Regression models

1) Simple Regression

$y = b_0 + b_1 x_1$ ← independent var

- y ↑ Dependent var
- b_0 ↑ Y intercept (constant)
- b_1 ↑ Slope (coeff)
- x_1 ↑ independent var

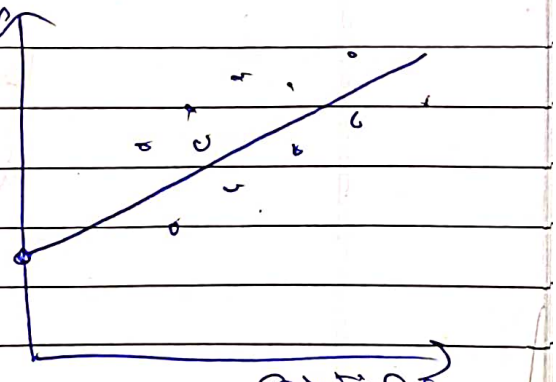
Suppose we want to predict potato yield based on fertilizer used

Potatoes (t) = $b_0 + b_1 \times$ fertilizer

Let $b_0 = 8$ tonnes

$b_1 = 3$ tonne / kg

Graph will be like ↑ potato



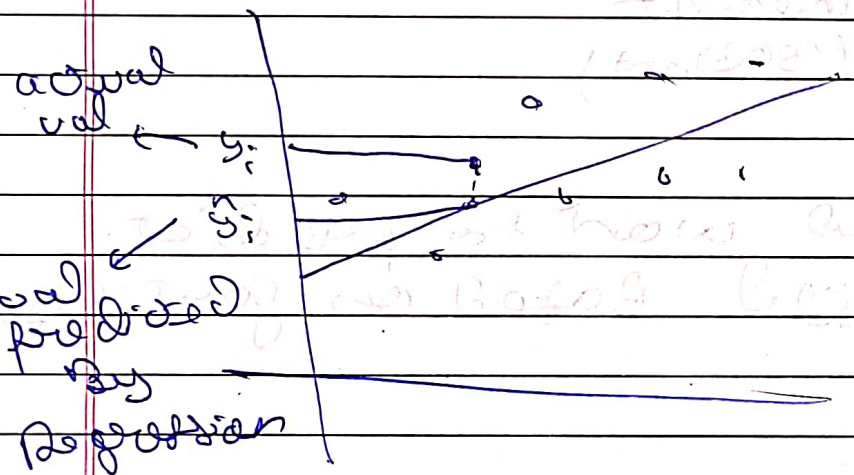
we can say that for every 1 kg of fertilizer potato inc by 3 t

fertilizer →

PAGE No.
DATE / /

Now How to find best Regression line

we do this By ordinary least square



e_i = $y_i - \hat{y}_i$
(Residual)

∴ we need to find

$\hat{y} = b_0 + b_1 x_1$

b_0, b_1 such that

$\sum (y_i - \hat{y}_i)^2$ is minimized

PAGE No.
DATE / /

→ multiple linear Regression

Let's say we have a list of company we have their R&D Spend, Administration, Marketing ~~and Admin~~ | Data and Profit

we want to build a model that will predict the Profit of the company.

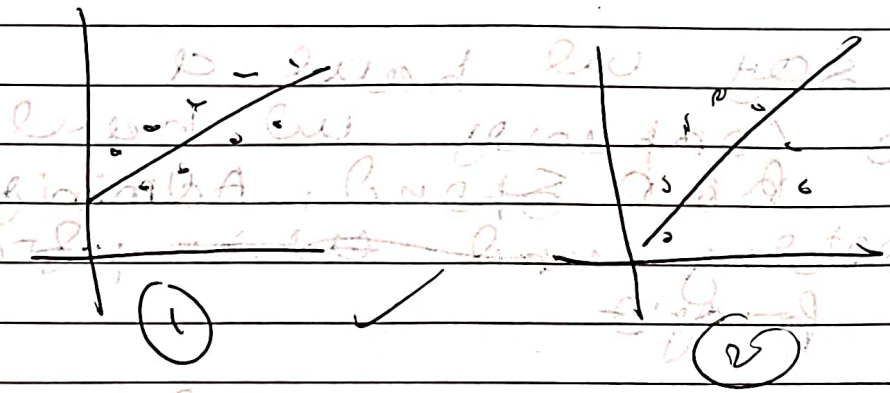
Equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$

- $\hat{y}$: y cap
- b_0 : constant
- b_1 : Slope 1
- x_1 : independent var 1
- b_2 : Slope 2
- x_2 : independent var 2

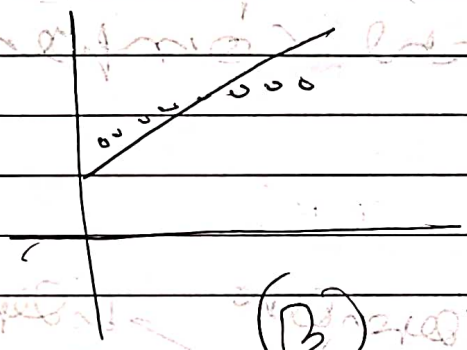
PAGE No.
DATE / /

Assumptions of linear Regression

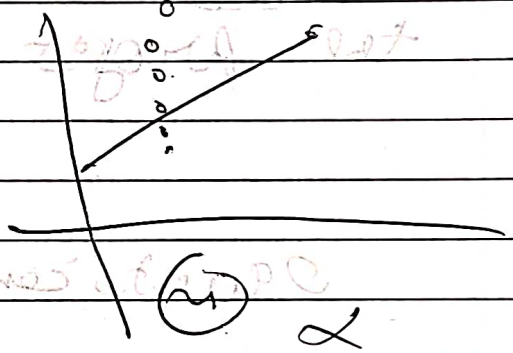


(1) ✓

(2) ✗



(3) ✗



(4) ✓

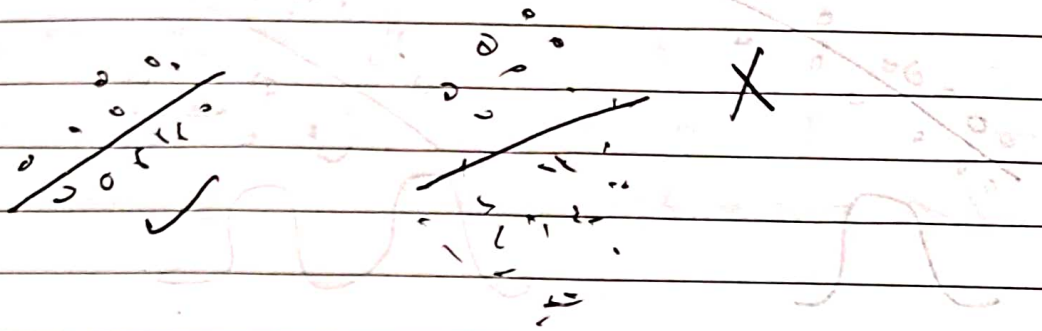
we should use linear Regression in 2, 3 and 4

to make sure if we can use linear Regression we have some checks

PAGE No.
DATE / /

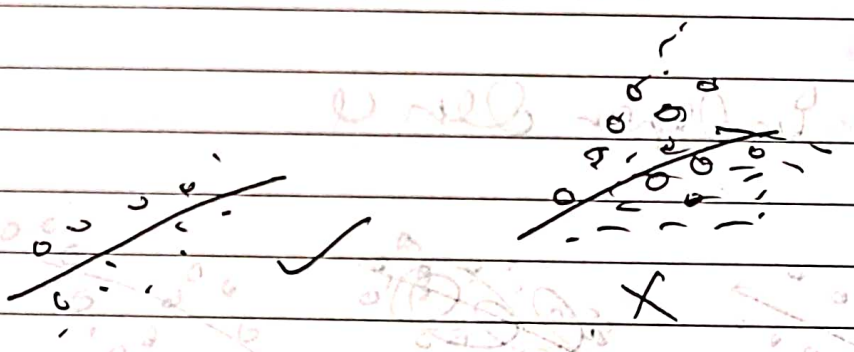
1 Linearity

(linear Relationship between X and Y)



2 Homoscedasticity

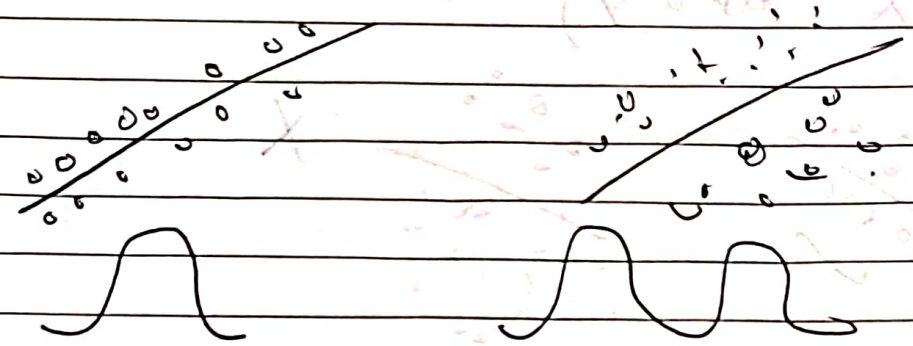
(Equal variance)



we don't want an inc or dec cone shape

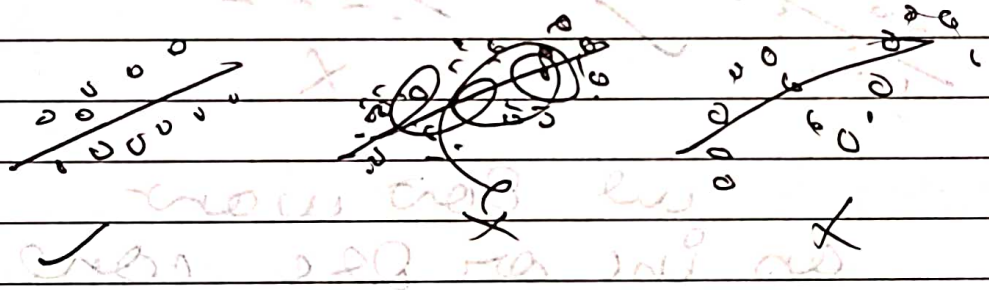
bcz spreader depend on iv dependent var

3] multivariate Normality



(we want normal distribution)

4] Independency



Don't want Patterns
(all var must be independent they should not depend on each other)

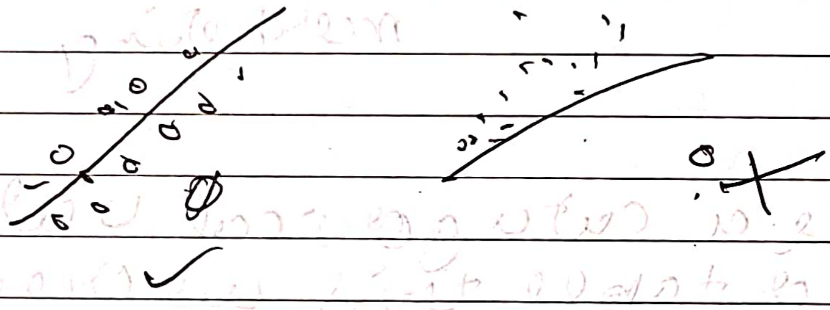
4] 5] Lack of ~~multi~~

5) Lack of multicollinearity

we want independent var not correlated with each other

$X_1 \not\sim X_2$ ✓ $X_1 \sim X_2$ ✗

6] outlier check (optional)



(So decide to remove outliers before or not (depends)

PAGE No.
DATE / /

Dummy variables

we have

Profit	R&D	admin	marketing	State

(New York, California)

to Build Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + ??$$

x_1 → R&D, x_2 → admin, x_3 → marketing

State is a categorical var
So to table this we create
Dummy variables

Suppose we have
2 categories in our
new york and california

So instead of first encode 2 new columns

P	R&D	admin	marketing	NY	CA
				1	0
				0	1
				0	0

1 & 0 for 1 means in NY and 0 means not in NY

So we only include NY and ignore CA

3. Regression Eq

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4D_1$$

this may seem Biased But when 'In NY' column is 0 the const and bo will adjust values of CA (Default val)

Dummy variable trap

if we include both the Dummy variables then there is no default set

and also

$$D_1 = 1 - D_2$$

So By having

$D_1 + D_2$ we get 1

→ P value

consider a coin toss

H0 : fair coin

↗

null hypothesis

Ha : not a fair coin

assuming coin is fair

now we toss a coin and get tails

1 time - 50%.
2 time - 25%.
3 time - 12%.
4 time - 6%.
5 time - 3%.
6 time - 1%.

significance value
$\alpha = 5\%$ (0.05)

if $p < \alpha$ reject null hypothesis

H_0 is wrong

very low (null hypothesis)

these are the p values

if we assume not fair coin

1 time 100%.
2 time 100%.
.
.
.
6 time 100%.

} p value

PAGE No.
DATE / /

→ Building a model

if in multiple linear regression we don't consider all the independent variables because

1) if too many variables will not yield optimal ~~Result~~ Results

2) model becomes too complex

→ 5 methods

1) all-in
2) Backward elimination
3) Forward selection
4) Bidirectional elimination

} Step-wise Regression

PAGE No.
DATE / /

5] Score comparision

→ All in

take all variables

1] if we have good prior knowledge then only use

2] it is compulsory

3] preping for Backward elimination

→ Backward Elimination

STEP-1 :- Select a significance level (α = 0.05)

STEP-2 :- fit all in

STEP 3 :- consider highest p value

if $P > SL(\alpha)$ then STEP 4
or Exit

STEP 4:- Remove var without var with highest p val

STEP 5:- Fit the model without var
(go to step 3)

Forward Selection

STEP 1:- Select & significance val ($\alpha = 0.05$)

STEP 2:- Fit all possible simple regression and select one with lowest p val

STEP 3:- Keep the var and fit all models with extra predictors

Step 4:- Consider lowest p value if $P < SL$ then go to Step 3 otherwise FIN (keep the previous model)

Bidirectional Elimination

Step 1:- Select a significance level to enter and to stay in the model

eg SLENTER: 0.05, SLSTAY: 0.05

Step 2:- Perform step of forward selection

(P < SL enter to enter)

new var

Step 3:- Perform steps of Backward elimination

(old var may need P < SLSTAY to stay)

If no new var can enter or exit then FIN

all possible models

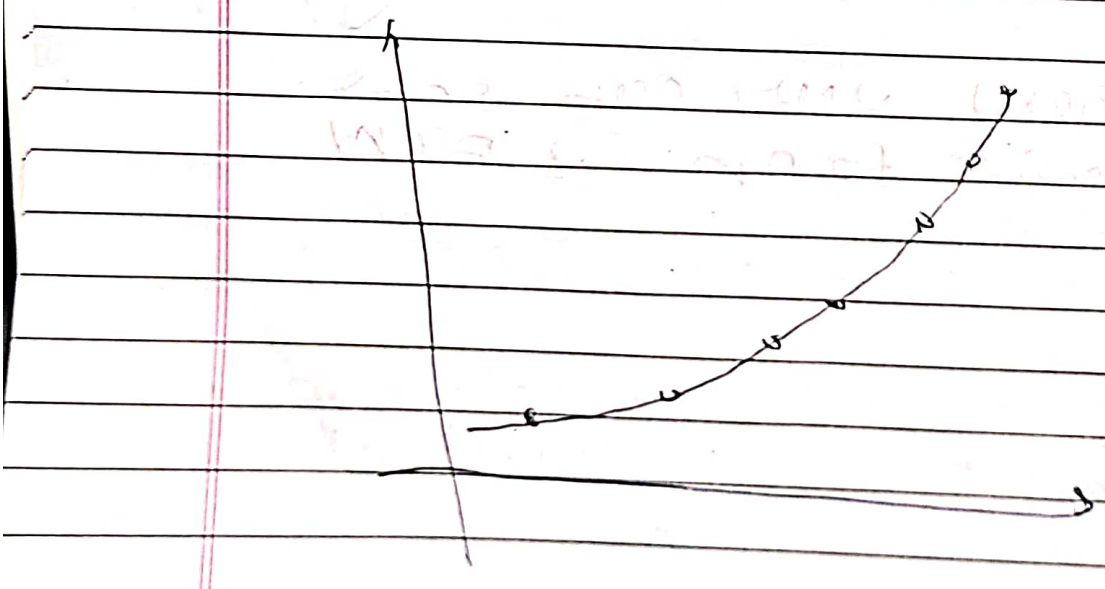
Step 1:- Select criteria
e.g. goodness

Step 2:- construct all models $(2^N - 1)$

Step 3:- Select one with best criteria.

→ Polynomial linear Regression

if we have a Data set like



a curve would fit better than a line

equation

$$y = b_0 + b_1 x_1 + b_2 x_2^2 + \dots$$

why still called linear

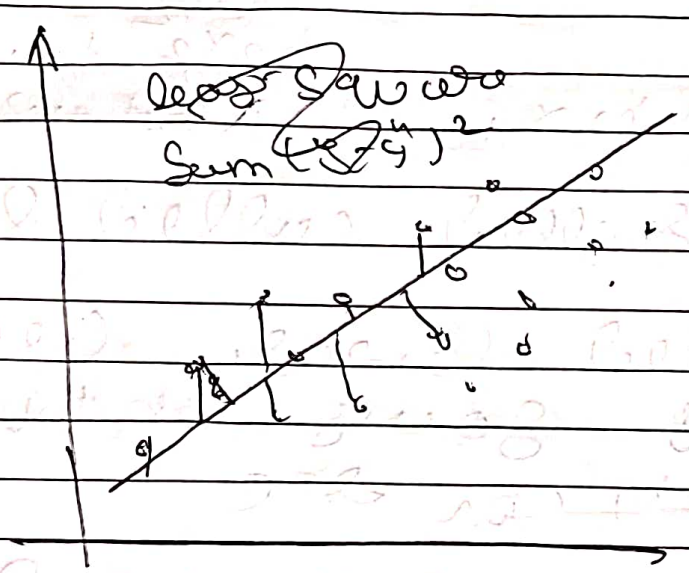
it is called linear because we focus focus on the coeff b_0, b_1, b_2 etc

if coeff are represented as linear combination we call it linear Regression

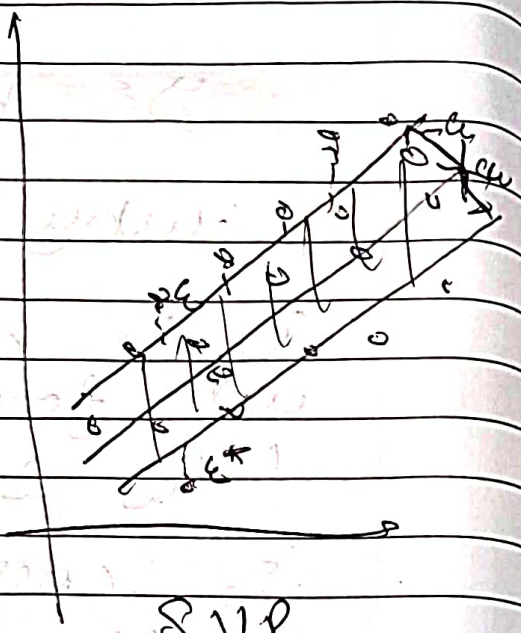
y is linear comb of X

→ Polynomial linear Regression is special case of multiple linear Regression

→ Support vector regression [linear]
is we have



Simple reg



SVR

Both the data are same

In SVR instead of a simple line we have a tube called as ϵ-insensitive tube. Here we ignore errors of data points in the tube (like error of least-square method)

PAGE No.
DATE / /

for linear reg we have

$\sum (y - y^n)^2$

for SVR we have have

$\frac{1}{2}||w||^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \rightarrow min$

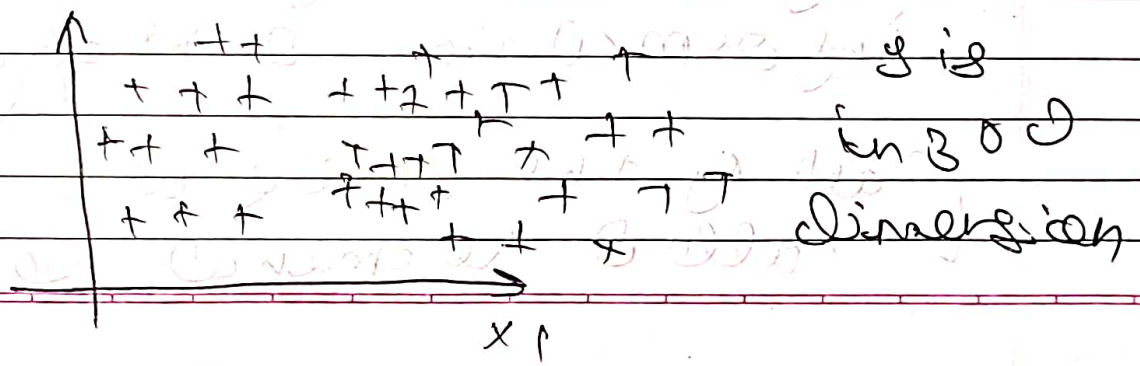
ξ_i — Points above tube

ξ_i^* — Points below tube

~~non linear SVR~~

→ Decision Tree Regression

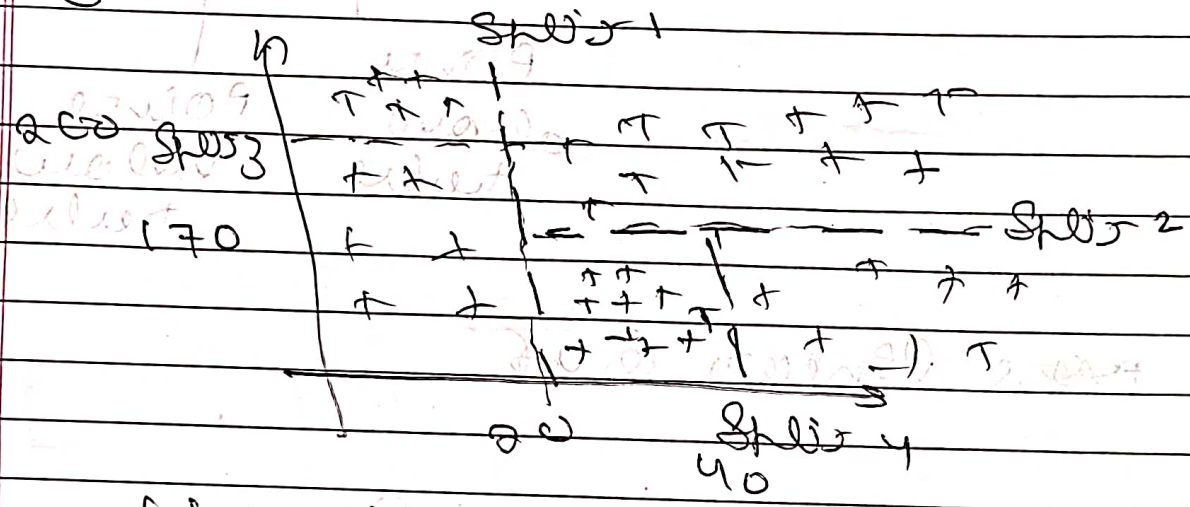
Suppose we have 2 independent variables x_1 and x_2 and we need to predict y



PAGE No.
DATE / /

now a decision tree algorithm will split the data points such that some information is added and it stops when information added is less than %.

So



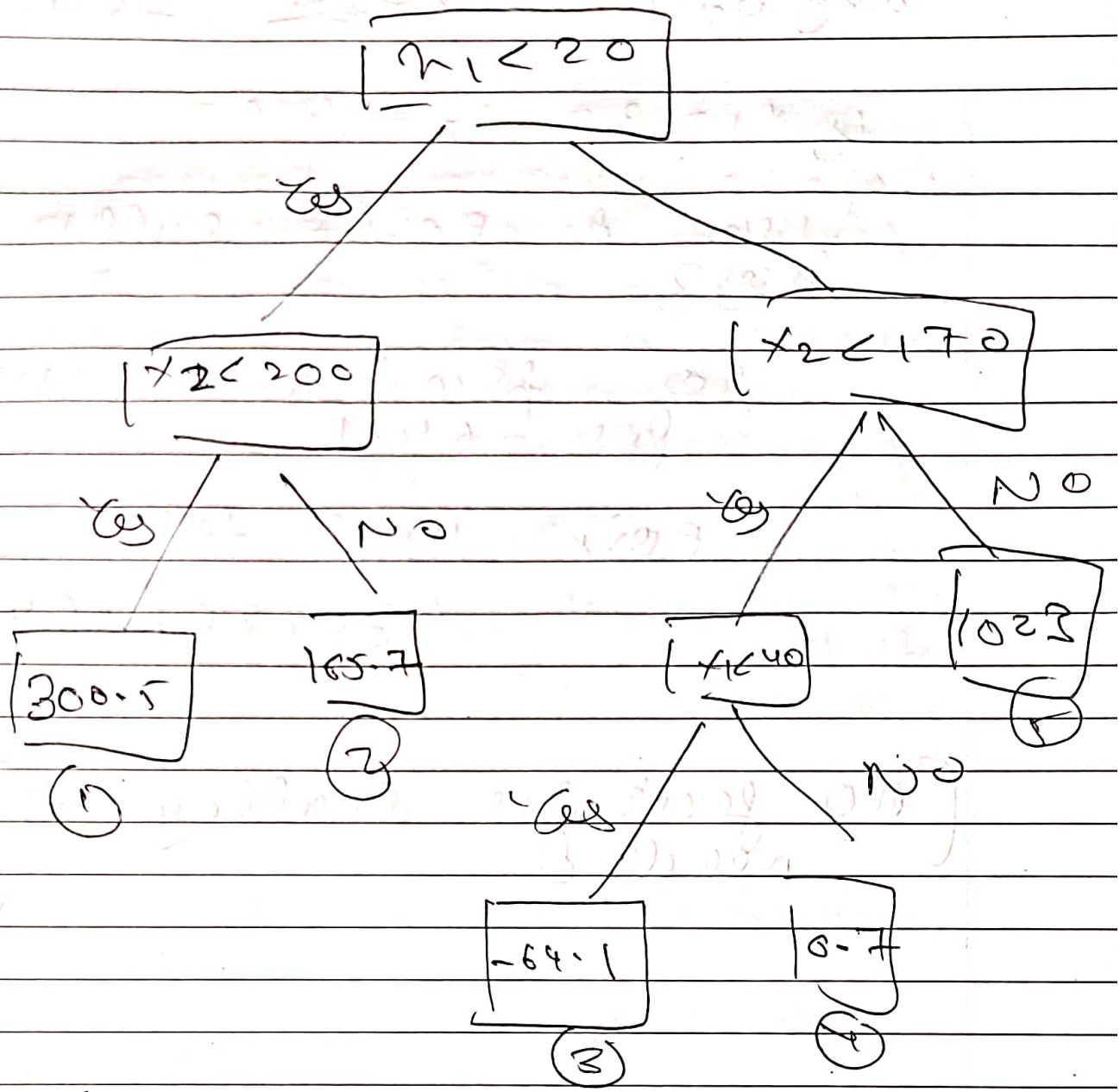
(this is kind of how it happens as we can see it does seems to make some sense)

algorithm uses information entropy

[all of the splits are called terminal leaves]

PAGE No.
DATE / /

decision tree



when we do traversal from any of the path from root to leaf we will receive a split

now every leaf node is value = avg of all the data points

PAGE No.
DATE / /

Let's say leaf 3 has
avg value -64.1

if $x_1 = 30$ $x_2 = 20$

then a point falls on leaf 3

so prediction will be -64.1

Point will be
(30, 20, -64.1)

[no feature scaling is needed]

PAGE No.
DATE / /

→ Random forest Regression

works on Ensemble learning

In this we do

1] Pick Random K data Points from training set.

2] we Build decision tree from these Random Points

3] we choose number N tree & we want to ~~build~~ a Build and repeat steps 1 and 2

4) for a new data point make each of N tree predict the value of y then we take avg of all predicted values and assign the avg to new data Point

By default we take 500 trees

this is better than decision tree because we don't rely on one tree so any change in data can impact single tree But it is very difficult to impact a lot of trees. So and since we take avg of multiple trees certainly impact less