

## Build a decision tree

### Results:

Single instance error rate is usually around 28%. After 100 instances I get an error rate of about 28.47%.

### How I build my tree and tuned it:

#### Building training and test sets:

I randomly selected 100 rows from the whole data set such that I would have 70 no results and 30 yes results, maintaining the ratio of the full data set. This made my results more consistent.

#### Building the tree:

I followed the ID3 algorithm of stopping if all cases are + or −, or if all attributes were traversed then selecting the most prevalent value.

#### Tuning the tree:

Following the algorithm got me pretty close to a 30% error rate which is as good as always guessing “no.” I added a condition to stop in the case gain isn’t greater than 0.04, which bumped my error rate down to about 29%. I also eliminated a handful of attributes and decreased my error rate down to about 28.5%. The one continuous attribute I kept, I threw the data in excel and picked what I thought to be clusters as breakpoints. Amount broke down into <2000, 2000<=4000, and >4000.

#### Eliminated attributes:

- months-loan-duration
- percent-of-income
- age
- housing
- job
- dependents