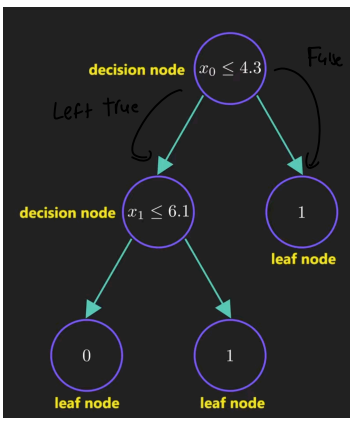
Random Forest Algorithm

The dataset

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

Binary Classifier $y \in (0, 1)$

Sample decision Tree from Static Dataset



If we change some values the tree won't work

(Highly sensitive to training data/high variance)

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	6.5	4.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

To build a Random Forest we build new datasets from our original data.

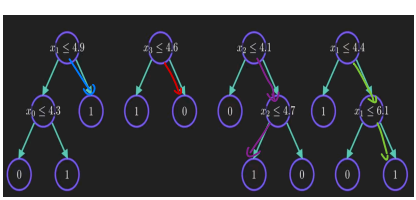
- Randomly select rows from our dataset, with replacement, and save # of rows as original # of rows (order matters)

# of rows

Bootstrapping

- We will train the trees using a subset of randomly selected features for each tree

$x_0\ x_1$ $x_2\ x_3$ $x_2\ x_4$ $x_1\ x_3$



1 0 1 1

* To make a prediction we pass a row into all the trees and not the results

Aggregation

Ex:

2.8	6.2	4.3	5.3	5.5
x_0	x_1	x_2	x_3	x_4

- Combine all the predictions by a majority winner gives 1

* Bootstrapping + Aggregation = Bagging