Credit card fraudlent

Transaction → ? → Yes / No

Historical

KNN → Yes | No

transa-history	F2	F3	state
—	—	—	ND
—	—	—	ND
—	—	—	ND
—	—	—	ND
—	—	—	Yes
—	—	—	Yes

(3) → 5000

link

Real

1000 → Yes 500

9000 → NO

DB

KNN ⇒ class-weight

[1) Oversampling
2) Undes]

↑

Imbalance

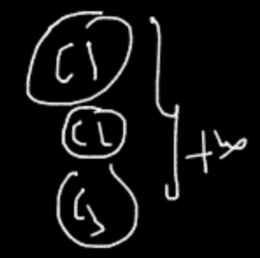
D_x (X, Y) → G [Yes, NO] → Binary classification

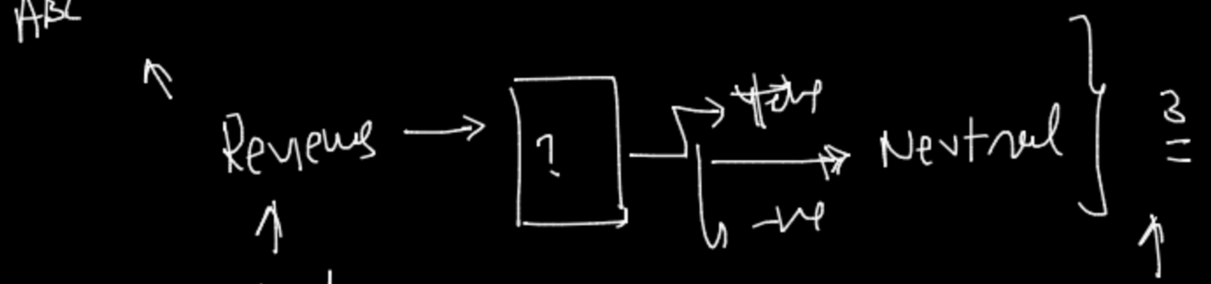
Input o/p

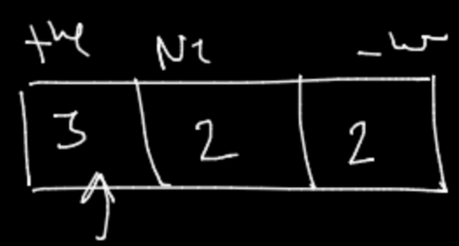
But

(two)

KNN →







multiclass - Classification

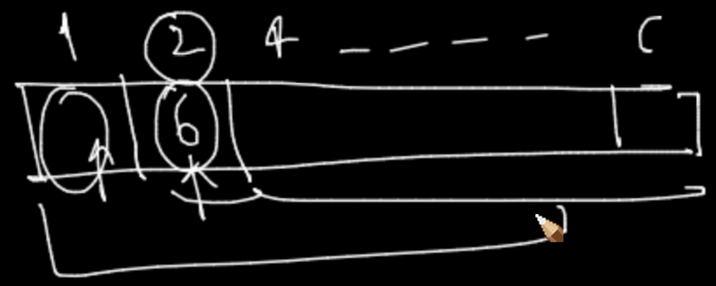
7 N N

$D_n = \{ X, Y \mid X \in R_q \; Y_i \in [0, 1, 2, \ldots 6] \}$

KNN

1 Vs Rest

Let 7-NN →



Train test set difference

$D_n\{X,Y\}$ → D_{train} randomly D_{test}

$D_n\{X,Y\}$ → D_{train} TBS

eg. Sales

Day	discount	Sales	volume
1	–	–	–
2	–	–	–
3	–	–	–
1	–	–	–
1	–	–	–
120	–	–	–

2 Months 60 days

60 days → test

Training

$P_{train}(X_{train}, Y_{train}) \rightarrow$ plot it

$\{+w, -w\}$ → 40% 60%

$(+w, -w)$

$\{Y_{test}, Y_{test}\}$

P_{test}

35% 65%

validation

Bad

X

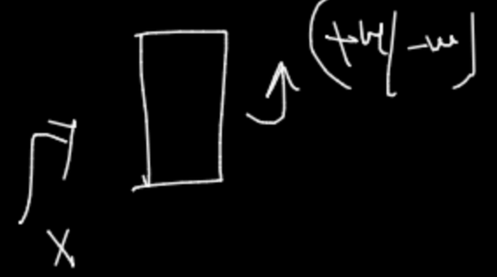
train U

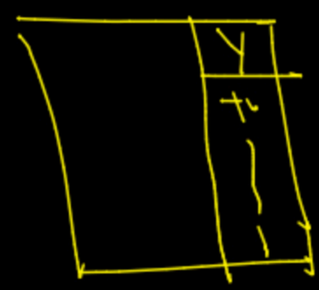
data

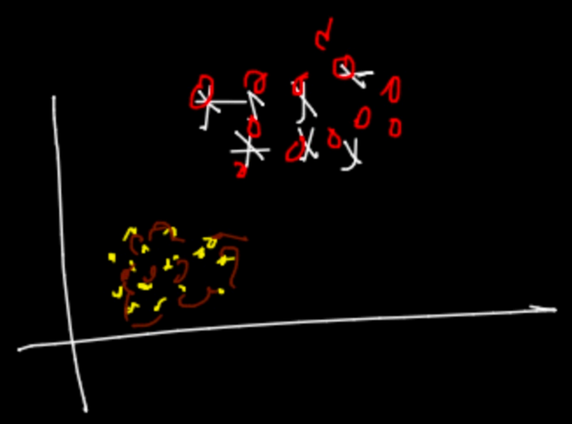
testing

train → pred X

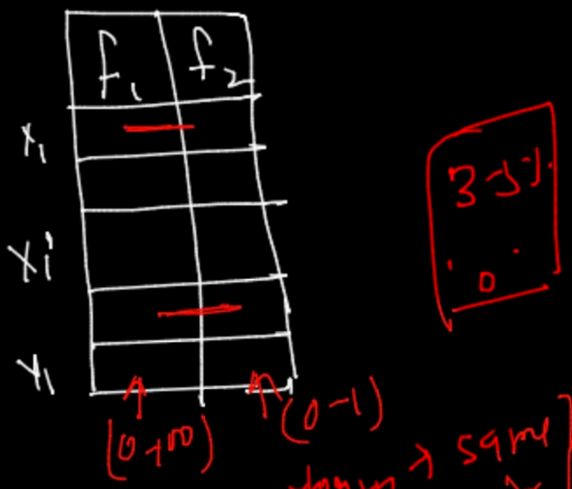
good

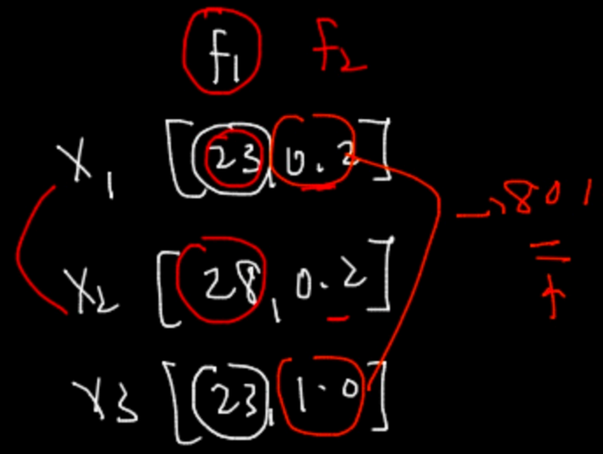






Impact of scale and column standardisation





uniform → same unit

KNN ← Data

test

distance (X1 X2) = 5

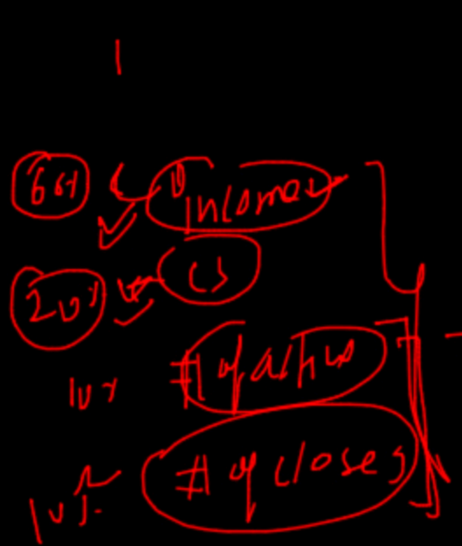
distance (X1 X3) = 0.8

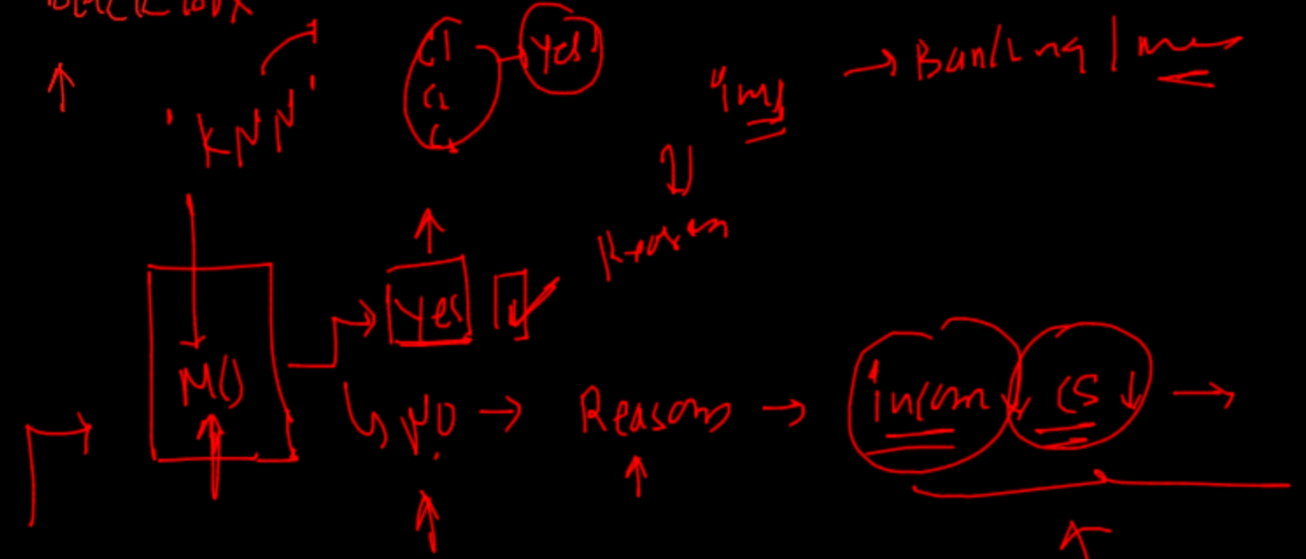
(0-1)

logically →

⇒ Model Interpretability and Black box

↑ 'KNN'

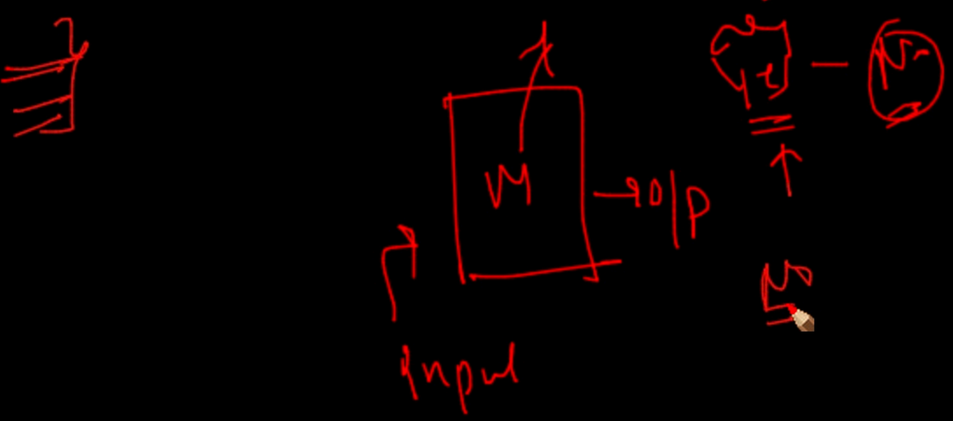


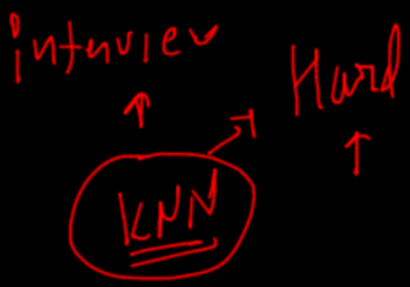


(Income, CS, # of achievements, # of classes)

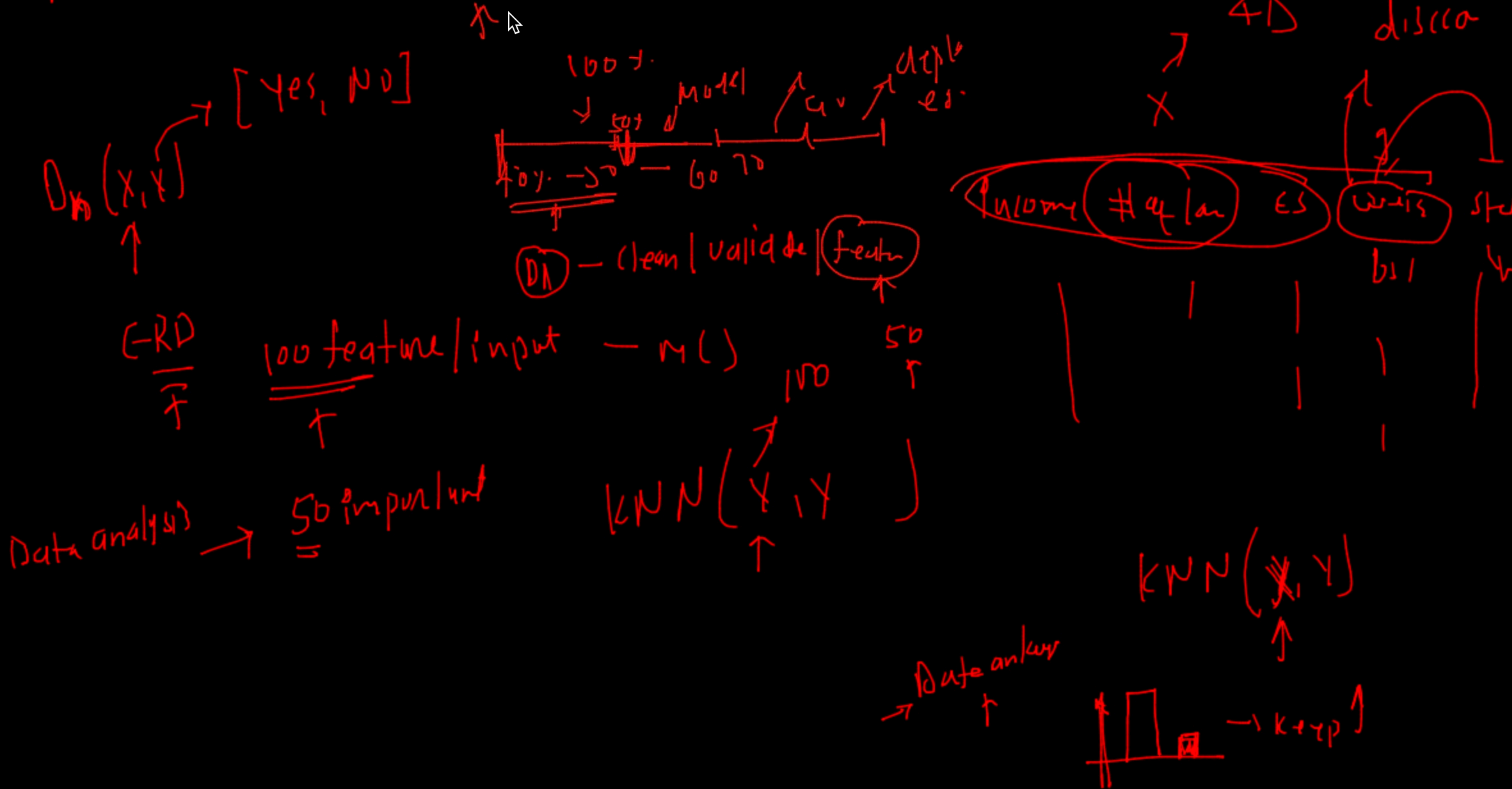
[25020 [1, 0]

KNN →

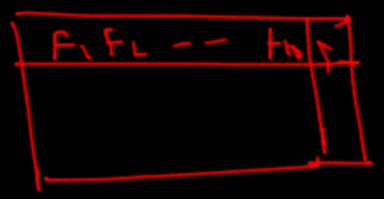


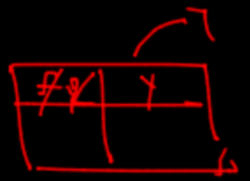


Feature Importance and Forward selection



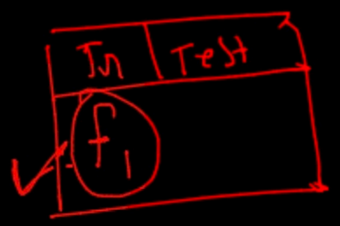
$f_1\ f_2\ f_3\ \cdots\ f_{10}$

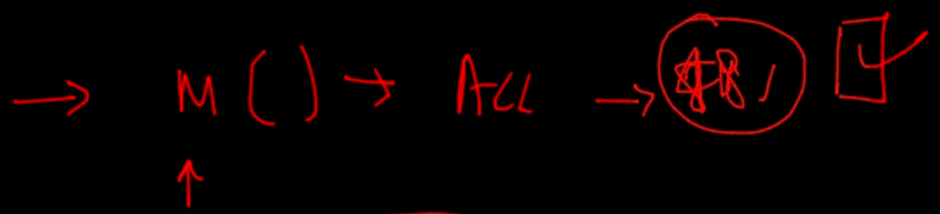


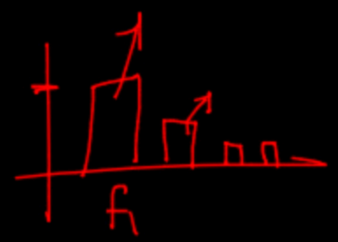


$D_n = \{T_R, T_F\}$

Step 1







important ← f2 — ACL → 51%

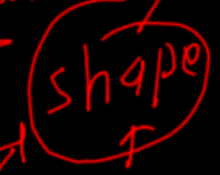
f3 → ACL + 38%

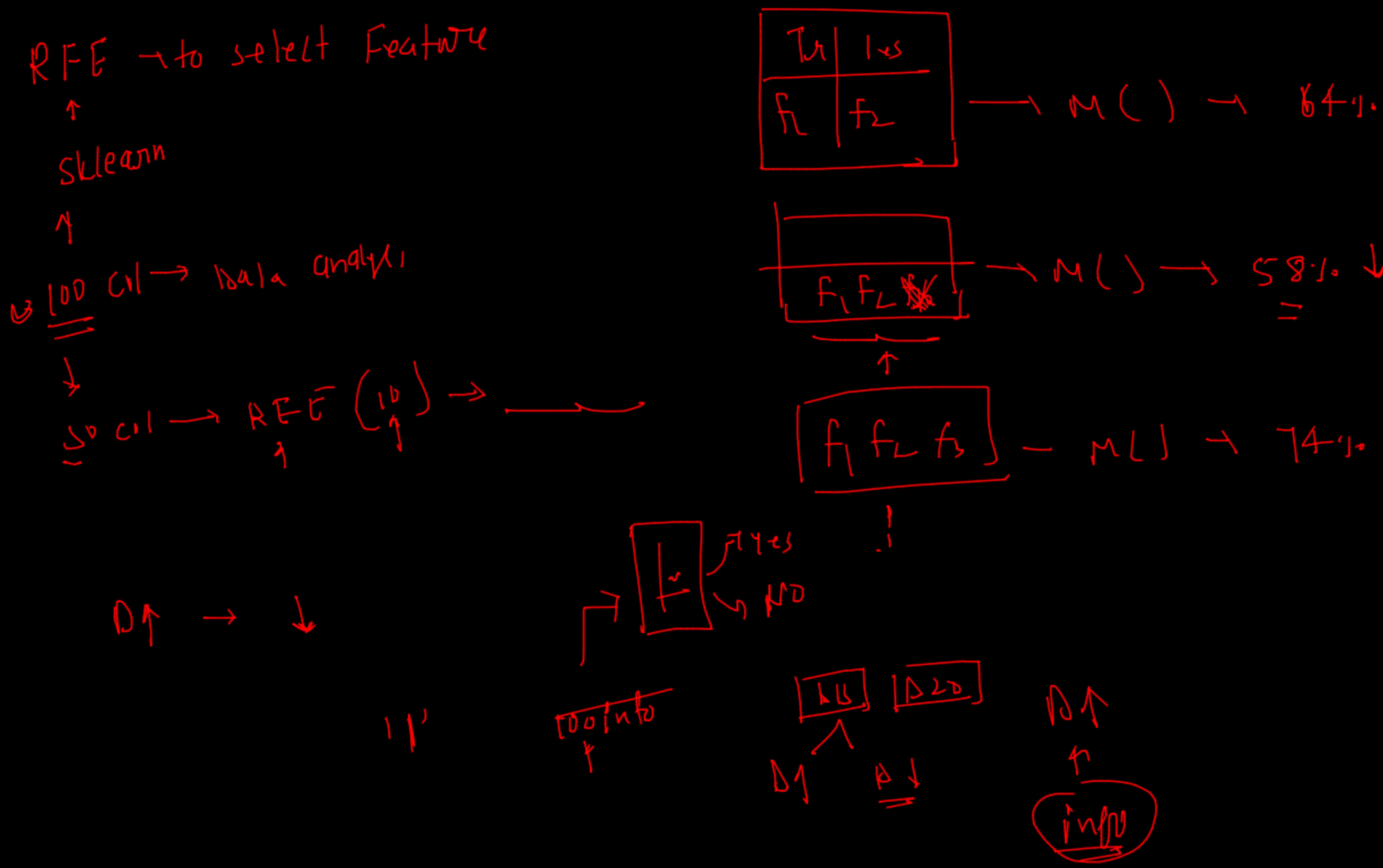
least → fx → ACL → 10%

f10 → ACL → 43%.

"Feature important"

evidently




RFE → to select Feature
sklearn
100 col → Data analysis
50 col → RFE (10)
Tr
Tes
f1
f2
M( ) → 64%
f1 f2 → M( ) → 58% ↓
f1 f2 f3 → M( ) → 74%
yes
NO
100 info
info