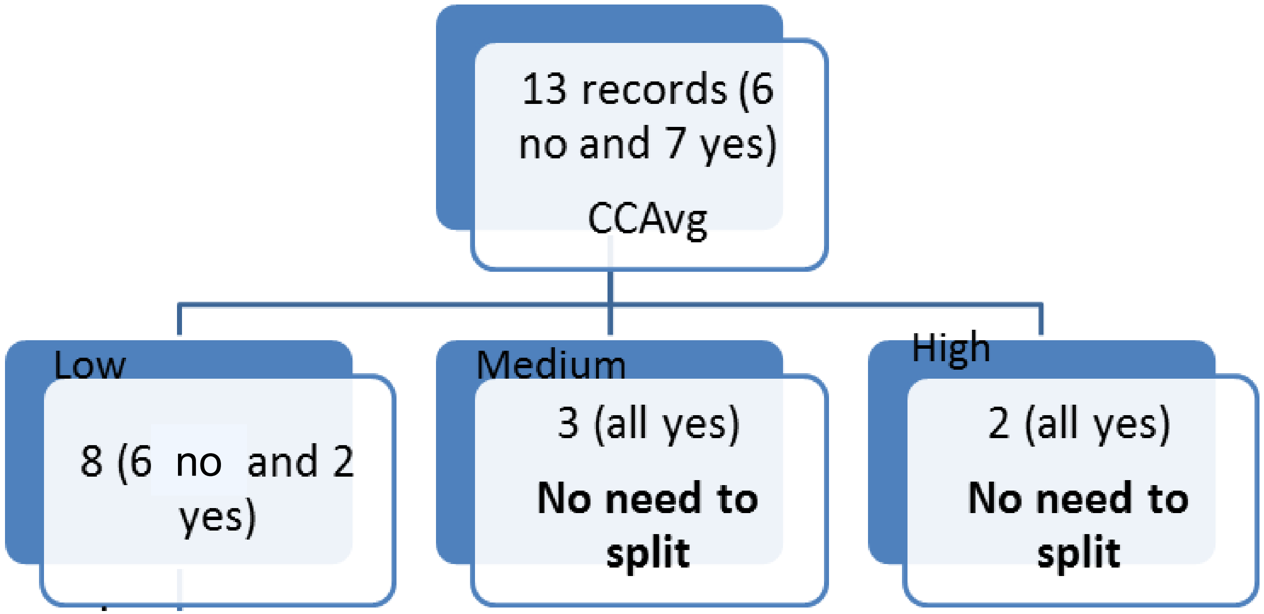
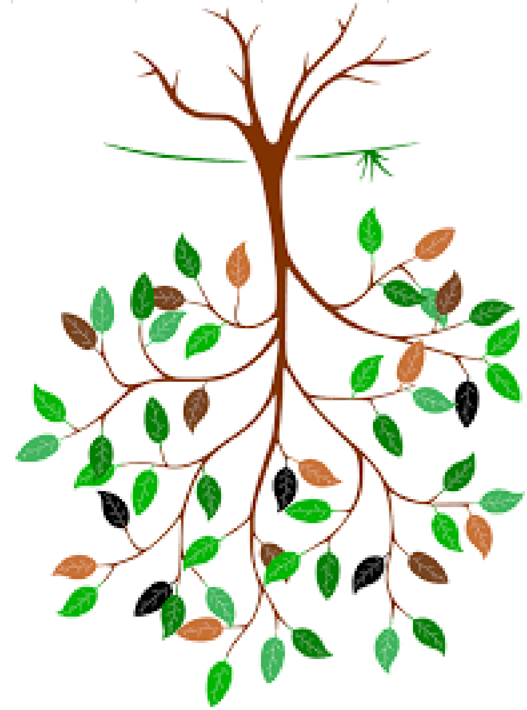
Data

ID	Age	Income	Family	CCAvg	Personal Loan
1	Young	Low	4	Low	0
2	Old	Low	3	Low	0
3	Middle	Low	1	Low	0
4	Middle	Medium	1	Low	0
5	Middle	Low	4	Low	0
6	Middle	Low	4	Low	0
10	Middle	High	1	High	1
17	Middle	Medium	4	Medium	1
19	Old	High	2	High	1
30	Middle	Medium	1	Medium	1
39	Old	Medium	3	Medium	1
43	Young	Medium	4	Low	1
48	Middle	High	4	Low	1

Constructing a Tree



ID	Age	Income	Family	CCAvg	Personal Loan
1	Young	Low	4	Low	0
2	Old	Low	3	Low	0
3	Middle	Low	1	Low	0
4	Middle	Medium	1	Low	0
5	Middle	Low	4	Low	0
6	Middle	Low	4	Low	0
10	Middle	High	1	High	1
17	Middle	Medium	4	Medium	1
19	Old	High	2	High	1
30	Middle	Medium	1	Medium	1
39	Old	Medium	3	Medium	1
43	Young	Medium	4	Low	1
48	Middle	High	4	Low	1



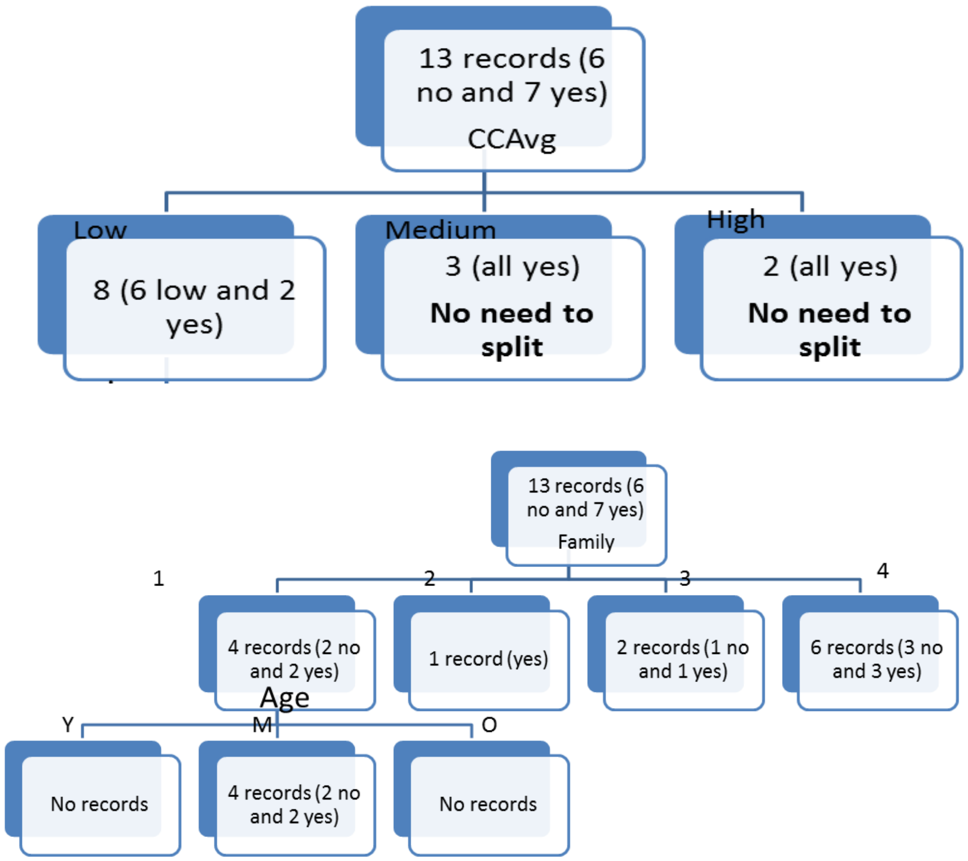
Nodes (root node): Test/Decision points
Leaves: Final Decisions / Conclusion
Branch: Collection of nodes and the leaf

Decision Trees with Different Attributes

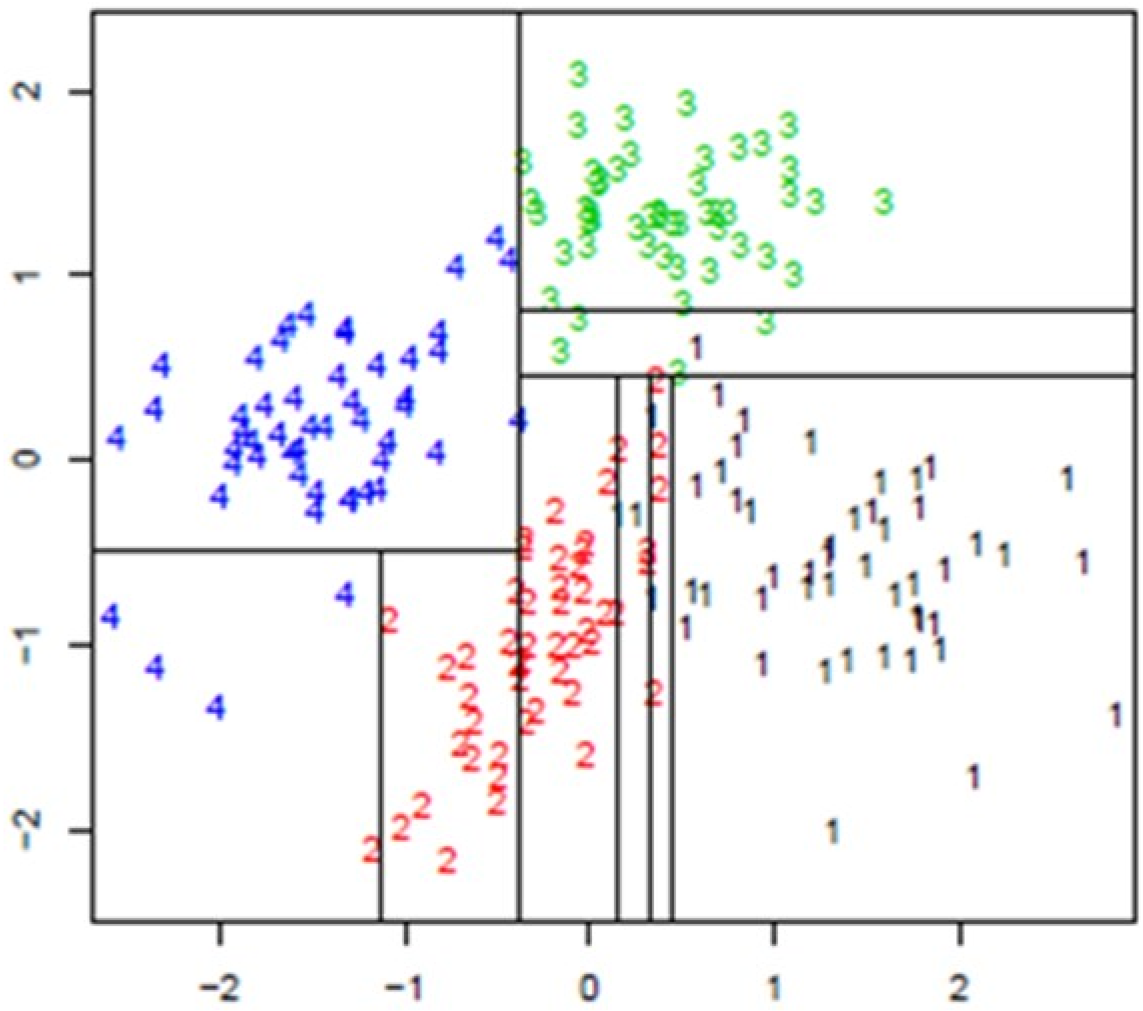
- Decision points (Divide-and-Conquer)
  - Deciding where to start (Selection of the root node)
  - Deciding when to stop (To avoid overfitting)

Trees are Rules Expressed as Disjunctive Normal Form

- *If* (ccAvg is Medium) *or* (CCAvg is High) *then* (Loan = Yes)
  - Within branch nodes are connected with “and” and branches with similar outcome are connected with “or”
- Disjunctive Normal Form
  - Disjunction *(or)* of conjunction *(and)* clauses



Geometry of Decision Trees: Axis Parallel Search



Two Aspects

- Which attribute to choose?

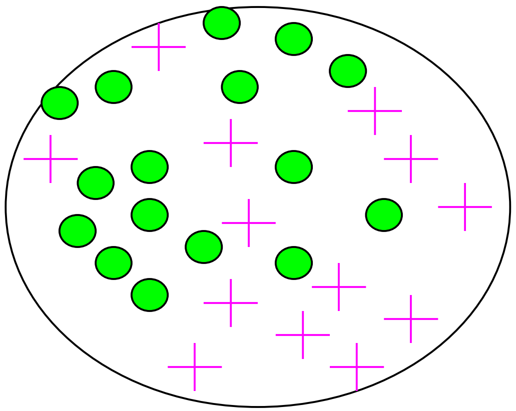
- Where to stop?

Attribute Selection Criteria

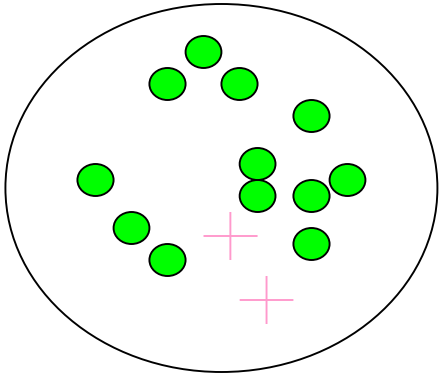
- Main principle
  - Select attribute which partitions the learning set into subsets as "pure" as possible
- Various measures of purity
  - Information-theoretic
  - Gini index
  - ...
- Various improvements
  - probability estimates
  - normalization
  - binarization, subsetting

Impurity

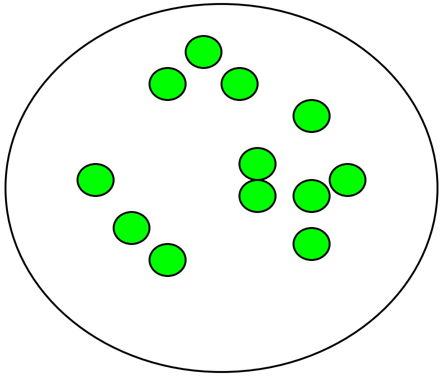
High impurity



Less impurity



Minimum impurity

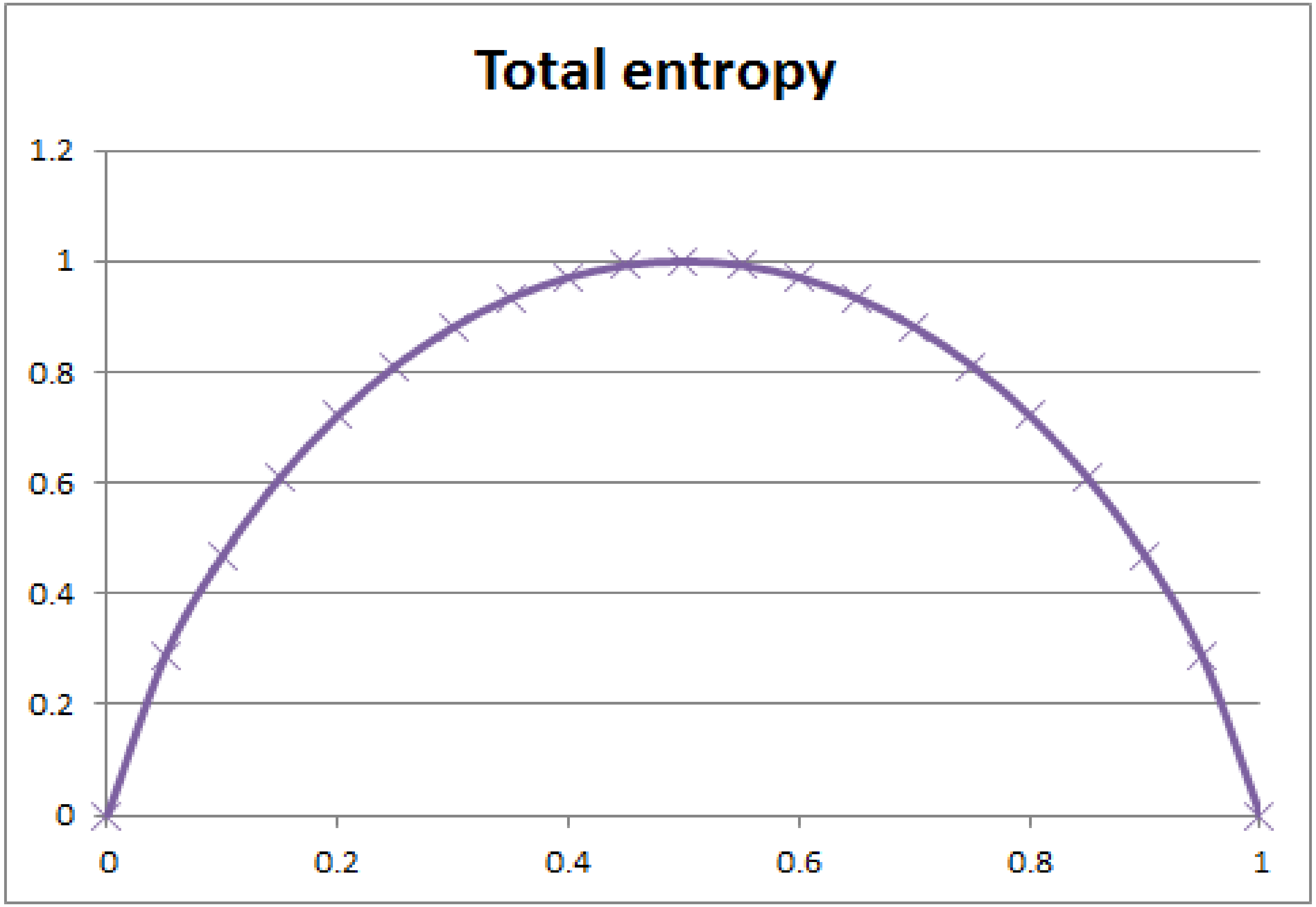


Classification Trees

- Entropy of information is a measure of the randomness or uncertainty or impurity of the outcome.

- Entropy
  - Let us say, I am considering an action like a coin toss. Say, I have five coins with **probabilities for heads** 0, 0.25, 0.5, 0.75 and 1. When I toss them, which one has highest uncertainty and which one has the least?

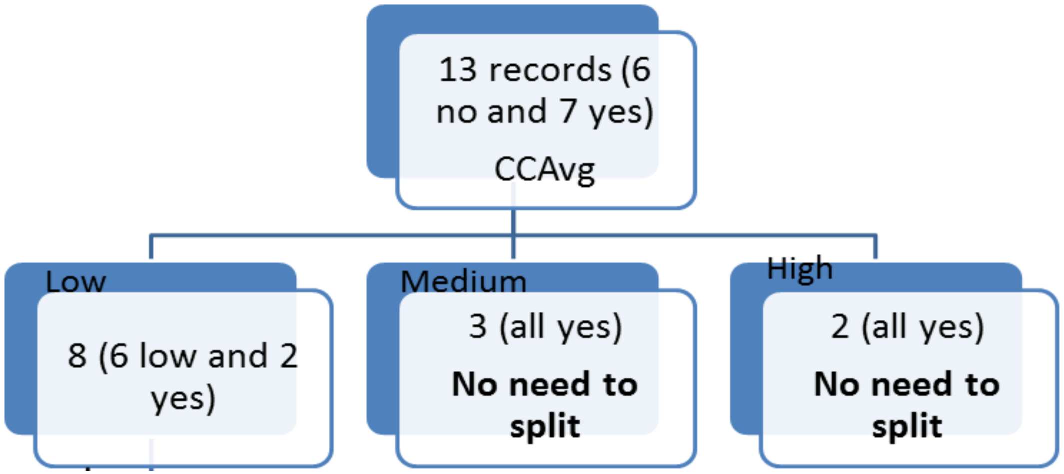
Entropy: A measure of randomness



Entropy and Information Gain (C5.0)

- $\mathrm{H} = -\sum_i p_i \log_2 p_i$

- Information gain = Entropy of the system before split – Entropy of the system after split



(Recall a posteriori or Frequentist approach to calculating probabilities)

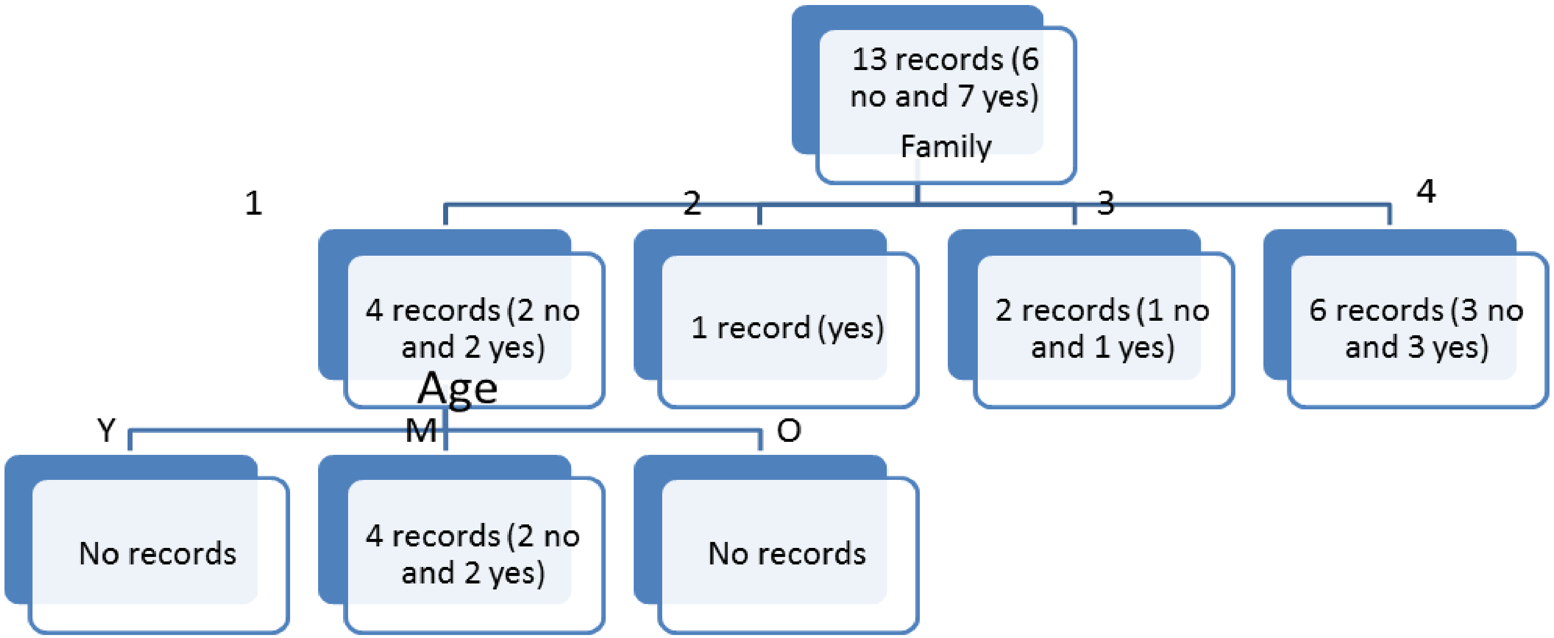
Entropy before split in our example

$$H = -\frac{6}{13} * log_2 \frac{6}{13} - \frac{7}{13} * log_2 \frac{7}{13} = 0.9957$$

Entropy (Weighted) after split on CCAvg

$$H = \frac{8}{13}\left(-\frac{6}{8} * log_2 \frac{6}{8} - \frac{2}{8} * log_2 \frac{2}{8}\right) + \frac{3}{13}\left(-\frac{3}{3} * log_2 \frac{3}{3}\right) + \frac{2}{13}\left(-\frac{2}{2} * log_2 \frac{2}{2}\right)$$
$$= 0.4992$$

Information Gain = 0.9957-0.4992 = 0.4965



Similar calculation for information gain when splitting on Family gives

Information Gain = 0.0726

Sometimes Information Gain Fails

ID	Age	Income	Family	CCAvg	Personal Loan
1	Young	Low	4	Low	0
2	Old	Low	3	Low	0
3	Middle	Low	1	Low	0
4	Middle	Medium	1	Low	0
5	Middle	Low	4	Low	0
6	Middle	Low	4	Low	0
10	Middle	High	1	High	1
17	Middle	Medium	4	Medium	1
19	Old	High	2	High	1
30	Middle	Medium	1	Medium	1
39	Old	Medium	3	Medium	1
43	Young	Medium	4	Low	1
48	Middle	High	4	Low	1

Let us do information gain for split on ID

Entropy after split

Now, the system will have 13 splits, one for each ID.

Entropy = -1*LOG(1,2) = 0

Is ID the root attribute?

- An attribute with many more states is likely to have less variation in each state. So, it will always give better information gain.

- So, we need to normalize it to get something like information gain per state.

Information Content

- Information content is defined as $= -\sum f_i log f_i$. We only want to know fraction of the members in a state divided by the total members.
- **Information content of ID**: It has 13 states. So, the information content

  = - 1/13*LOG(1/13,2)*13 = 3.7
- Information content of ccAvg = 1.33

Gain Ratio

- Information Gain is biased towards attributes with many values (levels)
- Gain Ratio normalizes Information Gain by dividing by the Information Content at the attribute

$$GainRatio(A) = \frac{Gain(A)}{InformationContent(A)}$$

- Attribute with the maximum Gain Ratio is selected as the splitting attribute

Gain Ratio

- Gain Ratio for ID = 0.27
- Gain Ratio for ccAvg = 0.37

Gini Index – Used in CART

$$1 - \sum_{1}^{m} p_i^2$$

It is computed on binary splits only.

So, if we take ccAvg (low, medium and high), it considers all binary options {Low}, {medium, high} or {medium}, {low, high}, etc.

Is a low or a high Gini preferred?

Gini Index

$$1 - \sum_{1}^{m} p_i^2$$

ID	Age	Income	Family	CCAvg	Personal Loan
1	Young	Low	4	Low	0
2	Old	Low	3	Low	0
3	Middle	Low	1	Low	0
4	Middle	Medium	1	Low	0
5	Middle	Low	4	Low	0
6	Middle	Low	4	Low	0
10	Middle	High	1	High	1
17	Middle	Medium	4	Medium	1
19	Old	High	2	High	1
30	Middle	Medium	1	Medium	1
39	Old	Medium	3	Medium	1
43	Young	Medium	4	Low	1
48	Middle	High	4	Low	1

$$\text{Gini Index before split} = 1 - \left(\frac{6}{13}\right)^2 - \left(\frac{7}{13}\right)^2 = 0.497$$

Gini after split with {Low} and {Medium,High}

$$= \frac{8}{13}\left(1 - \left(\frac{6}{8}\right)^2 - \left(\frac{2}{8}\right)^2\right) + \frac{5}{13}\left(1 - \left(\frac{5}{5}\right)^2\right) = 0.231$$

Calculated similarly for other binary splits

The one that gives the least Gini Index is picked