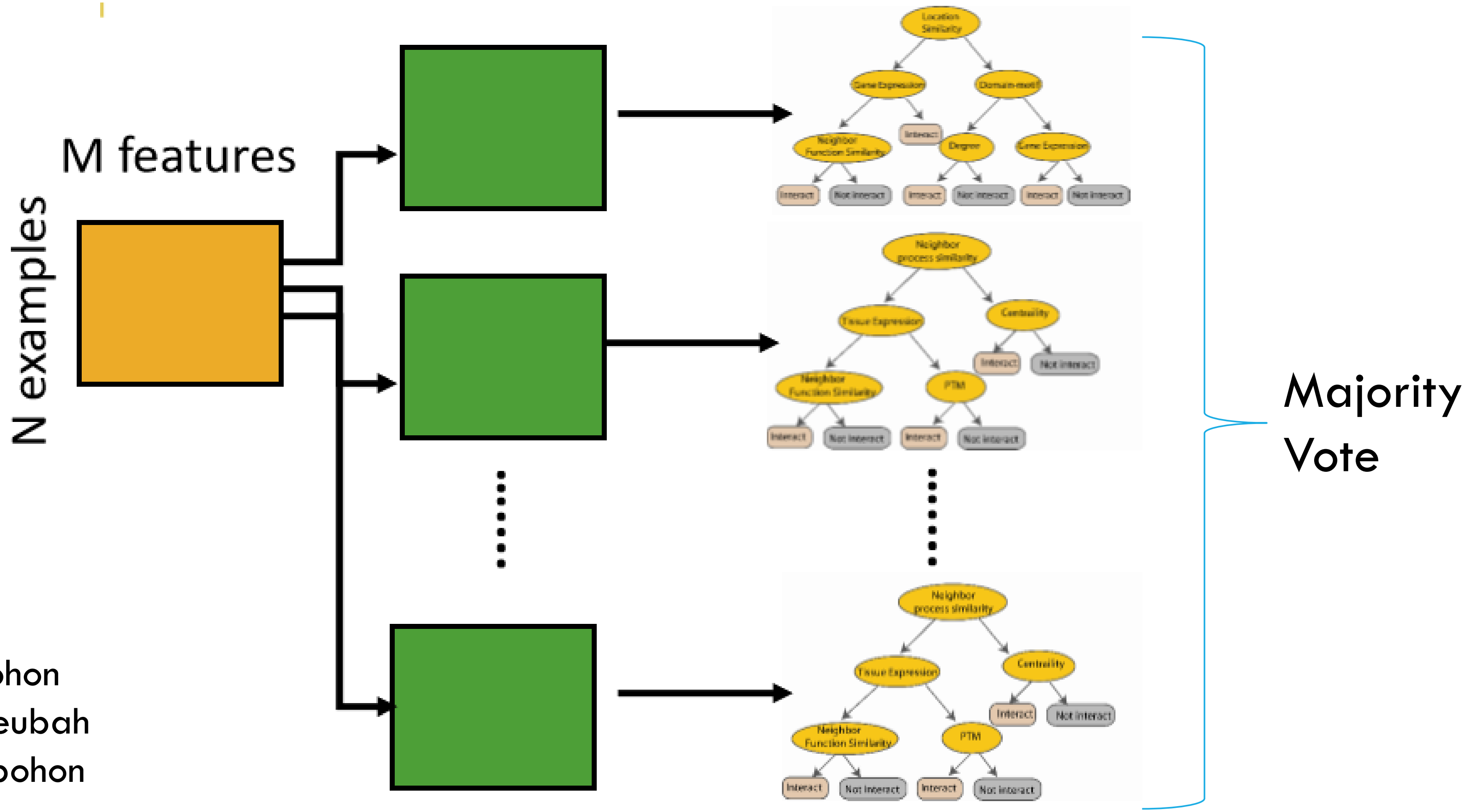
samples = 210
value = [22, 188]
class = tertarik

samples = 139
value = [109, 30]
class = tidak tertarik

samples = 232
value = [231, 1]
class = tidak tertarik

samples = 150
value = [132, 18]
class = tidak tertarik

RANDOM FOREST



Random Forest :

1. Tree cenderung overfit dan keragamannya tinggi
2. Menggabungkan banyak pohon
3. Pengambilan contoh acak peubah dan observasi untuk setiap pohon

EVALUASI

Ukuran Ketepatan Prediksi :

1. Akurasi, persentase yang terprediksi benar
2. Sensitivitas/Recall, persentase yang terprediksi benar pada kelas positif
3. Spesifisitas, persentase yang terprediksi benar pada kelas negatif
4. dan lain-lain

Metode Evaluasi :

1. Validasi (Train - Test)
2. Validasi Silang

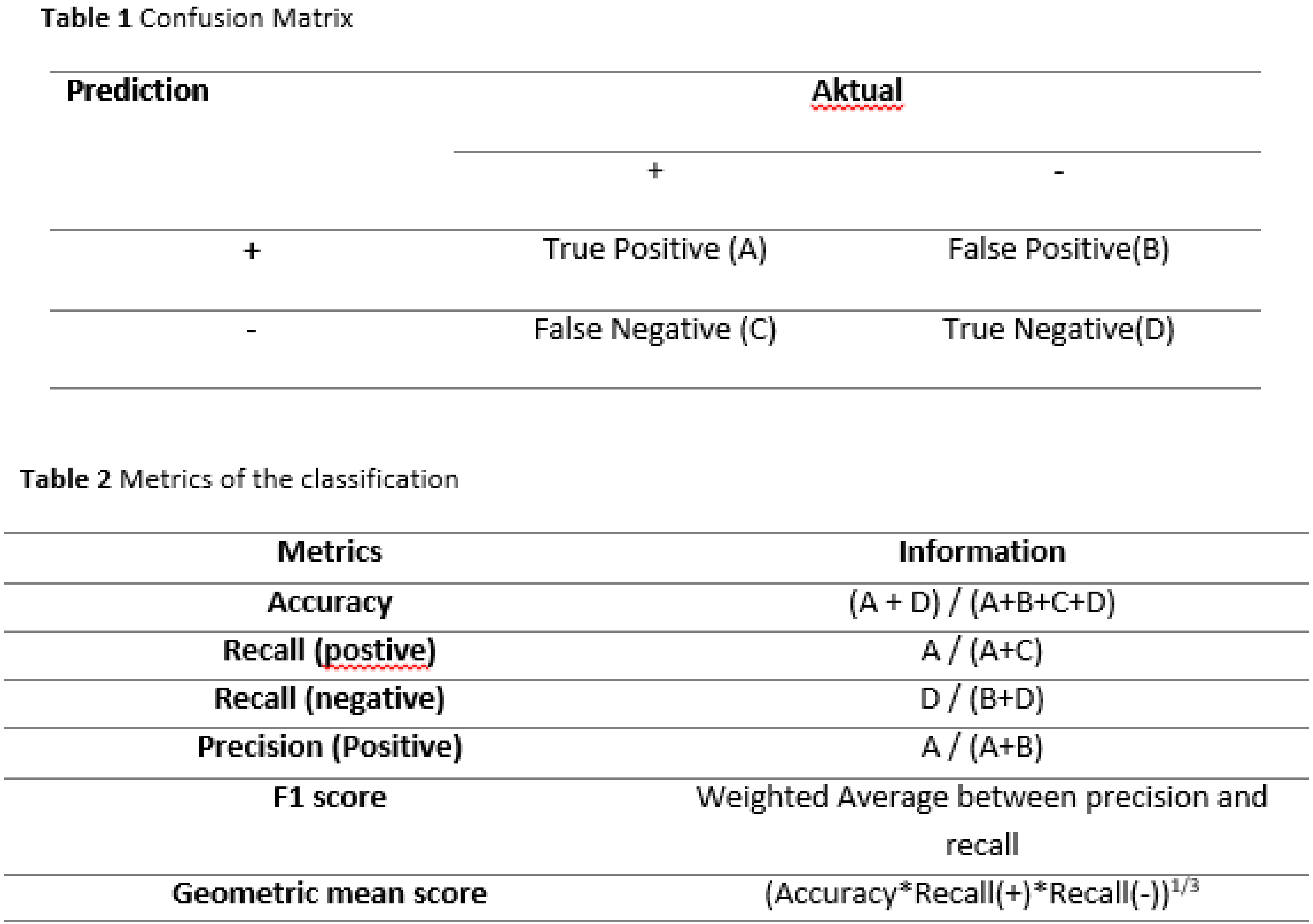


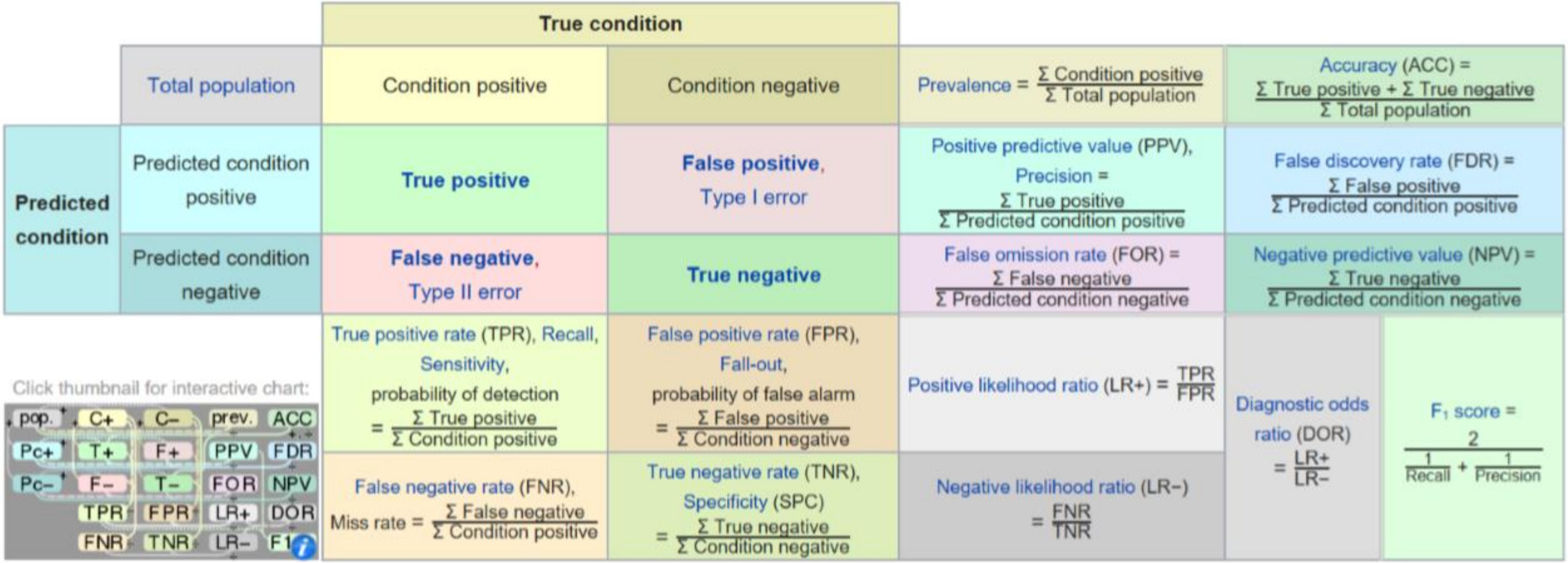
Table 1 Confusion Matrix

Prediction	**Aktual**	
	+	-
+	True Positive (A)	False Positive(B)
-	False Negative (C)	True Negative(D)

Table 2 Metrics of the classification

Metrics	**Information**
Accuracy	(A + D) / (A+B+C+D)
Recall (postive)	A / (A+C)
Recall (negative)	D / (B+D)
Precision (Positive)	A / (A+B)
F1 score	Weighted Average between precision and recall
Geometric mean score	(Accuracy*Recall(+)*Recall(-))$^{1/3}$

METRICS



		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
	Click thumbnail for interactive chart: pop. C+ C− prev. ACC Pc+ T+ F+ PPV FDR Pc− F− T− FOR NPV TPR FPR LR+ DOR FNR TNR LR− F1	True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

CONTOH

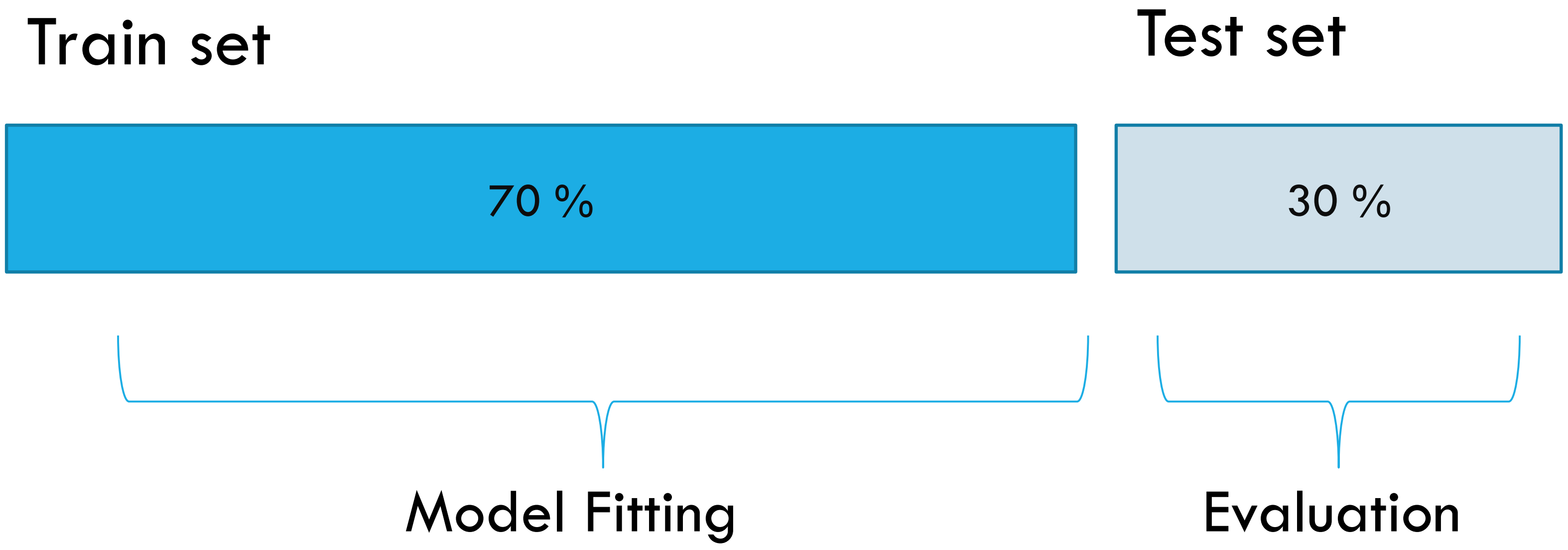
No	Perdiksi	Aktual
1	1	1
2	1	0
…	…	…
…	…	…
553	0	0
554	0	1

Matriks Klasifikasi

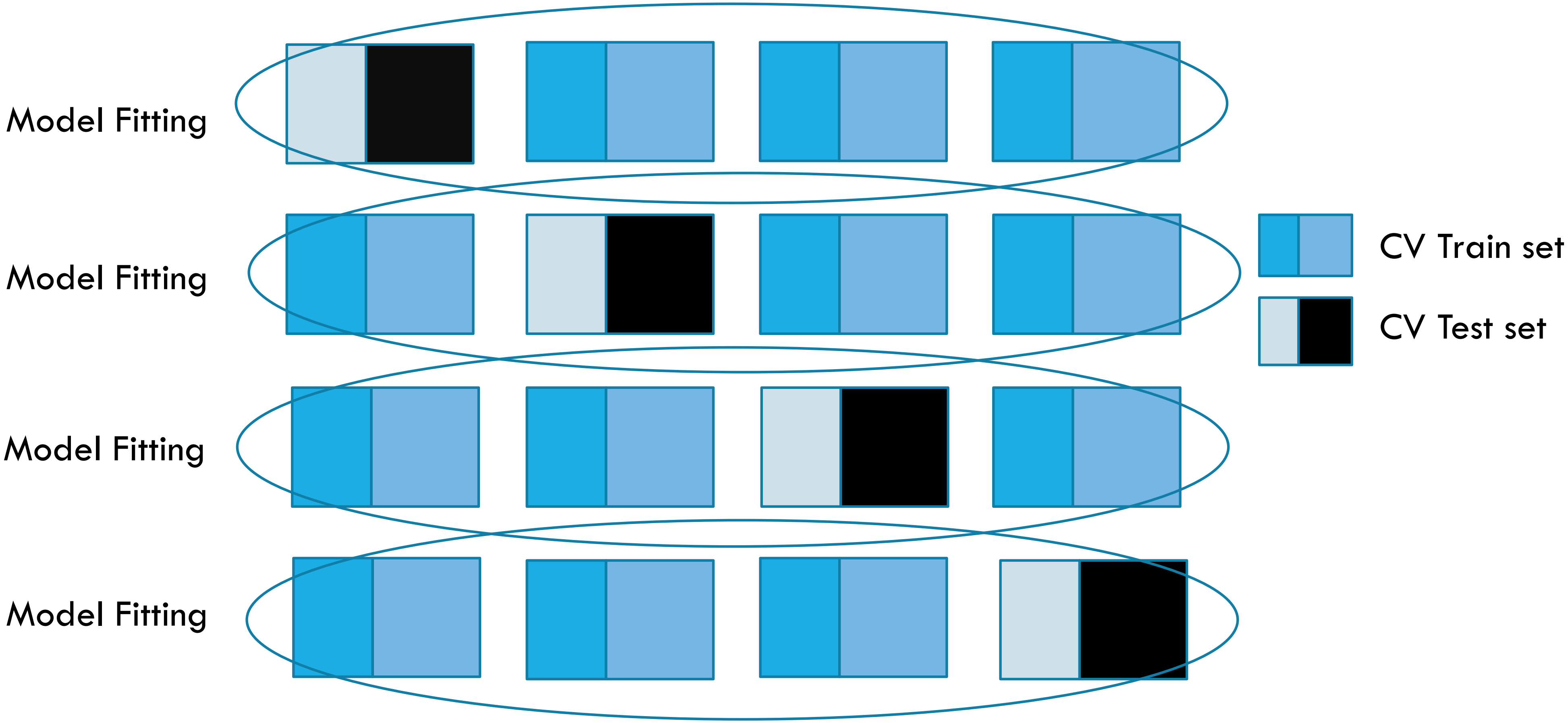
Prediksi	Aktual	
	1	0
1	121	23
0	10	400

Akurasi :
94.04 %
Sensitivitas :
92.36 %
Spesifisitas :
94/56 %

VALIDASI



VALIDASI SILANG



Model Fitting

Model Fitting

Model Fitting

Model Fitting

CV Train set

CV Test set

TASK

Prediksi variabel default pada data titanic dan upload di kaggel.com