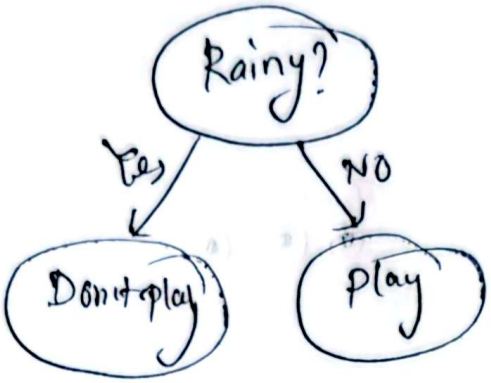
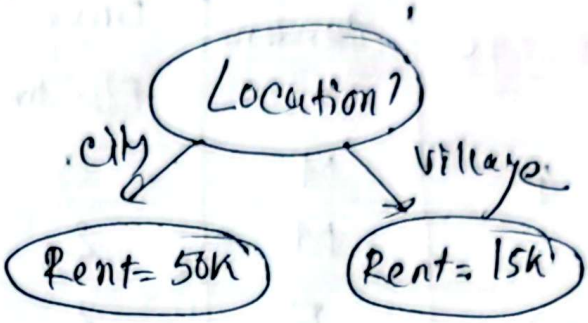
Regression Tree

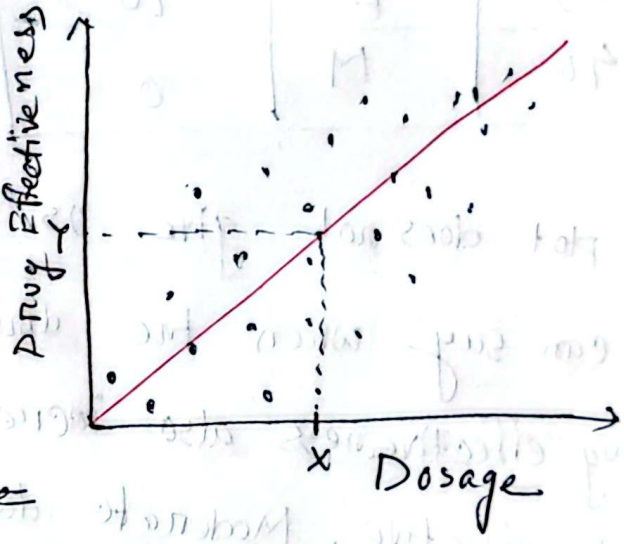




When a decision tree classifies things into categories — Classification Tree

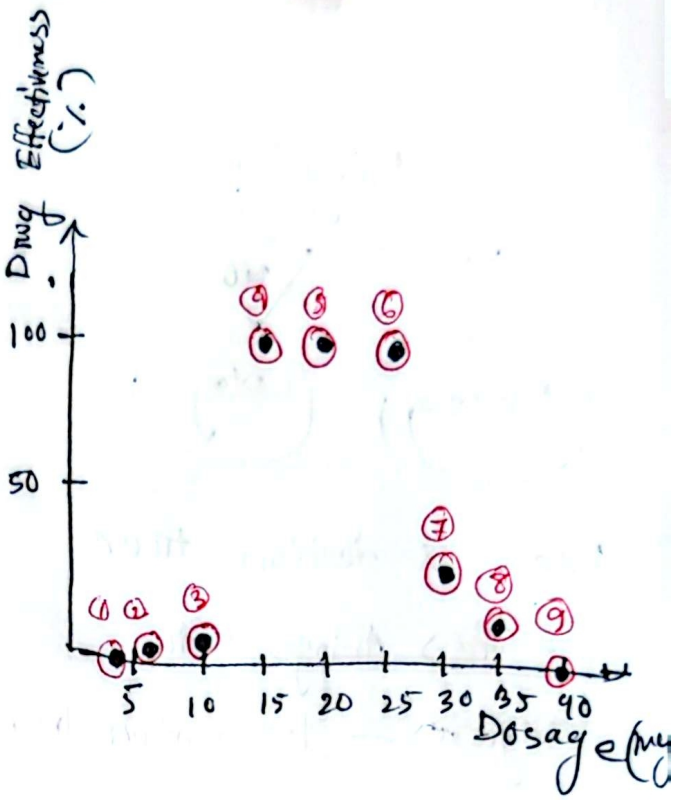
When decision tree predicts numeric values — Regression tree.

If we have a dataset with Drug "dosage" feature and Drug "Effectiveness" as target, which plots like the right picture, then we can easily fit a best fit line. Cause here ~~true~~ when the dosage increases the drug effectiveness also increases. Then for any drug dosage = X, we can easily predict the effectiveness of the drug = Y from this line. This is what we do in Linear Regression.

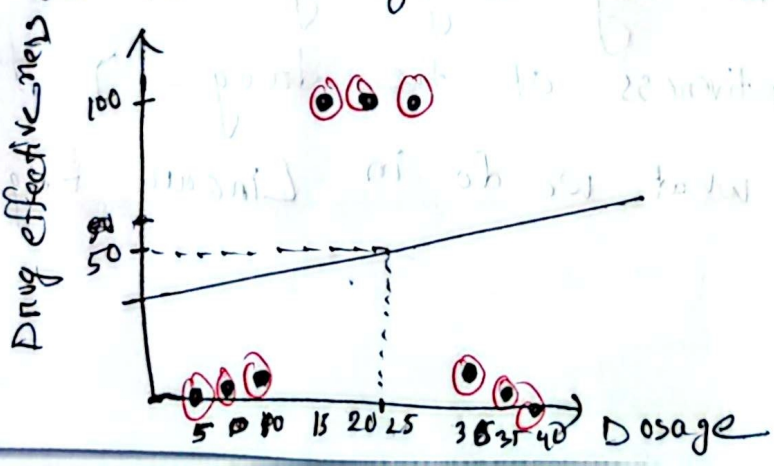


But if we have a dataset like below →

	Dosage	Genden	Drug Effectiveness
①	4	M	0
②	6	M	2
③	10	F	3
④	15	F	100
⑤	20	M	100
⑥	25	M	100
⑦	30	F	20
⑧	35	F	10
⑨	40	M	0

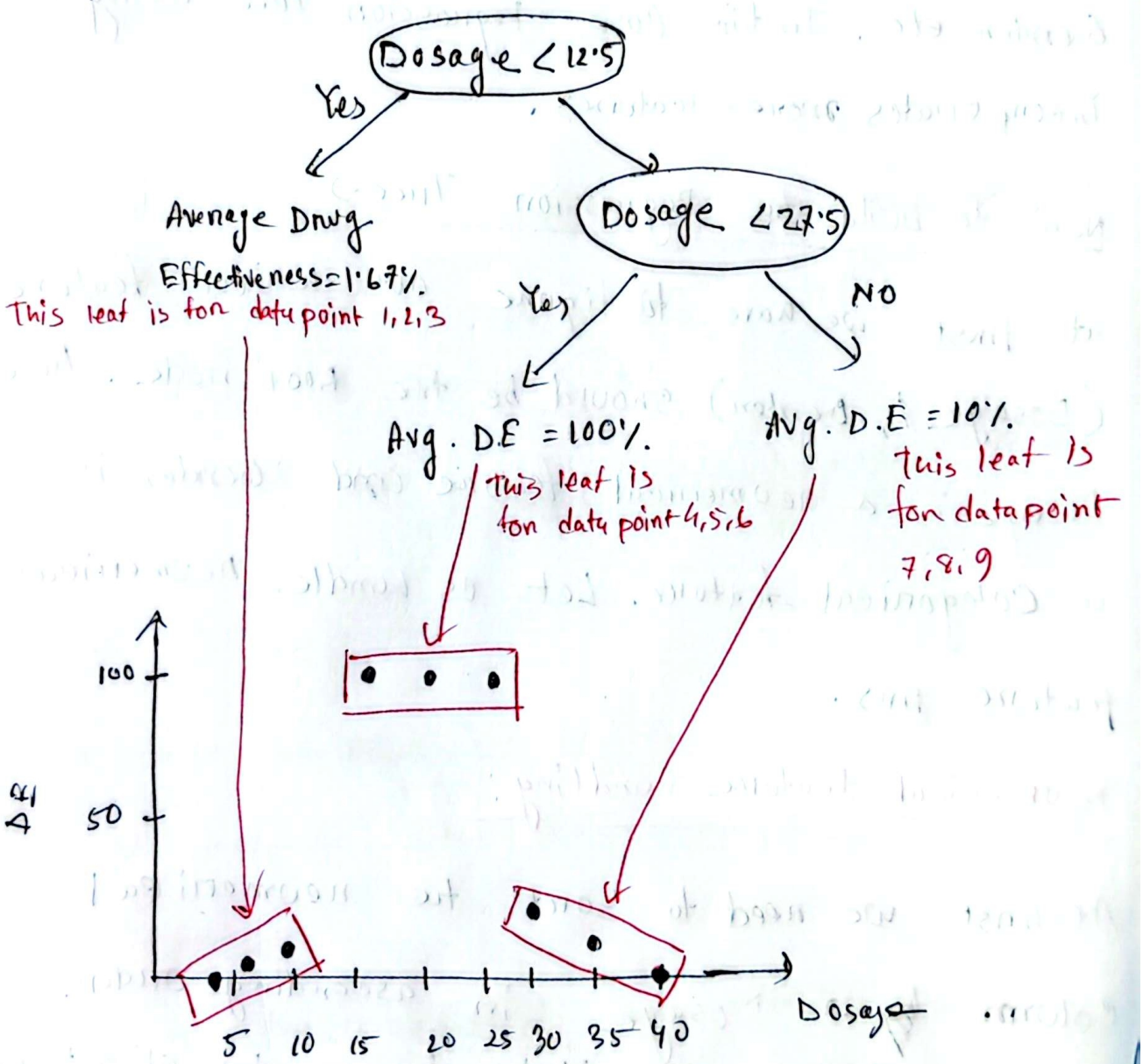


The plot does not give us an uniform cluster where we can say when the drug dosage increases the drug effectiveness also increases. Here low dosage → not effective, Moderate dosage → works well, and high dosage → works moderate to does not work at all. If we fit a straight line here →



The straight line will not be able to predict drug effectiveness. Here if we predict for dosage = 23 mg, we get effectiveness = 50% (approx). But we know that age dosage should give us 100% effectiveness.

So, we can use Regression Tree here to make prediction as the target column is numerical. Each leaf of the regression tree predicts numeric values.



Now if we want to predict on dosage = 23 mg, from the decision tree we get the drug effectiveness = 100%.

Even though we can predict this values only by looking the graph/plot, but it gets much more complicated when there are more features in the dataset like Gender etc. In this case Regression tree easily incorporates more features.

Now to build the Regression Tree ⇒

at first we have to figure out which feature (Dosage ⓑ, Gender) should be the Root node. Here Dosage is a neumerical feature and Gender is a categorical feature. Let us handle neumerical feature firs.

Neumerical Feature handling :

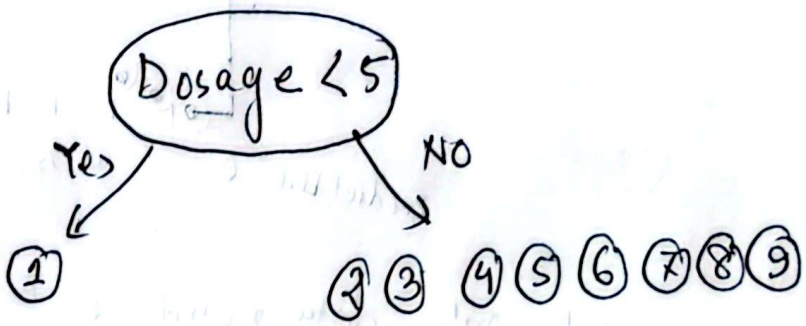
At first we need to sort the neumerical column ⓑ → Dosage, in ascending order.
Then we need to find the midpoints between two adjacent points.

Dosage → 4 6 10 15 20 25 30 35 40

Mid points → 5 8 12.5 17.5 22.5 27.5 32.5 37.5

Now we have to set every mid point as a threshold for the dosage column to see how well it splits the data. To measure which threshold split better, we will calculate SSR = sum of square Residual.

Let us take the first mid point 5 as threshold for the root node →



Only the first data point is less than 5, so it comes to the left leaf. all other data points goes to the right branch.

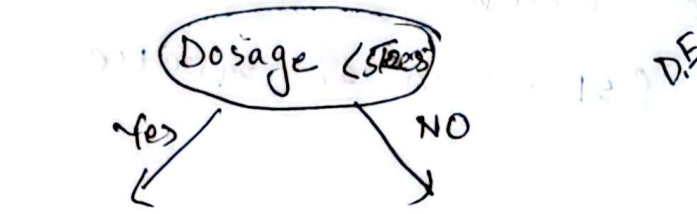
We calculate the predicted drug effectivess for every branch by taking the average of the

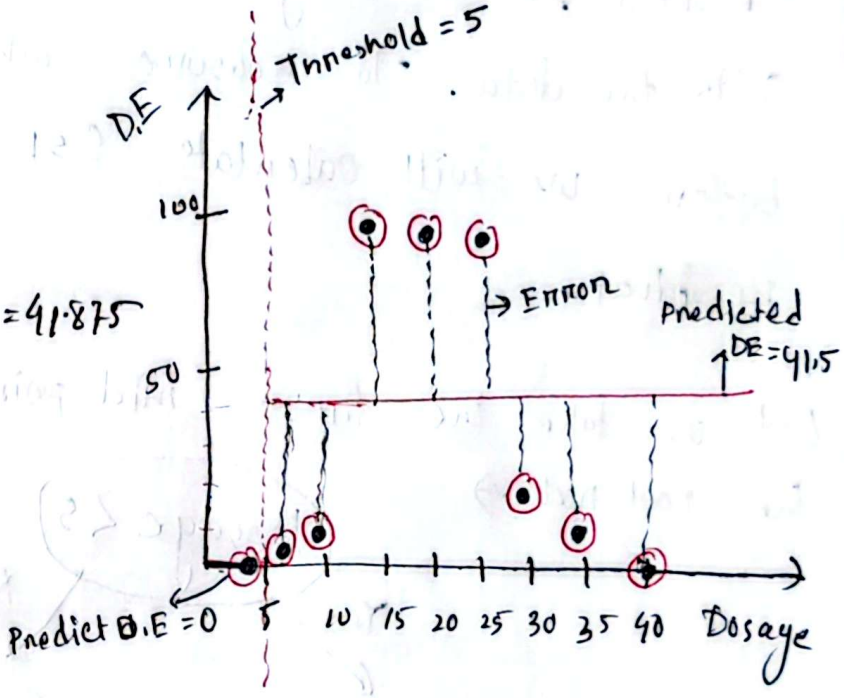
drug effectiveness of data points in each branch.

left branch → data point (1) → avenge D.E = $\frac{0}{1} = 0$

Right " → data point (2)(3)(4)(5)(6)(7)(8)(9) → avenge D.E = $\frac{2+3+100+100+100+20+10+0}{8}$

$= 41.875$.





In the left branch only data point 1 is present, which has actual D.E = 0. The predicted D.E = 0 as well.

If we calculate error between Actual D.E and predicted D.E using SSR we get →

$SSR_{left\ leaf} = (0 - 0)^2 = 0$

Actual D.E — Predicted D.E

In the right branch we have all the data points, and the actual drug effectiveness are ① → 2, ② → 3, ④ → 100, ⑤ → 100, ⑥ → 100, ⑦ → 20, ⑧ → 10, ⑨ → 0. But the predicted D.E = 41.875. The are huge errors present between actual D.E and predicted D.E, which we can calculate using SSR.

Actual Predicted

$$SSR_{Right} = (2 - 41.875)^2 + (3 - 41.875)^2 + (100 - 41.875)^2 + (100 - 41.875)^2 + (100 - 41.875)^2 + (20 - 41.875)^2 + (20 - 41.875)^2 + (0 - 41.875)^2$$

(1) (2) (3) (4) (5) (6) (7) (8)

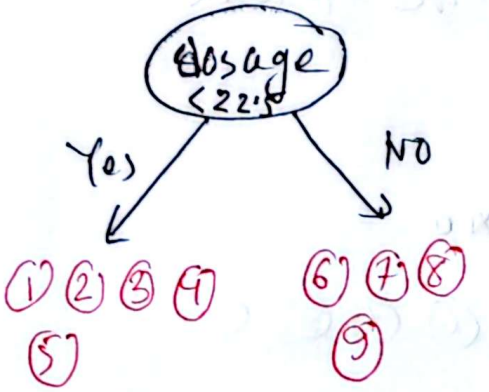
$$= 16485$$

So, $SSR(Dosage < 5) = SSR_{left} + SSR_{Right}$

$$= 0 + 16485$$

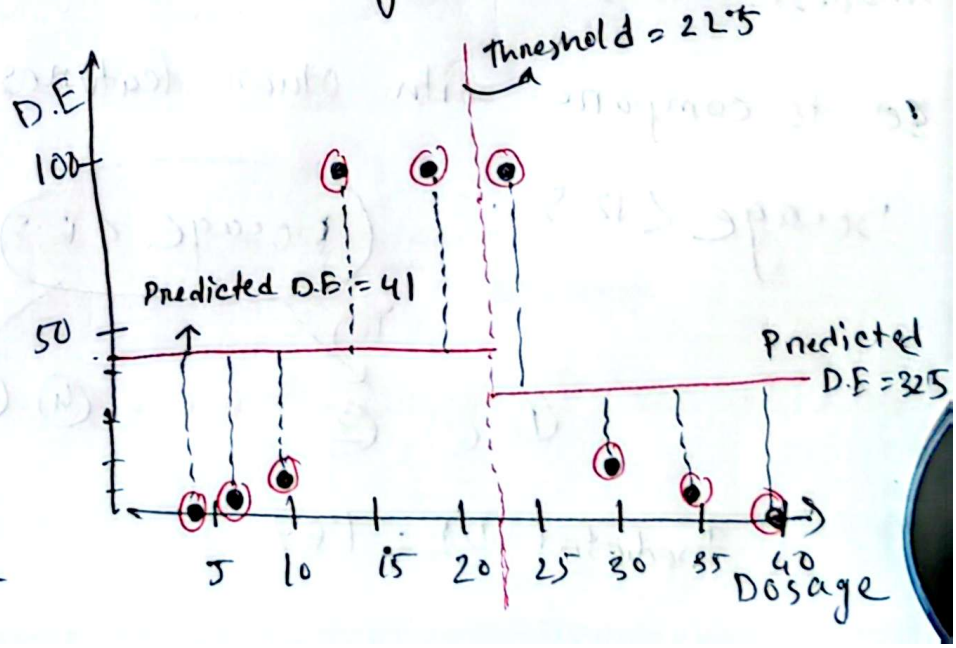
$$= 16485.$$

Now, let us calculate SSR for dosage < 22.5



Predicted D.E = 41

Predicted D.E = 32.5



SSR(Dosage < 22.5) = left branch $\begin{cases} (0-41)^2 + (2-41)^2 + (3-41)^2 + (100-41)^2 \\ + (100-41)^2 \end{cases}$

Right branch $\begin{cases} + (100-32.5)^2 + (12.0-32.5)^2 + (10-32.5)^2 \\ + (0-32.5)^2 \end{cases}$

$= 17883$

If we calculate SSR for all other threshold we get,

Dosage < 5 → 16485

Dosage < 8 → 14670

Dosage < 12.5 → 12355

Dosage < 17 → 17172

Dosage < 22.5 → 17883

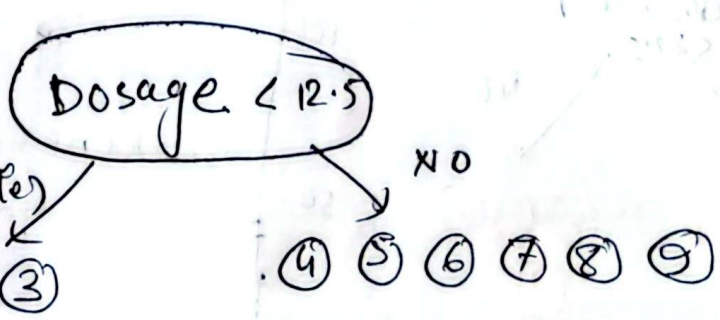
Dosage < 27.5 → 14709

Dosage < 32.5 → 15374

Dosage < 37.5 → 16485

Among this SSR (Dosage < 12.5) = 12355 is the smallest. So, this threshold will give us lowest error.

So, to compare with other features we pick Dosage < 12.5.

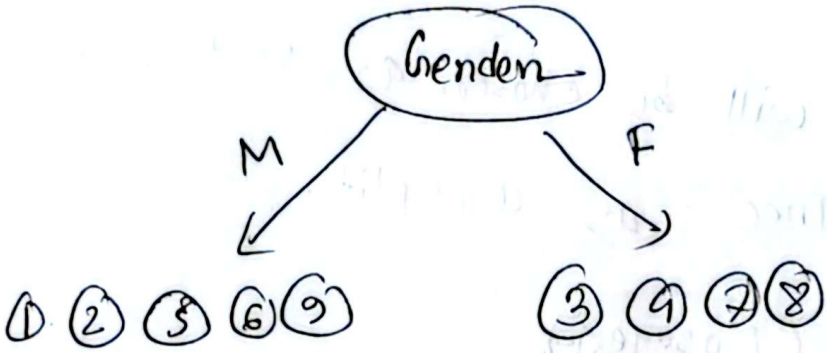


Predicted D.E = 1.67

Predicted D.E = 55

Categorical Feature handling :

handling categorical feature is easy as we can just split the datapoint according to the categories of this feature like below.



Average of Drug effectiveness of these points →

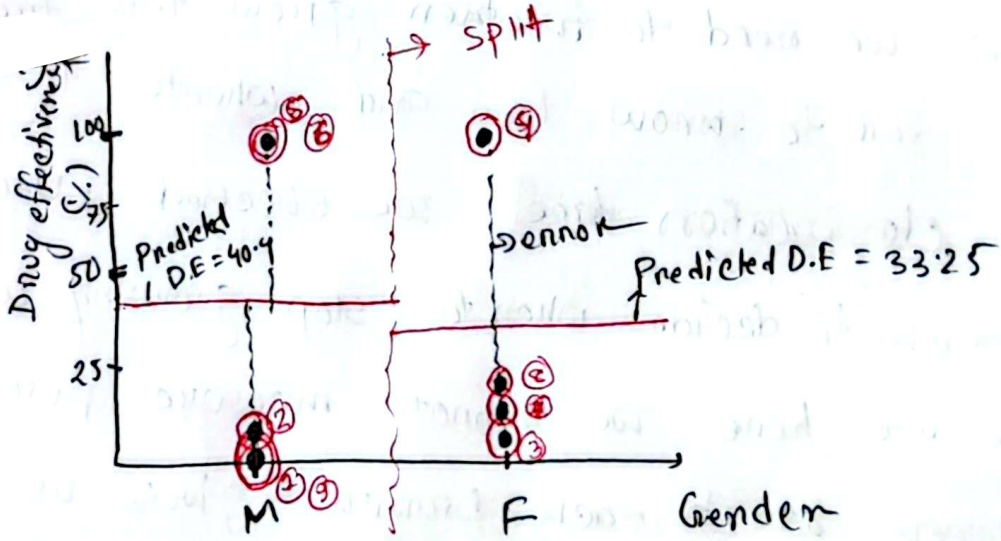
$$\text{predicte D.E} = \frac{0+2+100+100+0}{5}$$

$$= 40.4$$

Average of Drug effectiveness of these data points →

$$\text{predicted D.E.} = \frac{3+100+20+10}{4}$$

$$= 33.25$$



$$\therefore SSR(\text{Gender}) = (0-40.4)^2 + (2-40.4)^2 + (100-40.4)^2 + (100-40.4)^2 + (0-40.4)^2$$

Gender = M

$$+ (3-33.25)^2 + (100-33.25)^2 + (20-33.25)^2 + (10-33.25)^2$$

Gender = F

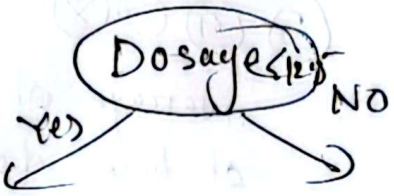
SSR(Gender) = 17929.25

SSR of all features:

SSR(Dosage < 12.5) = 12.355

SSR (~~Dosage~~ Gender) = 17929.25

∴ Dosage < 12.5 will be chosen as root node of the Regression Tree as a split.



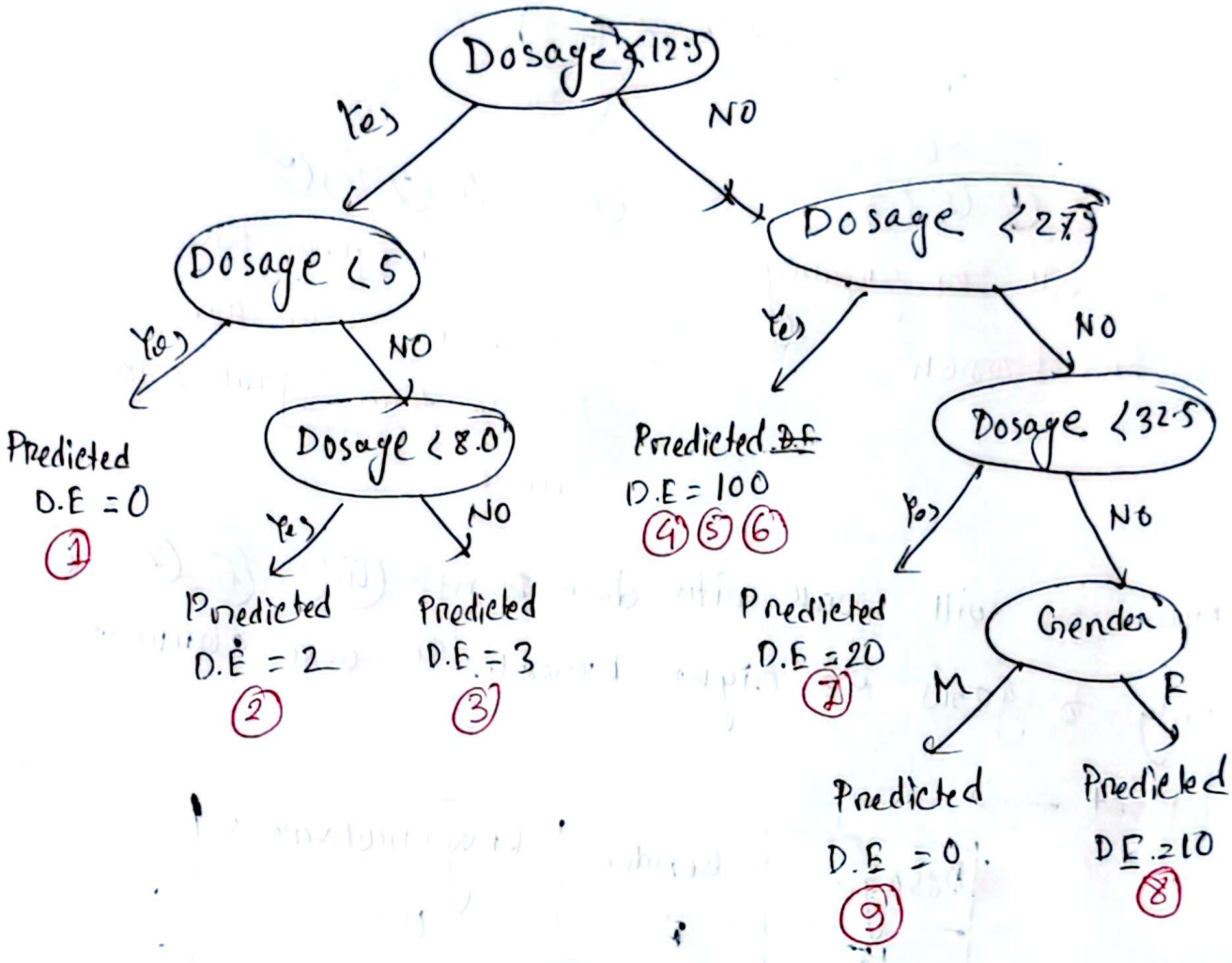
Predicted D.E = 1.67 ① ② ③

Predicted D.E = 55 ④ ⑤ ⑥ ⑦ ⑧ ⑨

Now we need to further grow this tree. But we need to know how and when to stop. In classification tree we checked purity of a branch to decide when to stop growing a tree. However here we cannot measure purity of a branch as each branch gives us a numerical prediction, rather than any class.

~~So to know~~ Moreover if we keep growing

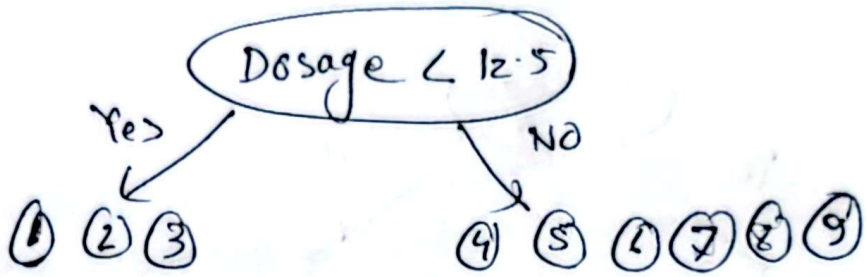
we might end up with a tree like this →



Here we can see, each leaf ends up with one sample/data point except the 'Yes' branch of 27.5 as (4) (5) (6) has d.E = 100. So, this Regression tree has overfitted into our training dataset and it will not work well for unknown data.

So To solve this issue, like the Classification tree we can set limit for leaf nodes like, "Do not split further if there is 3 or less than 3 data point in a leaf". If we enforce this limit

in ~~this~~ the ~~first~~ 2nd level of the tree →



we will stop growing this branch

we will grow this branch as we have more than 3 data points here

Now we will work with data point 4 5 6 7 8 9 only to grow the right branch. So our current dataset →

Dosage	Gender	Drug effectiveness
15	F	100
20	M	100
25	M	100
30	F	20
35	F	10
40	M	0

Again we have to calculate SSR (Gender) and SSR (Dosage < threshold) and compare those to decide the node of this branch. As Gender was not used as node before and dosage feature can have different thresholds, we will consider both.

Calculating SSR for Dosage feature →

Let us sort the dosage feature again and and find the midpoints →

Dosage → 15 20 25 30 35 40

17.5 22.5 27.5 32.5 37.5

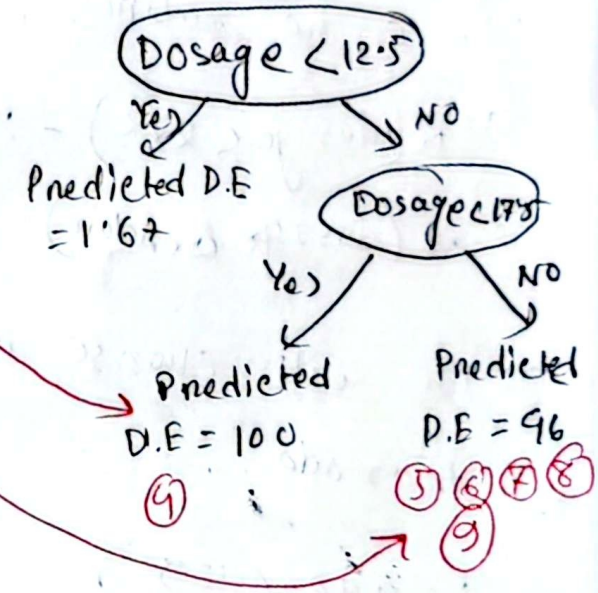
Now we have to calculate SSR by setting all of these midpoints as threshold.

Let us calculate for →

SSR (Dosage < 17.5)

$= (100-100)^2$
$+(100-46)^2+(100-46)^2+(20-46)^2$
$+(30-46)^2+(0-46)^2$

$= 9920$



Similarly if we calculate for all of the midpoints we get →

Dosage < 17.5 → 9920

Dosage < 22.5 → 6275

Dosage < 27.5 → 200 → Lowest

Dosage < 32.5 → 7425

Dosage < 37.5 → 13076

calculating SSR for Gender feature →

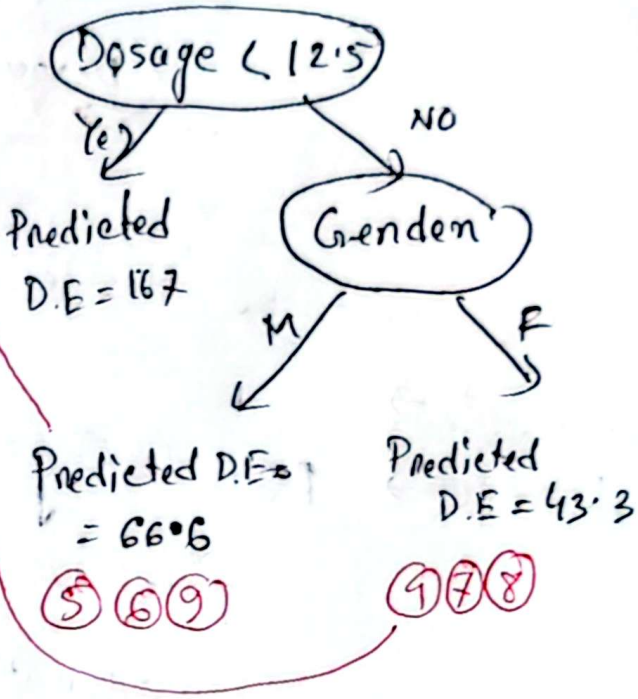
SSR (Gender) =

$= (100 - 66.6)^2 + (100 - 66.6)^2 + (0 - 66.6)^2$

$+ (100 - 43.3)^2 + (20 - 43.3)^2 + (10 - 43.3)^2$

$= 6666.68 + 4866.67$

$= 11533.35$

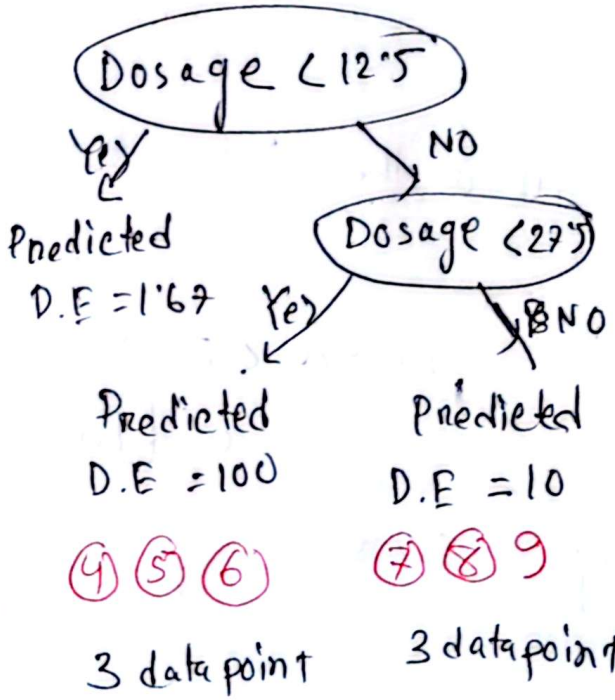


∴ SSR of the features for third level node →

SSR (dosage < 27.5) = 200 ⇒ Lowest

SSR (~~dosage~~ Gender) = 11533.35.

So, we will choose lowest SSR, dosage < 27.5 as the node..



As the leaf nodes have 3 data points each, which is our limit, we will not grow the tree further.