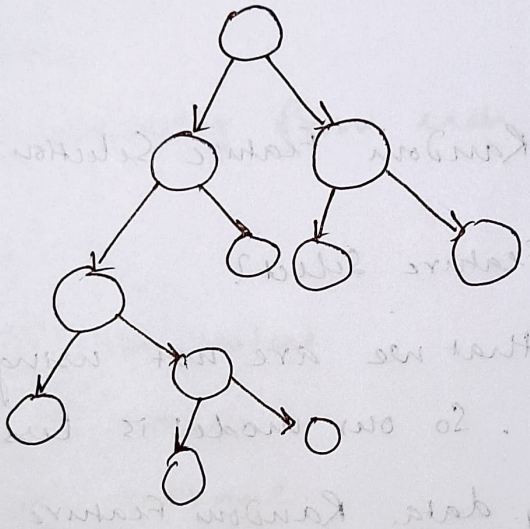
Random forests

→ Let Dataset be

id	x_1	x_2	x_3	x_4	x_5
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0

→ D (dataset)

Let the decision tree for this data be something like this



But for small change in data, ie dotted part, the tree becomes completely invalid. Moreover the algorithm of decision tree is characterised by low bias but high variance. To reduce this, random forests are used:

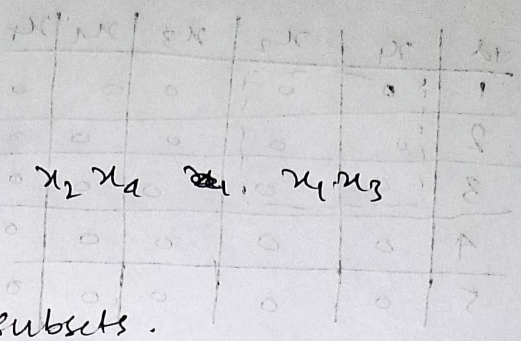
We, take rows from dataset randomly with replacement

id	id	id	id
2	2	4	3
0	1	1	3
2	3	3	2
4	1	0	5
5	4	0	1
5	4	2	2

This process is called BOOTSTRAPPING

→ Each bootstrapped dataset must have same no of rows to original

→ Then we randomly select features for each bootstrapped dataset.



→ $x_0 x_1$ $x_2 x_3$ $x_2 x_4$ $x_4 x_3$

→ Build trees for all subsets.

Q

• Why Forest?

→ More than 1 trees

• Why Random?

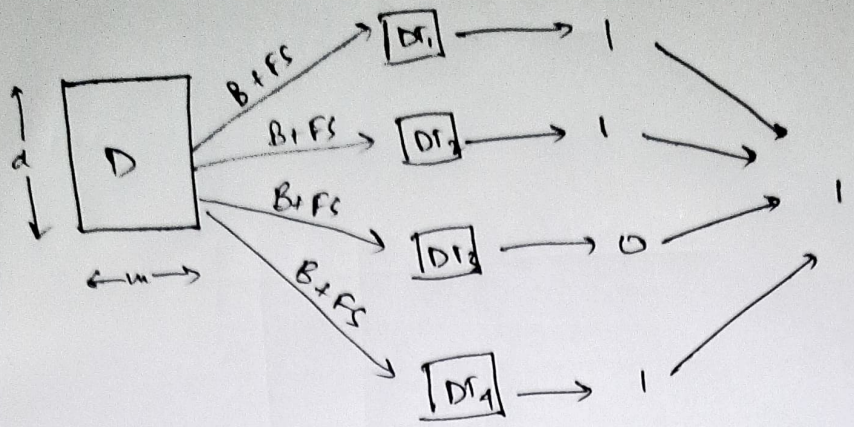
→ Bootstrapping & Random Feature Selection.

• Why Bootstrap & Feature Select?

→ Bootstrap ensures that we are not using same data for every tree. So our model is less sensitive to original training data. Random Feature selection helps to reduce the correlation between trees. If not used, all trees would have similar data and would produce very similar trees and that would have increased variance. Some trees will be giving bad predictions. Those kind will also give bad predictions in the opposite way, thus balancing them out.

• What's the ideal size of feature subset?

→ sqrt of total
log of total
} These work well



Bootstrapping + Aggregating = Bagging.

Classification;-

- Take the majority from each tree

Regression

- Mean of each value
- Mode of values
- Weighted average.

Note! → High variance of each DT averages out to be low variance as because:-

(i) Each tree recognizes few features

(ii) We take majority or mean

→ Change of training data will impact Random Forest much less, unlike DT.