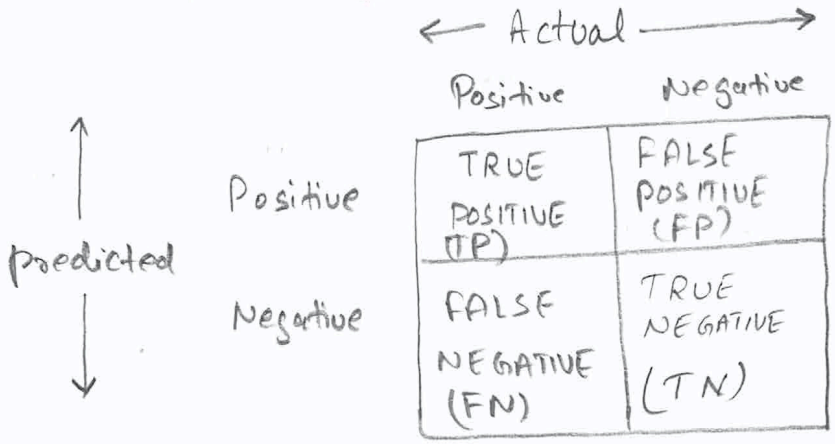
CONFUSION MATRIX



The size of Confusion matrix is determined by no. of thing we want to predict.

If we had to choose from 3 outcomes.

	thing 1	thing 2	thing 3.
thing 1	+	−	−
thing 2	−	+	−
thing 3	−	−	+

Confusion matrix tells us what our NL algorithm did right & what it did wrong.

Sensitivity (Recall) $= \frac{TP}{TP + FN}$

Proportion of actual positives that are correctly identified.

Specificity $= \frac{TN}{TN + FP}$

proportion of actual negatives that are correctly identified.

for 3 outcomes -

	OC1	OC2	OC3
OC1	2	4	6
OC2	8	10	12
OC3	14	16	18.

OC1

$$\text{Sensitivity} = \frac{2}{8+14} \quad ; \quad \text{Specificity} = \frac{(10+12+16+18)}{(10+12+16+18)+(4+6)}$$

OC2

$$\text{Sensitivity} = \frac{10}{4+16} \quad ; \quad \text{Specificity} = \frac{(2+6+14+18)}{(2+6+14+18)+(8+12)}$$

OC3

$$\text{Sensitivity} = \frac{18}{14+16} \qquad \text{Specificity} = \frac{(2+4+8+10)}{(2+4+8+10)+(6+12)}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

proportion of positive predictions were actually correct.

$$\text{Recall} = \frac{TP}{TP+FN}$$

what proportion of actual positives was correctly identified.

Normal Distribution:

1) ND are always centred around average.
2) Width of Curve = Std. Dev

Normal Curves are drawn such that 95% of measurements fall b/w +/- 2 Std. dev. around the mean.

Mean, $\bar{x} = \mu = \frac{x_1 + x_2 + \dots + x_n}{n}$.

Variance = How data is spread around μ $= \frac{\sum (x_i - \mu)^2}{n}$

Std. Dev. $\sqrt{\frac{(x_i - \mu)^2}{n}}$

↳ if data had been for apples, the variance = 100 (apples)^2
∴ we cannot plot this x-axis. ∴ Std. dev.

Populations (vs) Sample:

A population is the collection of all items of interest to our study & is denoted as 'N'. The mean, variance, std. dev ... are population parameters

Sample is subset of population and its denoted = 'n'.

mean, variance ... ⇒ are Statistics.

Population - Calculated

$\mu = \frac{\text{sum of all } x_i}{n}$

Variance: $\frac{\sum (x - \mu)^2}{n}$

Std. dev: $\sqrt{\frac{\sum (x - \mu)^2}{n}}$

Sample = Estimate (since we don't have all data)

$\bar{x} = \frac{\text{sum of } x_i}{n}$

Variance = $\frac{\sum (x - \bar{x})^2}{n-1}$

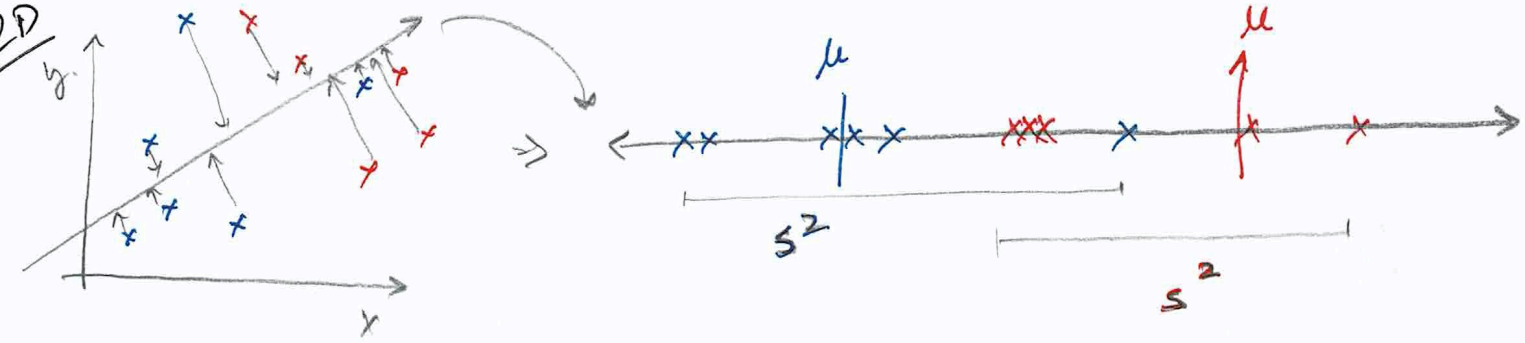
Std. Dev: $\sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$

LINEAR DISCRIMINANT ANALYSIS.

PCA: Reduces dimensions by focusing on variables with most variation.

LDA: Maximizing the separability b/w 2 categories to make best decisions.

2D



LAD creates a new axis & projects data on it.

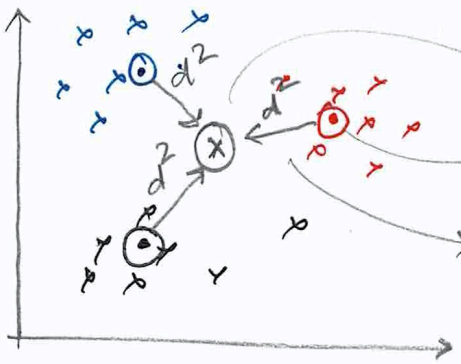
LDA creates this axis -

1) MAX. the dist b/w μ & μ

2) MIN. variation (scatter) b/w s^2 & s^2

$$\Rightarrow \frac{(\mu - \mu)^2}{s^2 + s^2} = \frac{d^2}{s^2 + s^2}$$

3D



Point that is central to all data

point central to each category.

MAX. the distance b/w each category

while MIN. scatter

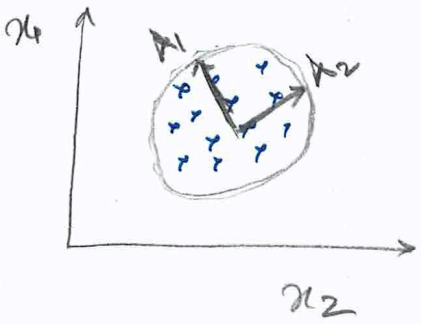
$$= \frac{d^2 + d^2 + d^2}{s^2 + s^2 + s^2}$$

PCA

PC1 accounts for the most variation in data

(A) PCA looks at variables with most variation

(B) PCA - un supervised. ML technique

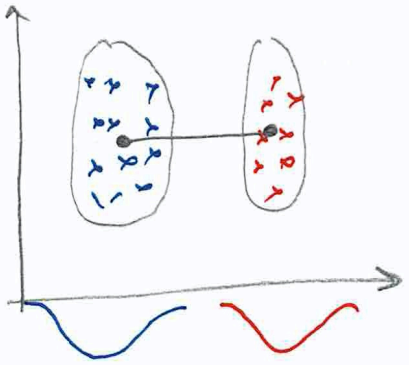


LDA

(A) LD1 (first new axis) accounts for most variation b/w categories.

(A) LDA MAX. the separation of categories.

(Y) LDA - supervised ML technique

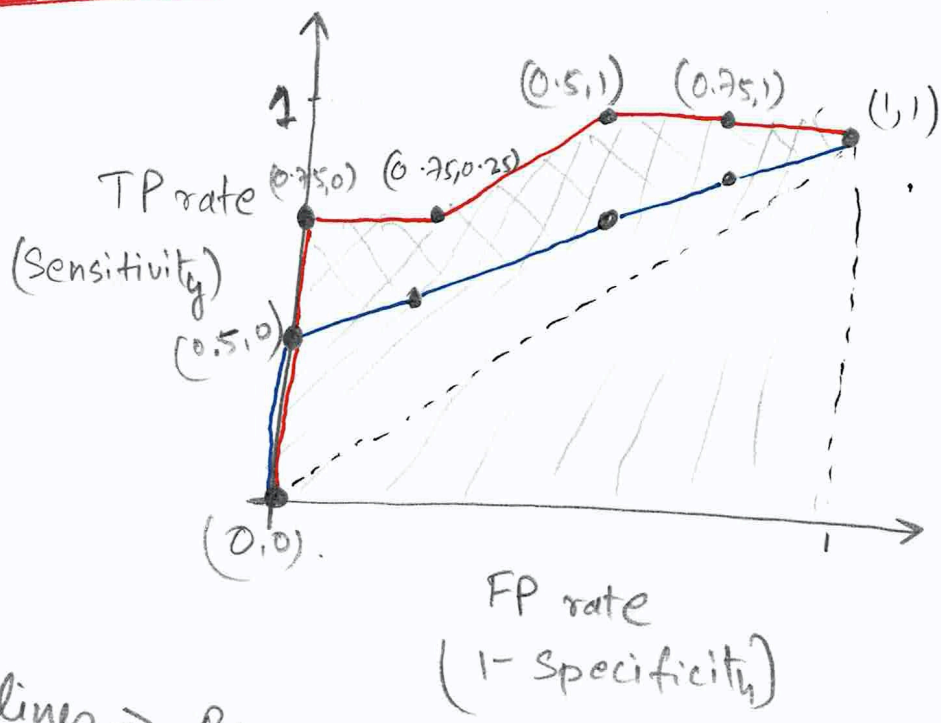


Good projection separates the classes well.

ROC & AUC

ROC - Receiver Operator Characteristics.

AUC: Area under Curve.



the red and blue lines ⇒ ROC graph.

Area under ROC graph ⇒ AUC.

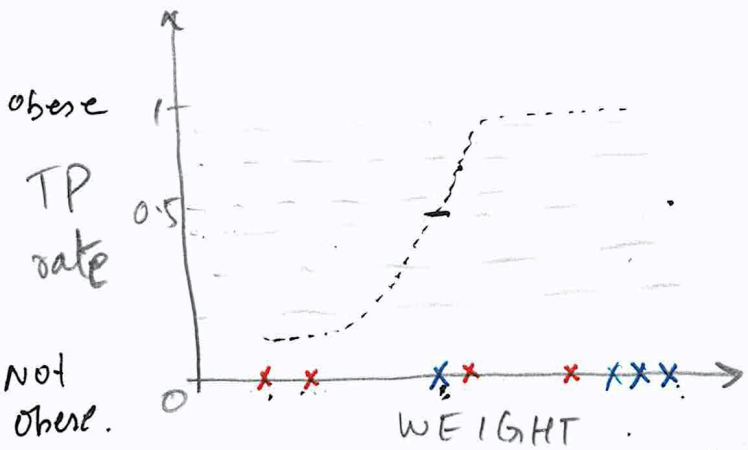
$$TP\ rate = \frac{TP}{TP + FN}$$

Proportion of 'T' samples are correctly classified.

$$FP\ rate = \frac{FP}{FP + TN}$$

Proportion of 'F' samples that were incorrectly classified.

ROC Graph:

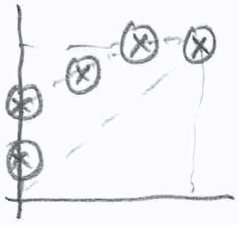


To test the effectiveness of the Logistic Reg. model - we can set a classification threshold=0.5

ie, > 0.5 - obese

< 0.5 - not obese.

Threshold	Confusion Matrix	TP rate	FP rate	Graph.
0	4 | 4 0 | 0	1	1	
0.1	4 | 3 0 | 1	1	0.75	
0.2	4 | 2 0 | 2	1	0.5	
0.3	3 | 1 1 | 3	0.75	0.25	
0.7	3 | 0 1 | 4	0.75	0	

This how the ROC Graph is obtained 

AUC can be used compare two ROC graphs.

From graph on previous page - RED ROC is greater than BLUE ROC

∴ Red ROC is BETTER.

red ROC - logistic regression
blue ROC - random forest

} ROC can be used to compare 2 ML's.

ODDS, LOG(ODDS), ODDS RATIO, LOG(ODDS RATIO).

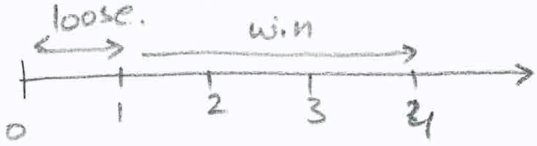
$$\text{odds} = \frac{\text{Some thing happening.}}{\text{Some thing NOT happening}} = \frac{\times\times\times\times\times}{\times\times\times}$$

$$\text{Probability} = \frac{\text{Some thing happening.}}{\text{everything that could happen}} = \frac{\times\times\times\times\times}{\times\times\times\times\times\ \times\times\times}$$

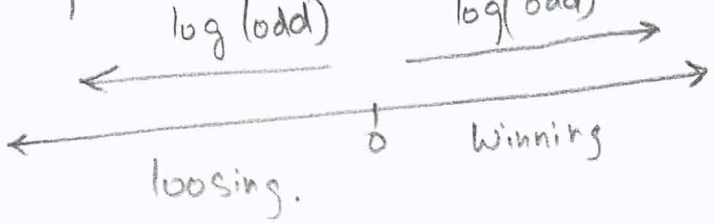
$$\text{odds} = \frac{\text{probability}}{1 - \text{probability}}$$

log (odds) - makes values symmetrical & easier to interpret.

Eg: When a team is playing very bad, the odds of losing may fall b/w 0 & 1 always ; the odds of winning may be >1



taking the log( )



$$\text{ODDS RATIO} = \frac{\frac{xx}{xxxx}}{\frac{xxx}{x}} = \frac{2/4}{3/1} = 0.17.$$

When denominator > numerator ⇒ value b/w 0 & 1

∴ log (odds ratio) ⇒ nice & symmetrical.

Eg:

Has mutated gene	has cancer: Y	has cancer: N
Y	23	117
N	6	210

Use odds ratio to determine relation b/w gene & cancer

Q: If someone has mutated gene, are the odds of them having cancer higher?

if Gene = 'Y', odds they can have cancer = $\frac{23}{117}$

Gene = 'N', " " " " " " = $\frac{6}{210}$

$$\text{odds RATIO} = \frac{\frac{23}{117}}{\frac{6}{210}} = 6.88$$

(odds are 6.88 times greater that someone with muted gene also has cancer.)

$\frac{\log(6.88)}{1} = 1.93$ ← Relation b/w gene/can

* odds ratio - larger value means gene is good predictor of cance

To find find if this relation is statistically significant - i.e odds ratio, log (odd ratio)

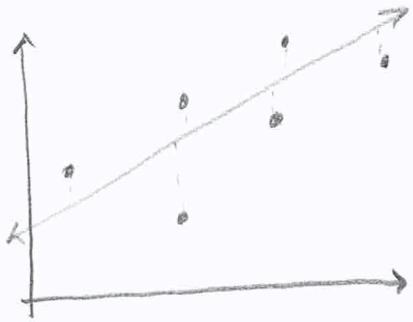
1) Fishers Exact Test.
2) Chi-Square test
3) The Wald test.

BIAS & VARIANCE TRADE-OFF

Bias: the inability for a ML algorithm to capture the true relationship is called BIAS.

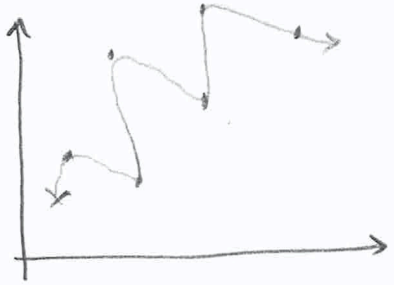
Variance: the difference in fits b/w datasets (test & train dt's) is VARIANCE.

training set



High Bias

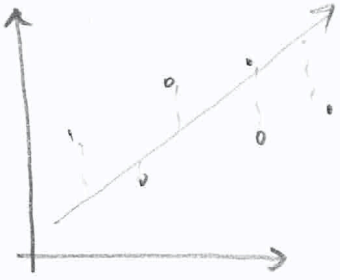
OVERFIT



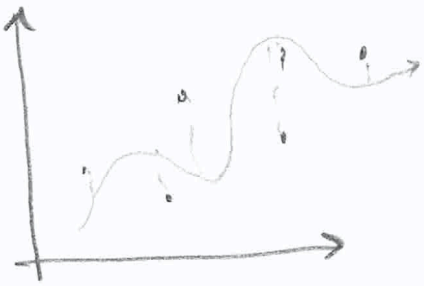
b/c it fi
the
training d
really we

Low Bias

testing set



low variance



high variance.

Best ML model: low bias (accurately model true relationship)
low variability (consistant prediction across various dt's)

Regularization, Boosting, Bagging } methods to find the sweet spot b/w simple & Complicated models.
- to find ML model with low bias/variance

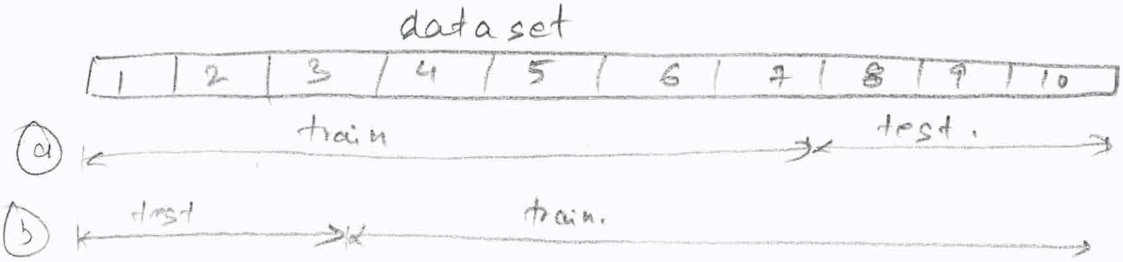
4/20

CROSS VALIDATION.

How do we decide which machine learning mode to use?

- Cross Validation allows us to compare different ML models & get a sense of how well they will perform.

In supervised ML, data set is divided - test & train.



How do we know that first or last 80% of the data for training & last 20% of data for testing is the best option.

∴ Cross validation chooses all of them, one at a time & then summerizes the result.

Since the data set is divided into 10-blocks it is called 10-fold Cross validation

K-fold Cross validation

K - no. of blocks the df is divided into.

Leave one out Cross Validation:-

when the 'K' (no of folds) = no. of observations

- this is only Record when df is small.

Definition.

K-fold - we split df into K different subsets. we use 'k-1' subsets to train df & leave last subset for test. then Average agents each fold & then finalize our model.

Hypothesis Testing

Hypothesis: A claim that we want to test. the current fact

Null Hypothesis: (H_0): Currently accepted value of a parameter.

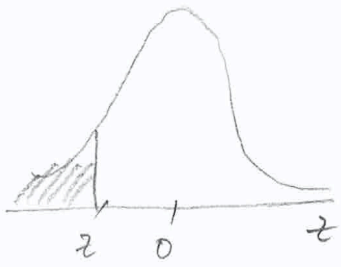
Alternative hyp (H_1/H_a): Research Hyp; Involves the claim to be tested. (what we are proposing to be true)

{possible outcomes: Reject H_0 : p-value < 0.05

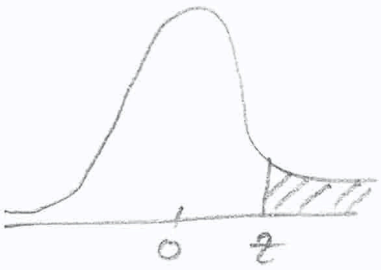
Fail to Reject H_0:

left-tail test

$$z = \frac{\bar{x} - \mu}{(\sigma/\sqrt{n})}$$

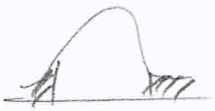


Right-tail test



eg: $z = -1.34$, $H_0: \mu \geq 0.5$, $H_a: \mu < 0.5$

$\therefore$ left tail test.

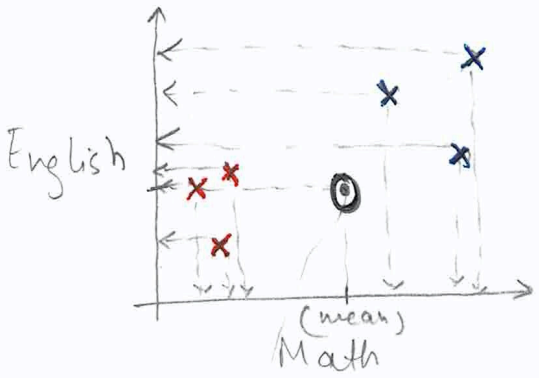
$H_a: \neq$ 

Principal Component Analysis (PCA)

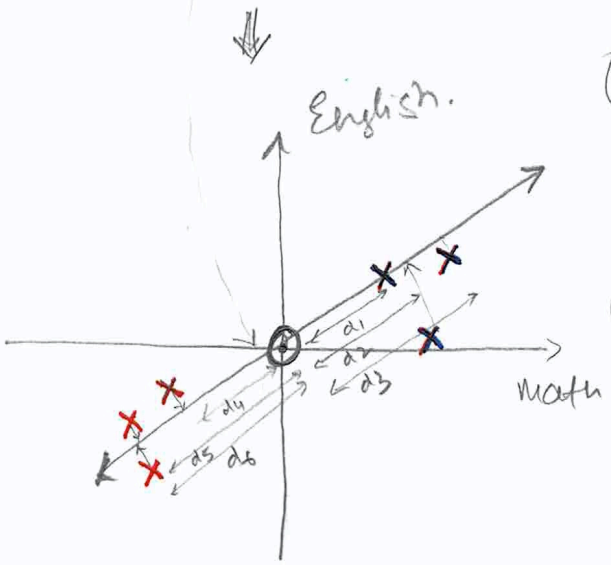
SVD: Singular Value Decomposition

df:

	Math	English	Science	History & variables
Stud 1	10	11	8	3
Stud 2	6	4	5	3.
Stud 3	12	9	10	2.5
Stud 4	5	7	6	2



① Find the mean of Math & English & project the points on x/y axis



② Shift the data so that the mean is in center.

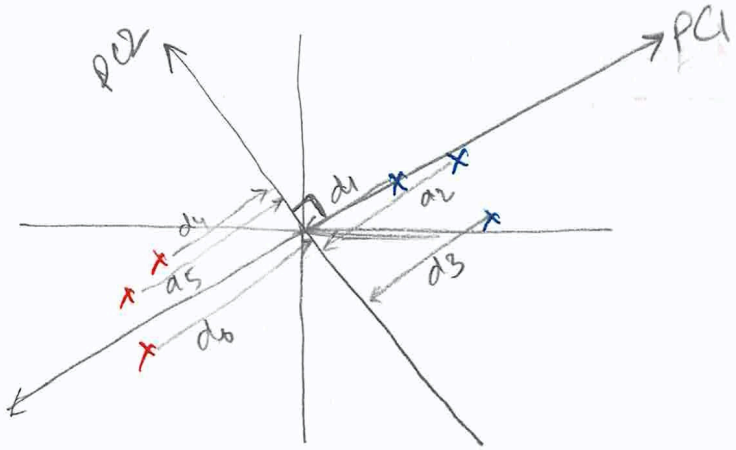
③ we try to fit a line that MAXIMIZES the distance b/w origin (mean) to the points.

ie $d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ = Sum of sq. d

= SS (distance)

= PC1

= Eigenvalue of PC1



(4) PC2 is at 90° of PC1.
Perpendicular

$\therefore d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = PL2$

= Eigenvalue for PC2

PC1 & PC2 now becomes x/y axis & its rotated.
We convert into variation around origin (0,0) ⇒

$\frac{SS(dist\ PC1)}{n-1}$ = variation of PC1 (eg. = 15)

$\frac{SS(dist\ PC_2)}{n-1}$ = variation of PC2 (eg. = 3)

⇒ PC1 accounts for $15/18 = 0.83 = 83\%$ of total variation &
P2C accounts for $3/18 = 0.17 = 17\%$ of total variation.

SCREE PLOT is a graphical resp of % of variation in PCs