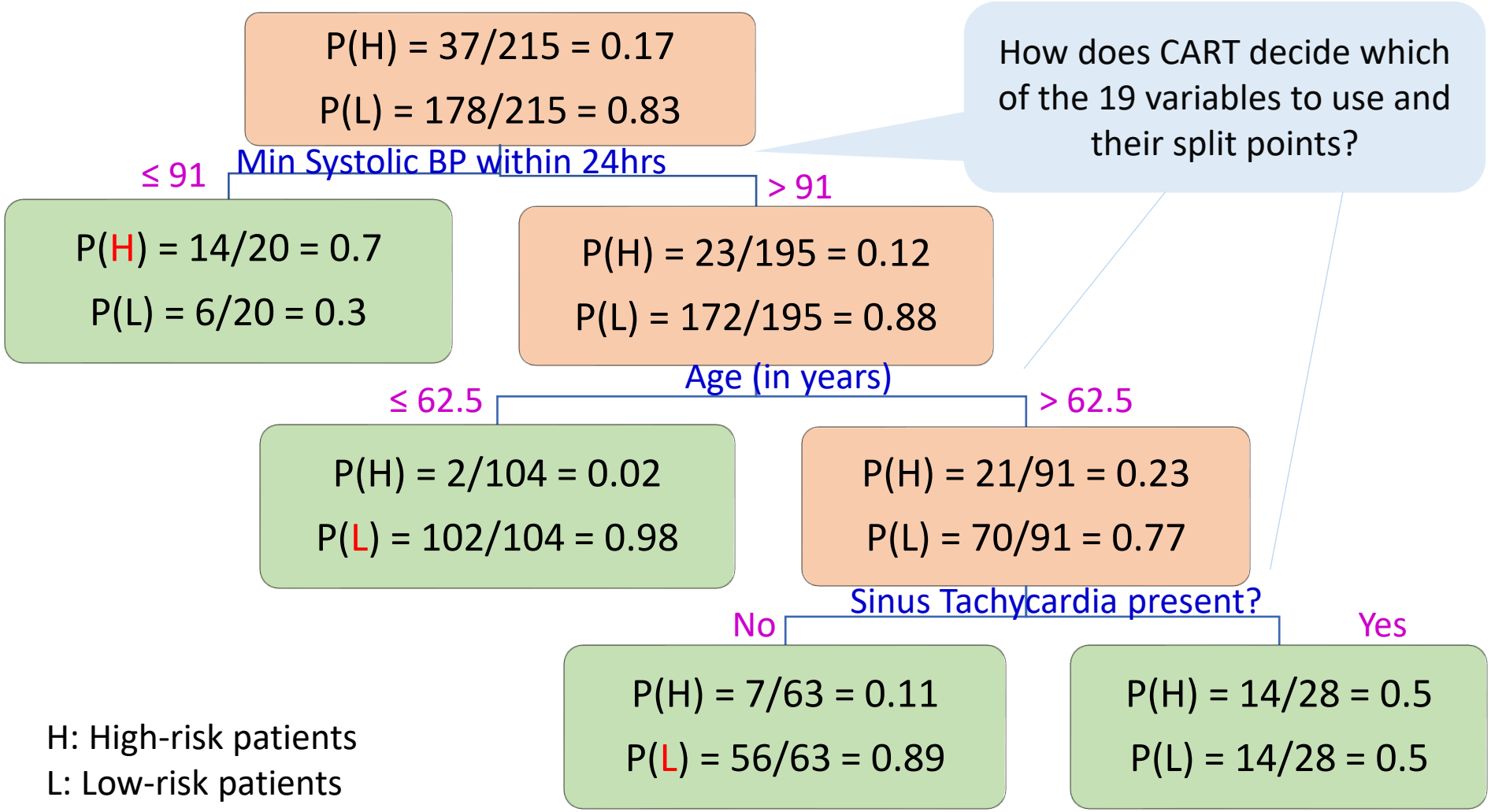
Node Purity, Misclassification Error and Gini Index (for Categorical Y only)

CART

Based on Chew C. H. (2020) textbook: AI, Analytics and Data Science. Vol 1., Chap 8.

CART Results for Heart Attack Prognosis

Source: Breiman, Friedman, Olshen, and Stone (1983). Classification and Regression Trees. Wadsworth.



H: High-risk patients
L: Low-risk patients

- There are 4 terminal nodes (i.e. 4 decision rules).
- Model predictions at terminal nodes are based on majority.
- Misclassification error at a terminal node = minority proportion.

Key Concepts in Phrase 1: Growing Tree to the Maximum

Categorical Y

- Node: $\hat{Y} = majority$
- Misclassification Error
  - Node level
  - Tree level
- Node Purity
- Gini Index
- Entropy

Continuous Y

- Node: $\hat{Y} = \bar{Y}$
- SSE at the node level
- MSE at the Tree level

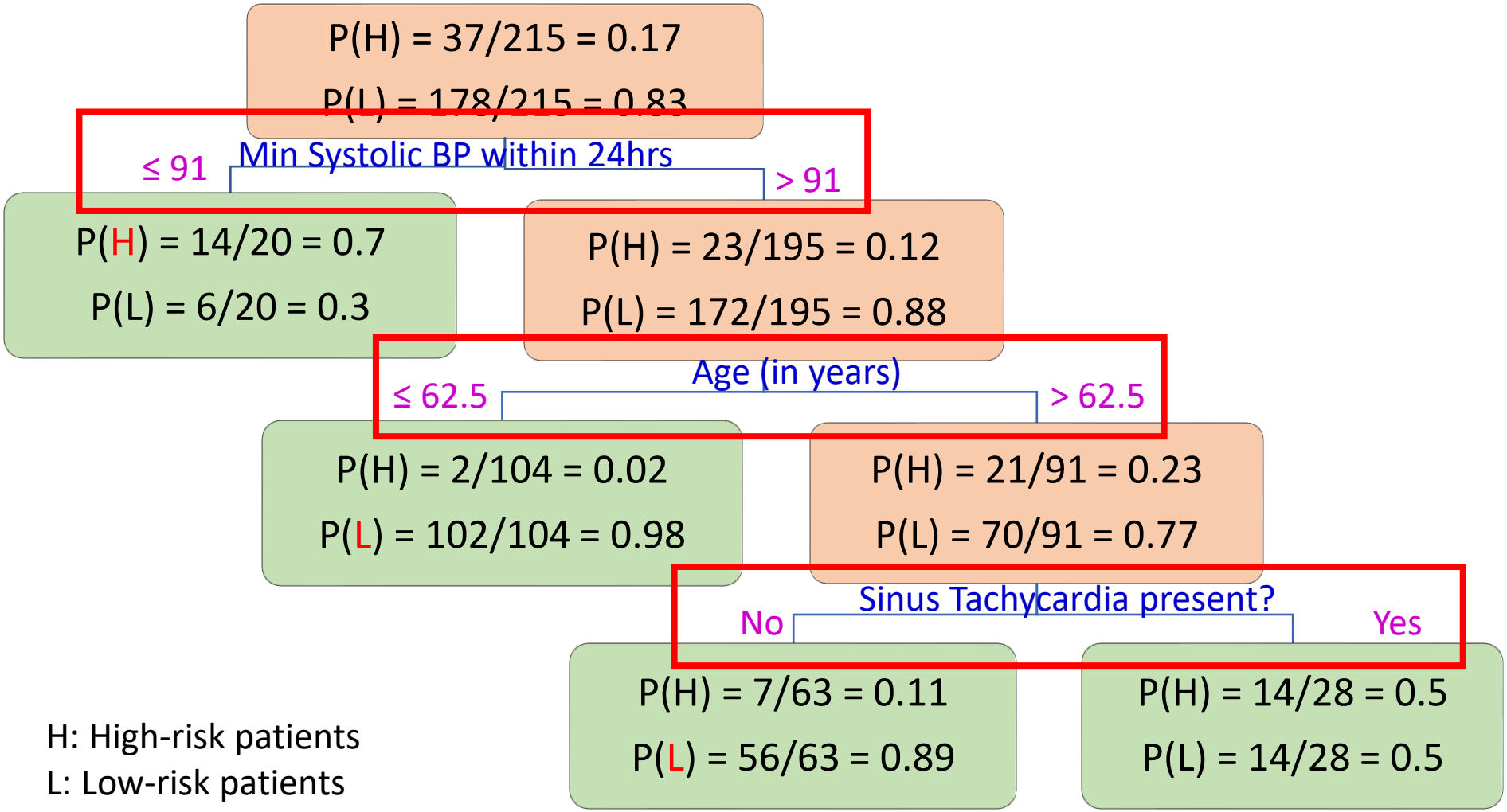
Surrogates

Choosing the Best Splitting Variable and Best Split Point

- At each node during growth phrase, CART must consider and test:
  - all X variables and,
  - all possible values in that X variable
  - to determine the best binary split.
- A split is good if it results in purer child nodes, on average.
- The best split will result in the purest possible child nodes, on average.
- The theoretical purest child node is the node with 100% in one Y category (i.e. 0% in all other Y categories).
- The theoretical most impure child node is the node with the same proportion in each of the Y category (i.e. uniform distribution for Y).
- CART model prediction accuracy directly affected by node purity.

CART Results for Heart Attack Prognosis

Source: Breiman, Friedman, Olshen, and Stone (1983). Classification and Regression Trees. Wadsworth.



H: High-risk patients
L: Low-risk patients

- A split is defined by a specific X variable and a specific value.
- Only the best split is used (and shown) to generate 2 child nodes.
- Result is purer when averaged across the 2 child nodes.

To automate the search for the best split, need to specify the selection criteria in terms of a formula

- Formulas to determine the best binary split:
  - Gini Index, or
  - Entropy.
- Gini index is preferred by Prof Leo Breiman.
  - Default setting in many software including Rpart package.
- Misclassification Error is not a good formula to determine the best binary split.
  - Set as an exercise question.

Example: Y = category A or B.

1. Best case: Highest purity occurs when 0% cat A and 100% cat B, or 100% cat A and 0% cat B.
2. Worst case: Highest impurity occurs when 50% cat A and 50% cat B.
3. Preference: Impurity of {40% cat A, 60% cat B > Impurity of {30% cat A, 70% cat B}
4. Symmetry: Impurity of {40% cat A, 60% cat B} = Impurity of {60% cat A, 40% cat B}

- All formulas used to determine the best split must demonstrate the above results.

Entropy, Gini and Misclassification Error

Let $p(y \mid t)$ denote the proportion of cases belonging to class y at tree node t.

The three measures can be simplified if y has only two possible outcomes {0, 1}.

Let p represent p(y = 1 | t), then 1 – p represent p(y = 0 | t). The 3 impurity measures become:

Entropy(t) =

If Y has k categories

If Y has only 2 categories

$$-\sum_{y=0}^{k} p(y \mid t)\log_2 p(y \mid t) = -\left[p(0 \mid t)\log_2 p(0 \mid t) + p(1 \mid t)\log_2 p(1 \mid t)\right] = -\left[(1-p)\log_2(1-p) + p\log_2 p\right]$$

*Note: 0*log(0) is defined as 0.*

If Y has k categories

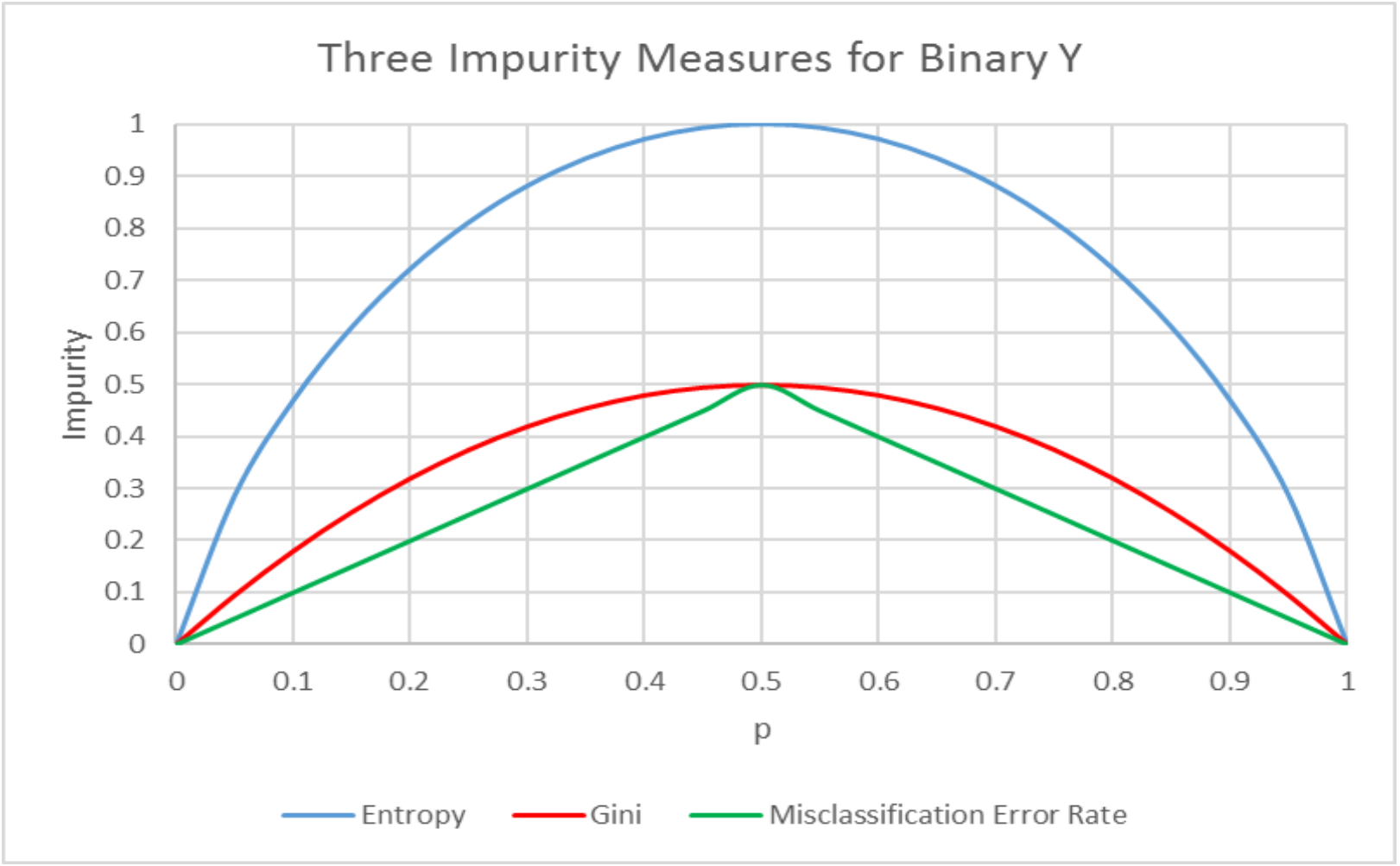
If Y has only 2 categories

$$\textbf{Gini(t)} = 1 - \sum_{y=0}^{k}[p(y \mid t)]^2 = 1 - \left[p(0 \mid t)^2 + p(1 \mid t)^2\right] = 1 - \left[(1-p)^2 + p^2\right] = 2p(1-p)$$

[Majority Rule] Misclassification error, r(t) = $1 - \max\{p(y \mid t)\} = 1 - \max\{p, 1-p\}$

If Y has only 2 categories

If Y has k categories

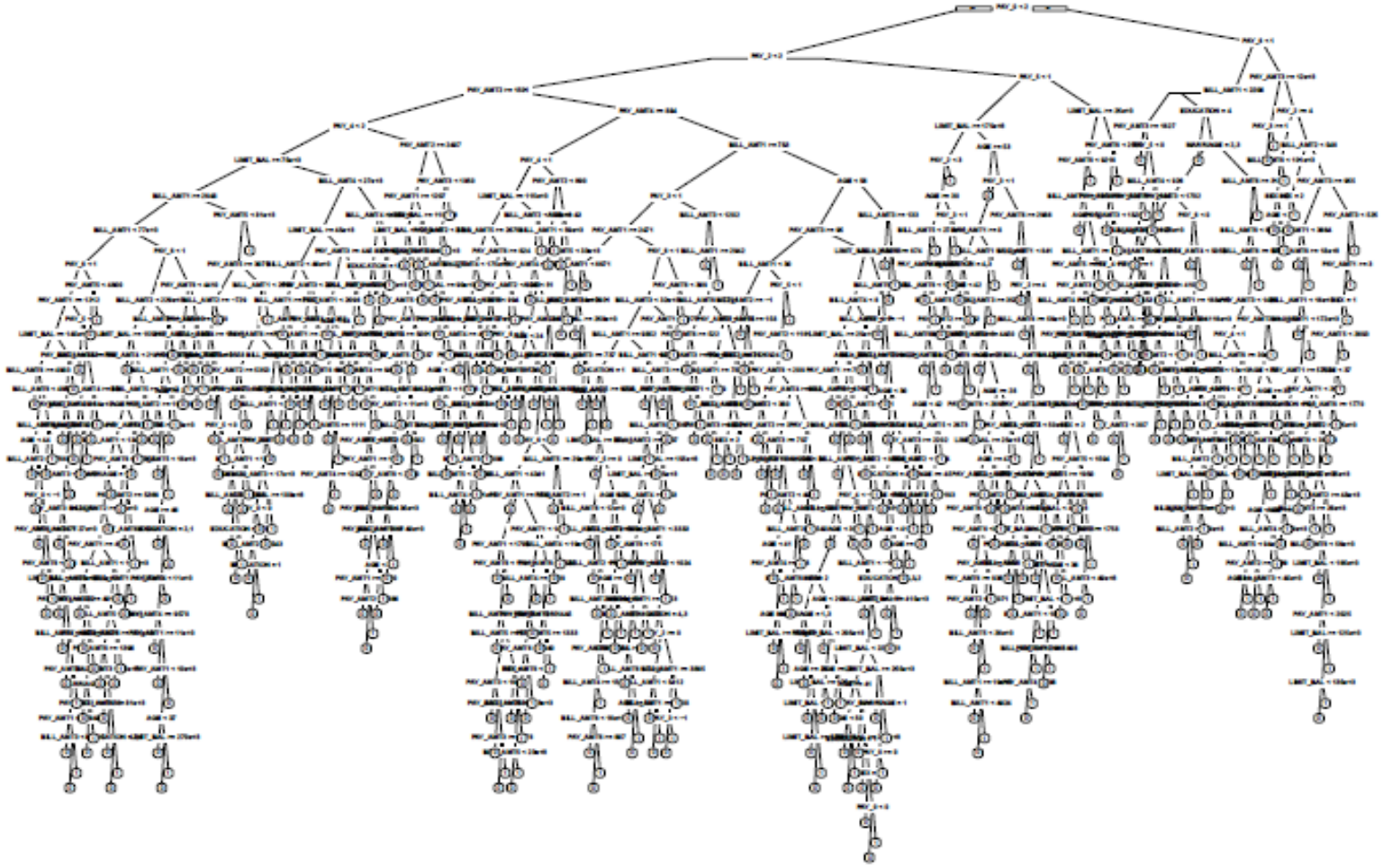


- Gini and Misclassification range from 0 (best) to 0.5 (worst).
- Entropy range from 0 (best) to 1 (worst).
- A numerical exercise with a small dataset will clarify how to use these formulas to determine the best split.

Growing the Tree to the maximum

- At each step, the best splitting variable and it's best split point is found, and applied to create two child nodes.
- The process continues, until a (lenient) stopping criteria is met.
- The result is a very large tree, which is likely to suffer from overfitting.
- Pruning is then required to prune the Tree to the minimum (root node) in Phrase 2.
  - Pruning generates a sequence of smaller and smaller Trees.
  - The optimal Tree is somewhere between the maximal Tree and the minimal Tree.

A Maximal Tree generated from a dataset with 21,000 cases and 23 X variables to predict Credit Risk



Next: Pruning the Tree

- Understand the Pruning process.
- Weakest Link Pruning.
  - Define the strength of a node's link to all its descendants.
- Completely determined by the data
  - No human judgement.
  - Pure Machine Learning.