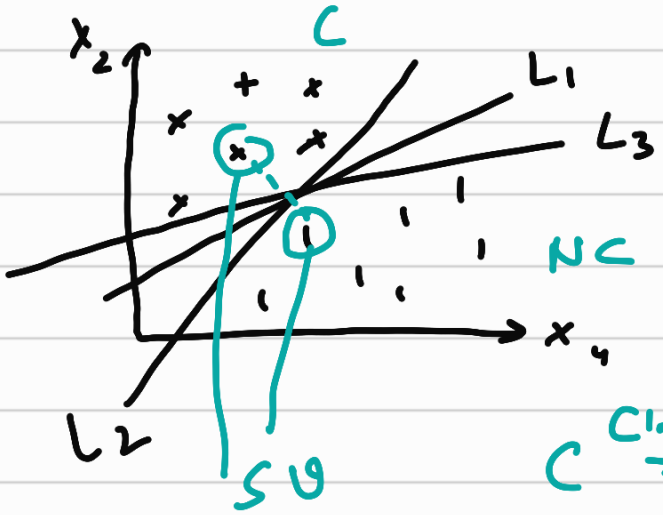
SVM → Support Vector Machine → identify best Line which separate Closed

Classification → Breast Cancer Pred

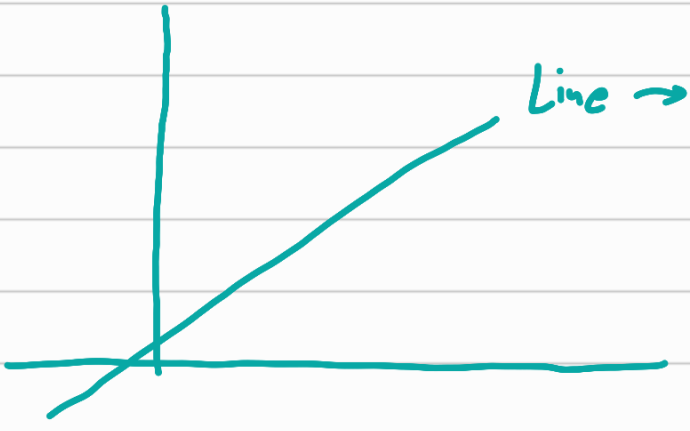


Close to
C → NC
NC → C

3 d. Plane
4.d / 5.d / n.d → Hyperplane

Line which has Maximum separation is better

Eqn ⇒ $W^T x + W_0 = 0$ ⇒ Generalized formula of L_1

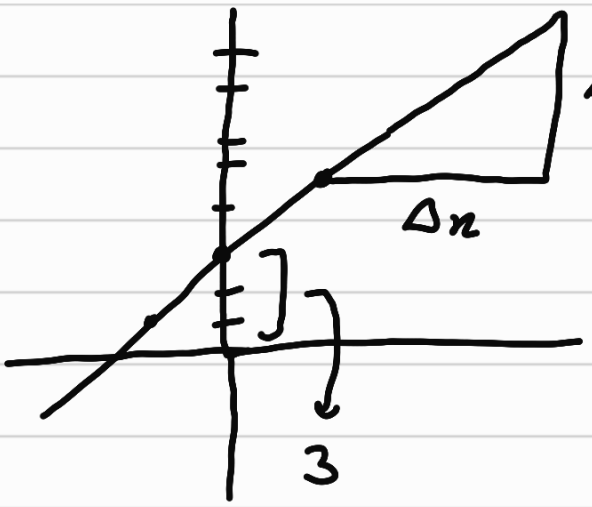


Line → $y = mx + c$
↓ Slope ↓ Intercept

$y = 2x + 3$

Plane

y	x
3	0
5	1
1	-1
7	2



$$m = \frac{\Delta y}{\Delta x} = 2$$

$$y = mx + c \Rightarrow ax + by + c = 0$$

$$y = \frac{-ax - c}{b}$$

$$y = -\frac{a}{b}x - \frac{c}{b}$$

$$ax_1 + bx_2 + c = 0$$

$$\boxed{w_1x_1 + w_2x_2 + w_0 = 0}$$

More Generalized eq^n of Line

3-d $\rightarrow$ $w_1x_1 + w_2x_2 + w_3x_3 + w_0 = 0$

$\downarrow$

n-d $\rightarrow$ $w_1x_1 + w_2x_2 + \cdots + w_nx_n = 0$

$\downarrow$

$$w \cdot x + w_0 = 0$$

$$w^T \cdot x + w_0 = 0$$

$w^T \cdot x = w \cdot x$

$= \|w\| \cdot \|x\| \cdot \cos\theta$

Confusion Matrix :-

Imbalanced dataset → Accuracy is Cursed.

Predicted \ Actual	0	1
0	TN	FN II
1	FP I	TP

NC → NC
↓
True -ve

C → C
⇓
True +ve

I is Critical in Medical Profession

$$\text{Recall} \rightarrow \frac{TP}{TP+FN} = \frac{\text{Correct Predicted +ve}}{\text{Actual Positive}}$$

$$\text{Precision} \rightarrow \frac{TP}{TP+FP} = \frac{\text{Correct Pred +ve}}{\text{Pred +ve}}$$

$$\text{F1-score} \Rightarrow \frac{2 \times P \times R}{P+R}$$

Area Under ROC :-

↗ Received operating Characteristic

Higher AUC
Better Model

ROC Curve

$$SN = \frac{TP}{TP+FN} = \frac{TP}{P}$$

$$SP = \frac{TN}{TN+FP} = \frac{TN}{N}$$

A	B	C	D	E	F
Classifier	Accuracy	Precision	Recall	F1	AUROC
kNN	78	76	76		
DT	79	74	74		
RF	88	67	61		

after Calculating all values then you take Model for further optimization

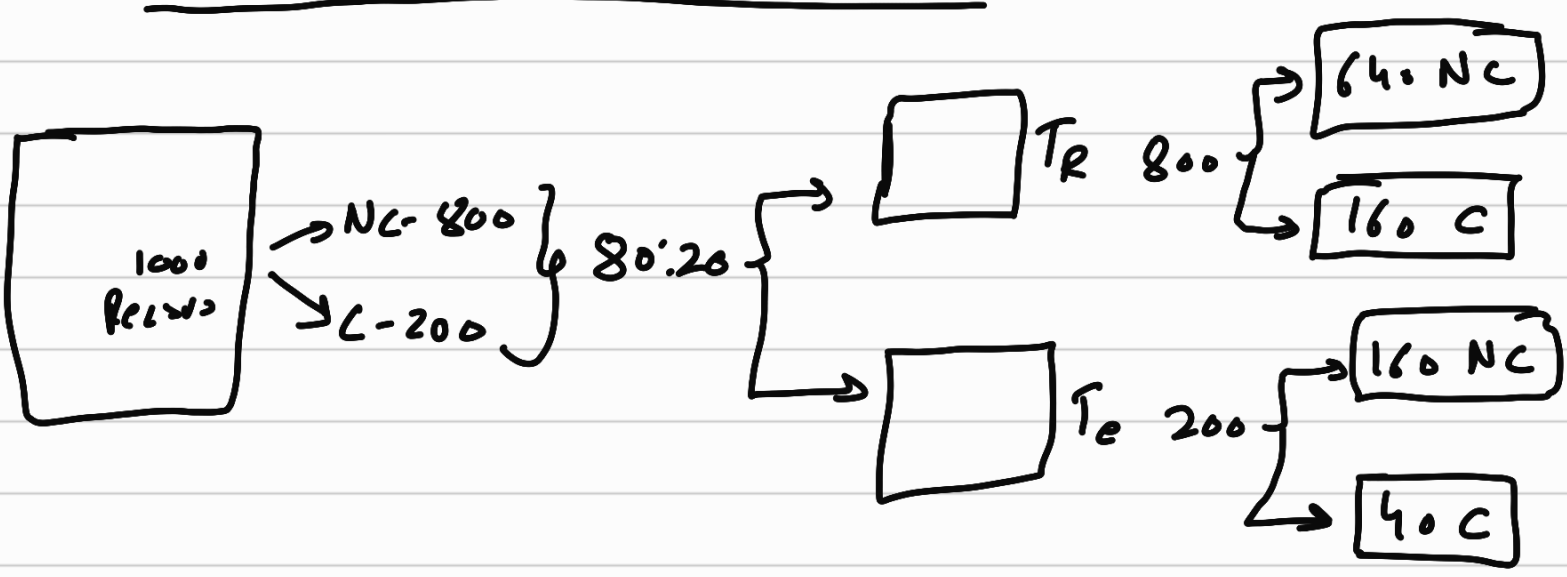
Generate Classification report using sklearn.

You should see which Model is performing better on minority Classes.

Issues in Classification:-

(1) Overfitting:- Training Accuracy is High
Testing Accuracy is Low
↓
Shuffling of data can be also tried

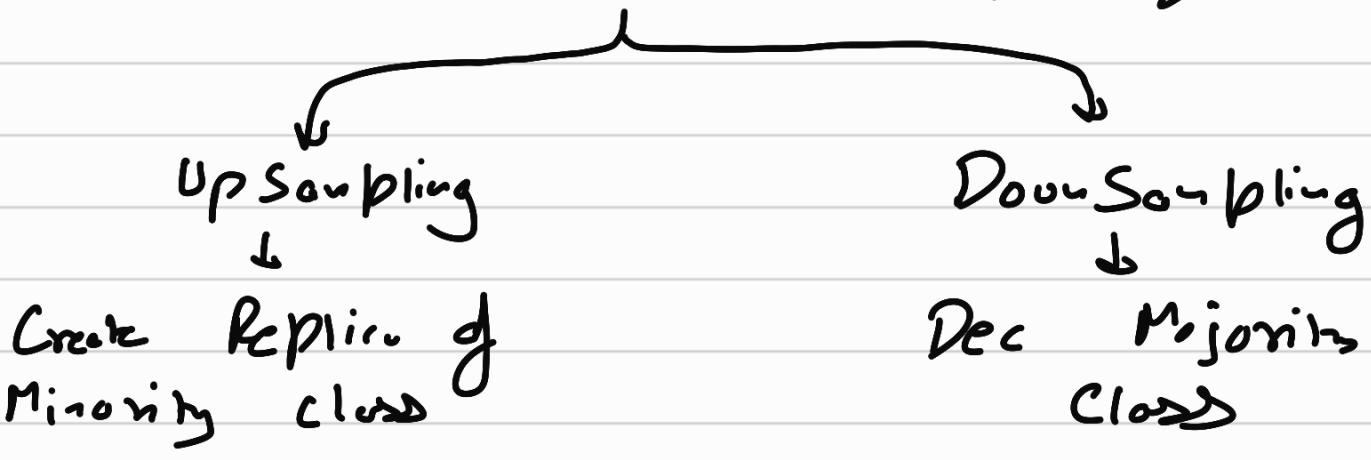
(2) Class Imbalance Problems:-



→ Model will learn less about Cancer Patients

→ So we have to balance the dataset

→ There are two techniques



You can apply both techniques & see Models are performing better on which Dataset.

Lib :- Resample, Smote, Smotein, adasin.

video :- Smote Satyajit.

1. Cross Validation
2. Increasing the data
3. Shuffling of data
4. Feature Selection

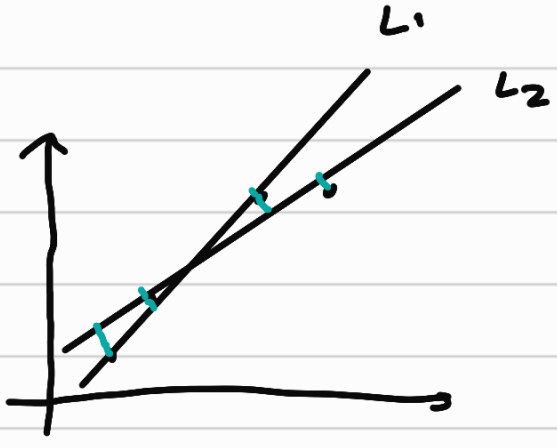
Regression :-

1. Simple Linear Regression (SLR)
2. Multiple Linear Regression (MLR)
3. Polynomial Regression

SLR

House Price Prediction

Area	Price
500	30 L
600	40L
1200	90L
1600	100 L
700 →	?



Cal dis b/w Lines & points add all distance line having Smallest Total distance is best fit.

For more than one feature use MLR

$$y = m_1x_1 + m_2x_2 \cdots m_nx_n$$