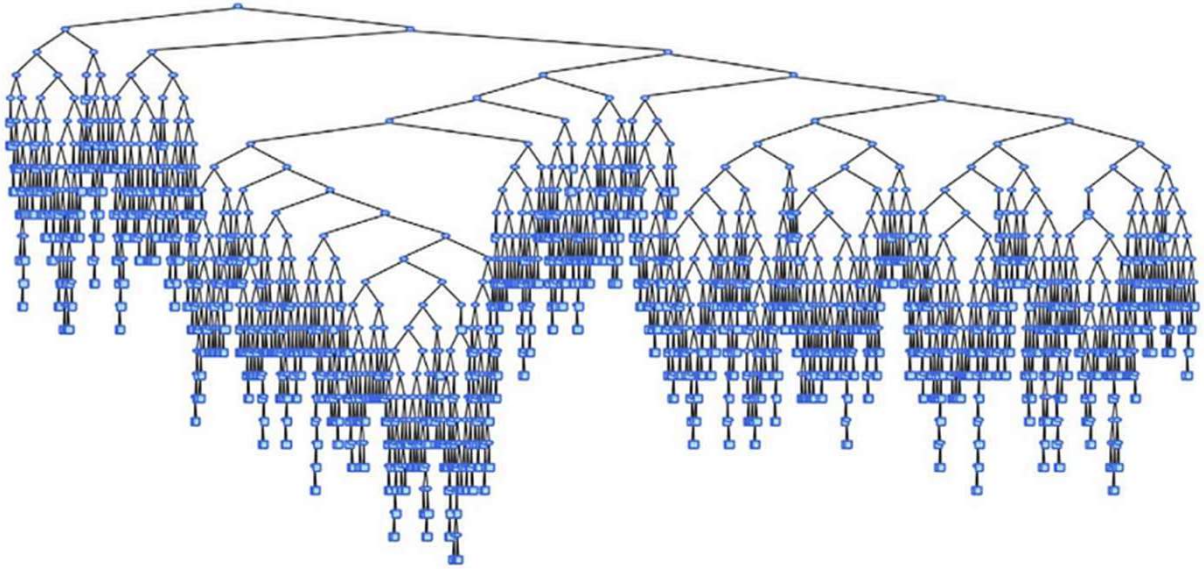
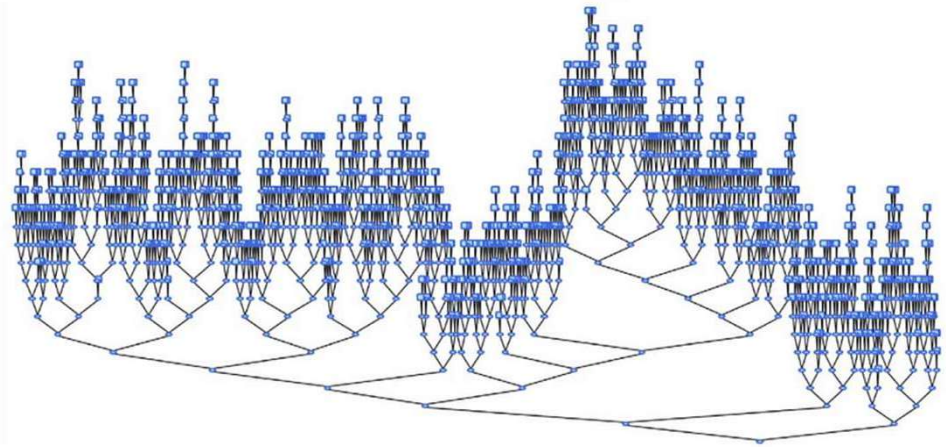
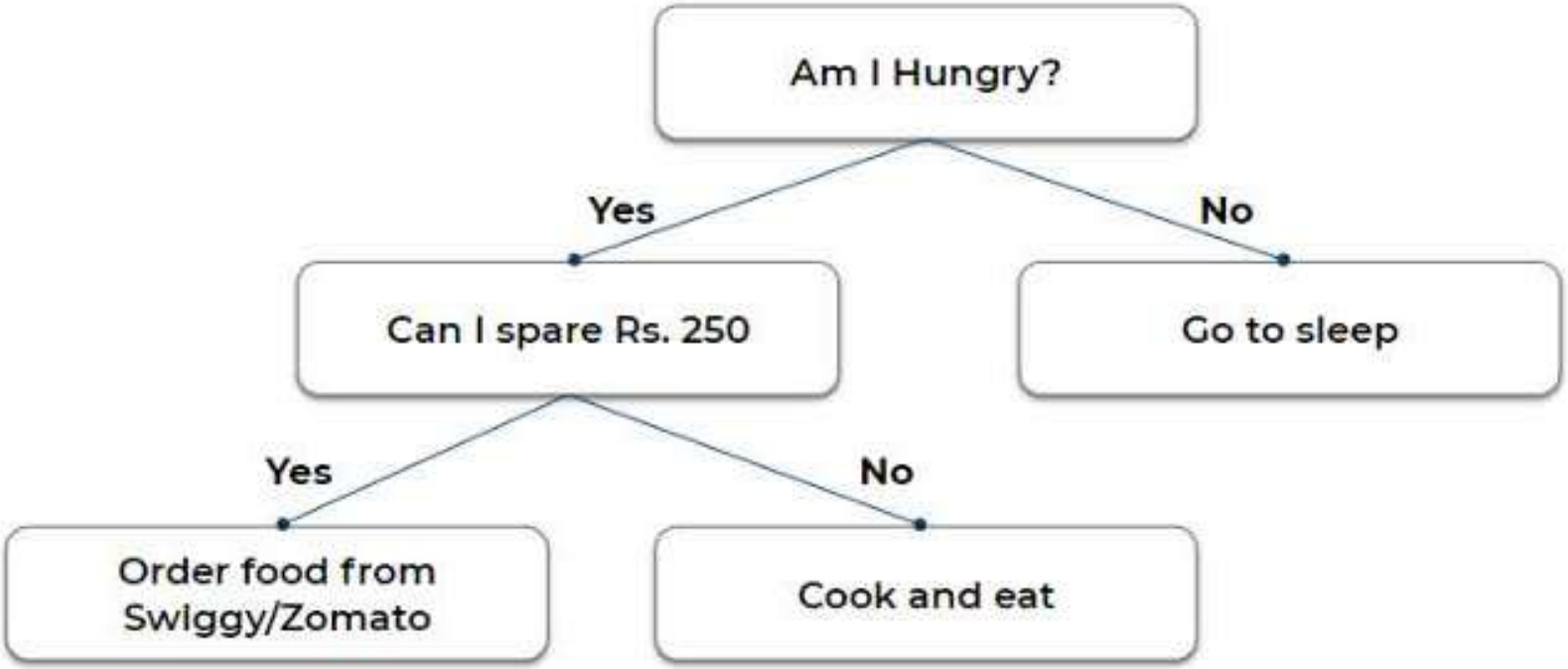
Decision trees

Do you see the tree?

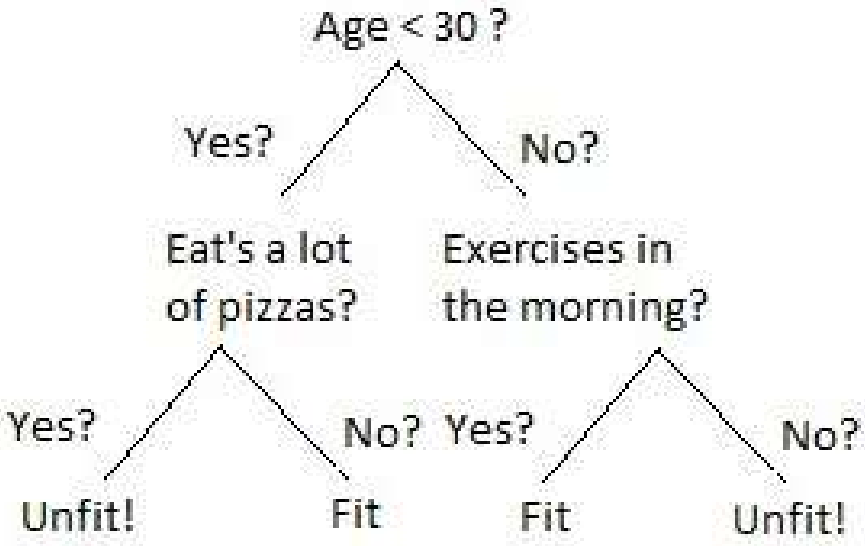


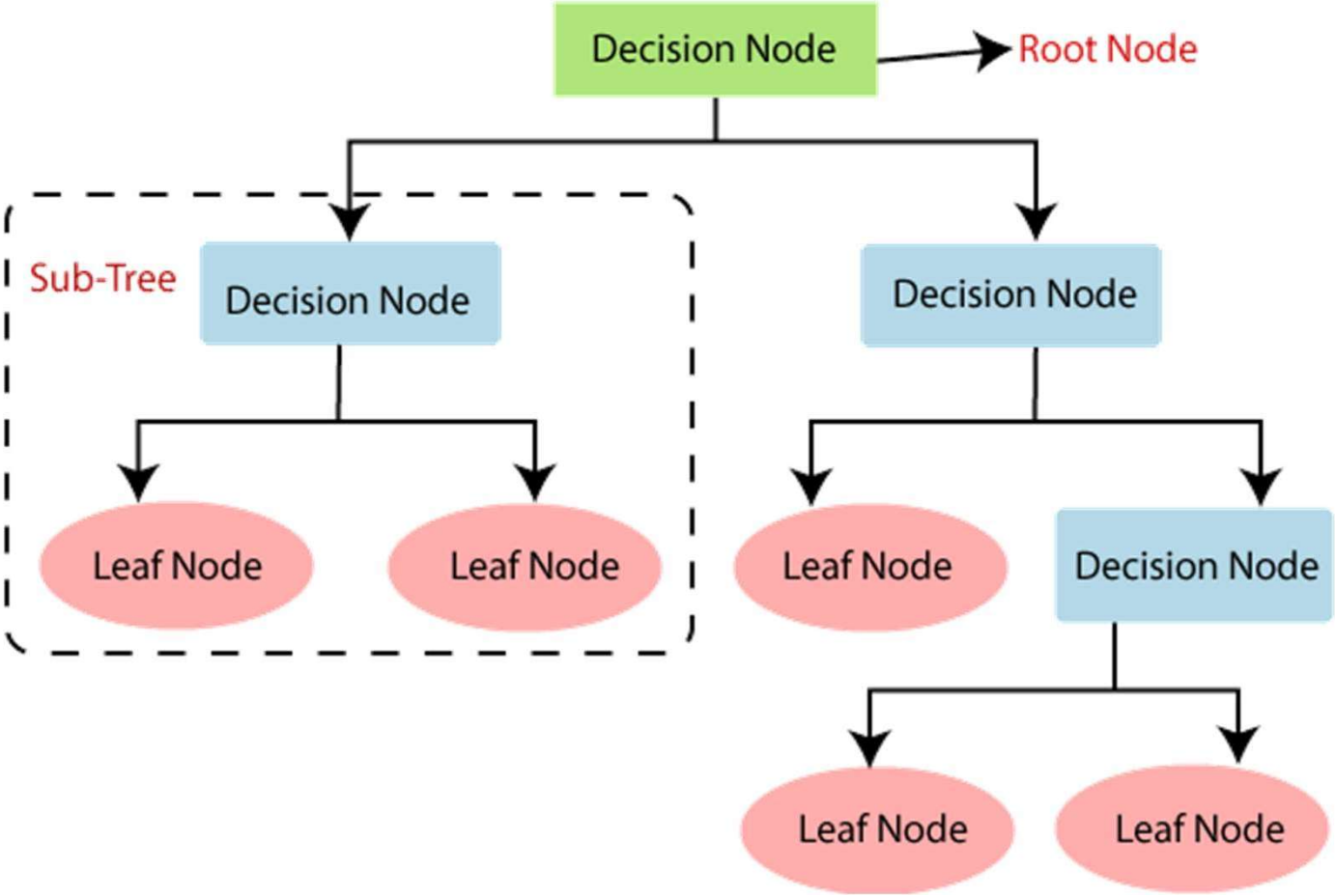
Do you see the tree?




Am I Hungry?
Yes
No
Can I spare Rs. 250
Go to sleep
Yes
No
Order food from Swiggy/Zomato
Cook and eat

Is a Person Fit?

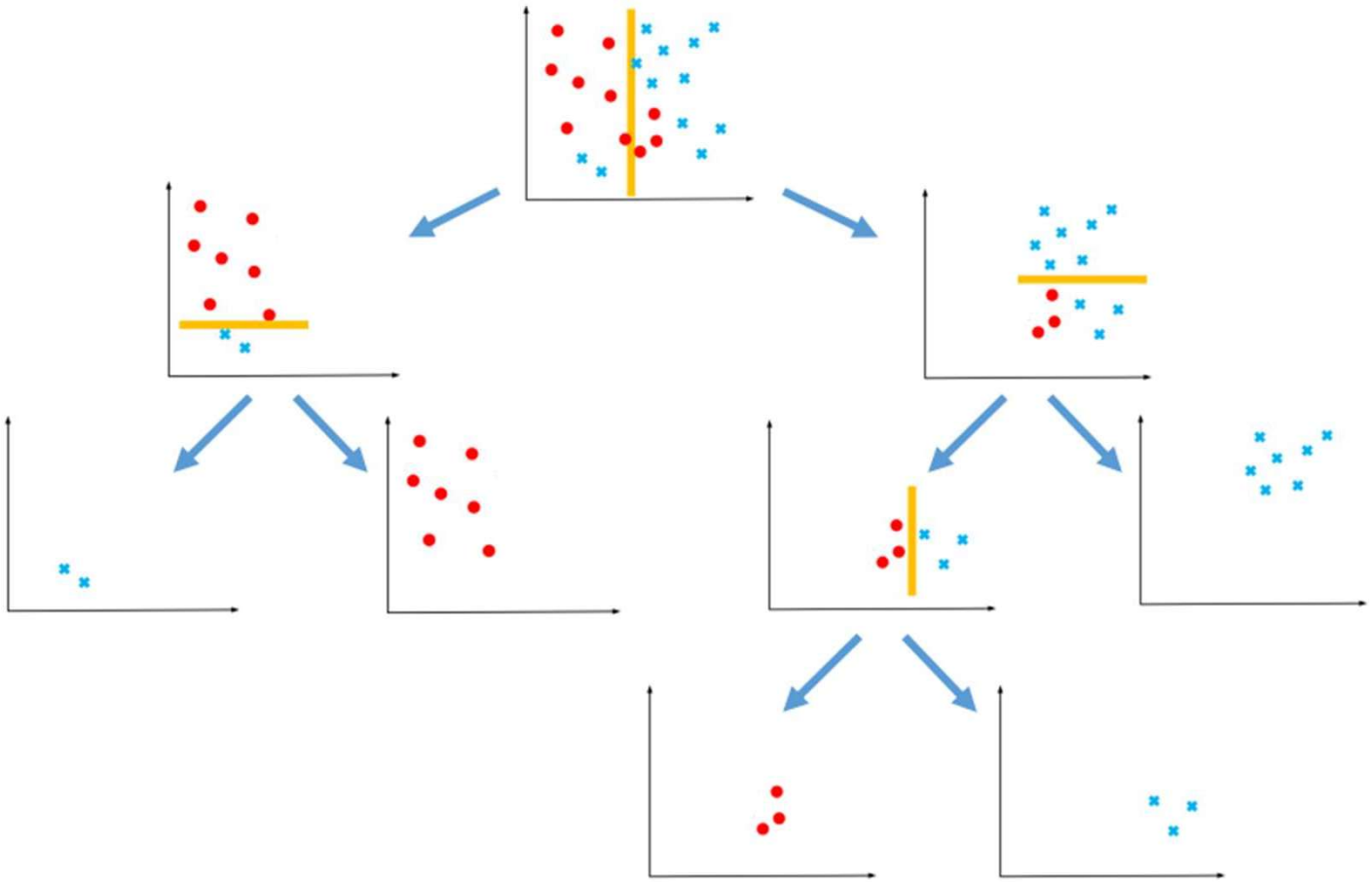



Decision Node
Root Node
Sub-Tree
Decision Node
Decision Node
Leaf Node
Leaf Node
Leaf Node
Decision Node
Leaf Node
Leaf Node

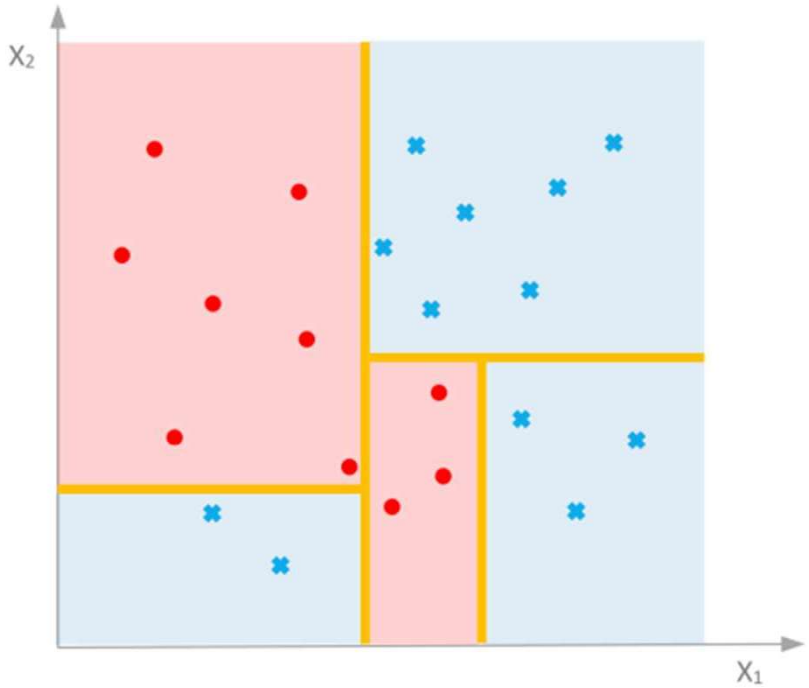
What are decision trees?

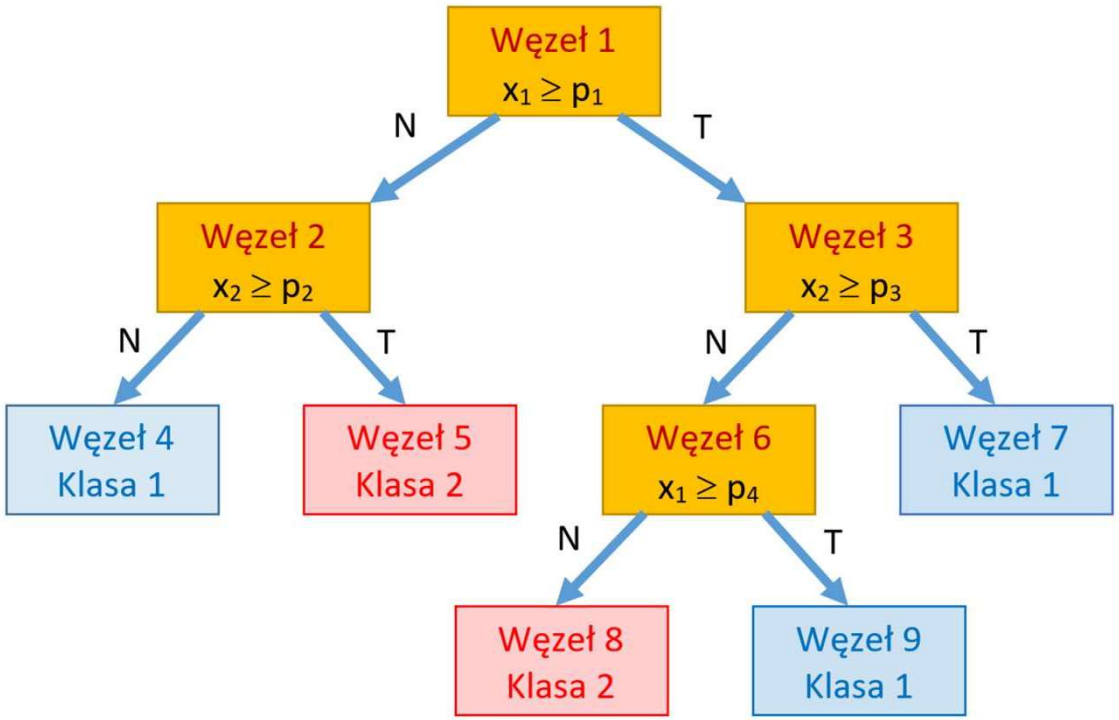
A decision tree is a **supervised learning** technique that has a pre-defined target variable and is most often used in **classification problems**. This tree can be applied to either **categorical** or continuous input & output variables. The training process resembles a flow chart, with each internal (non-leaf) node a test of an attribute, each branch is the outcome of that test, and each leaf (terminal) node contains a class label. The uppermost node in the tree is called the root node.

Decision trees – building the model

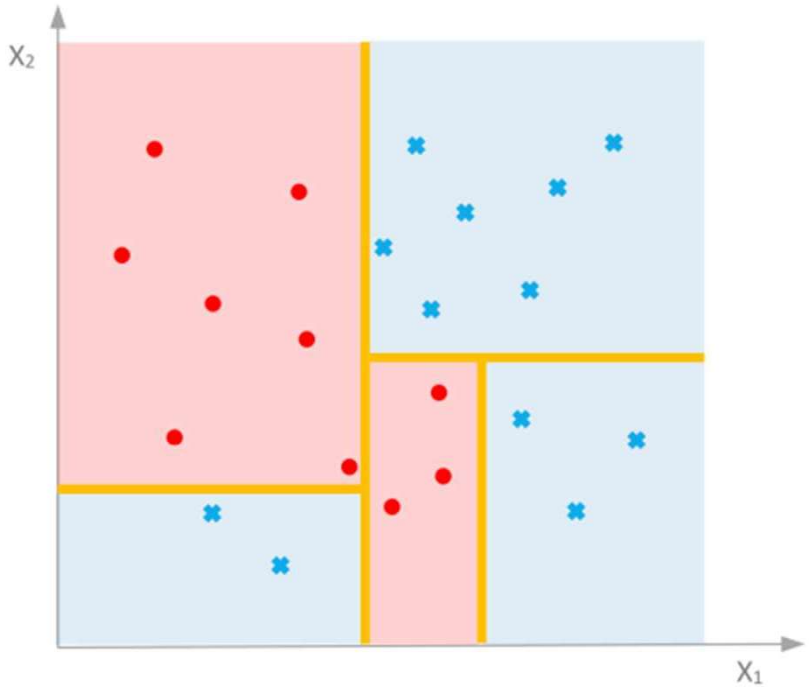


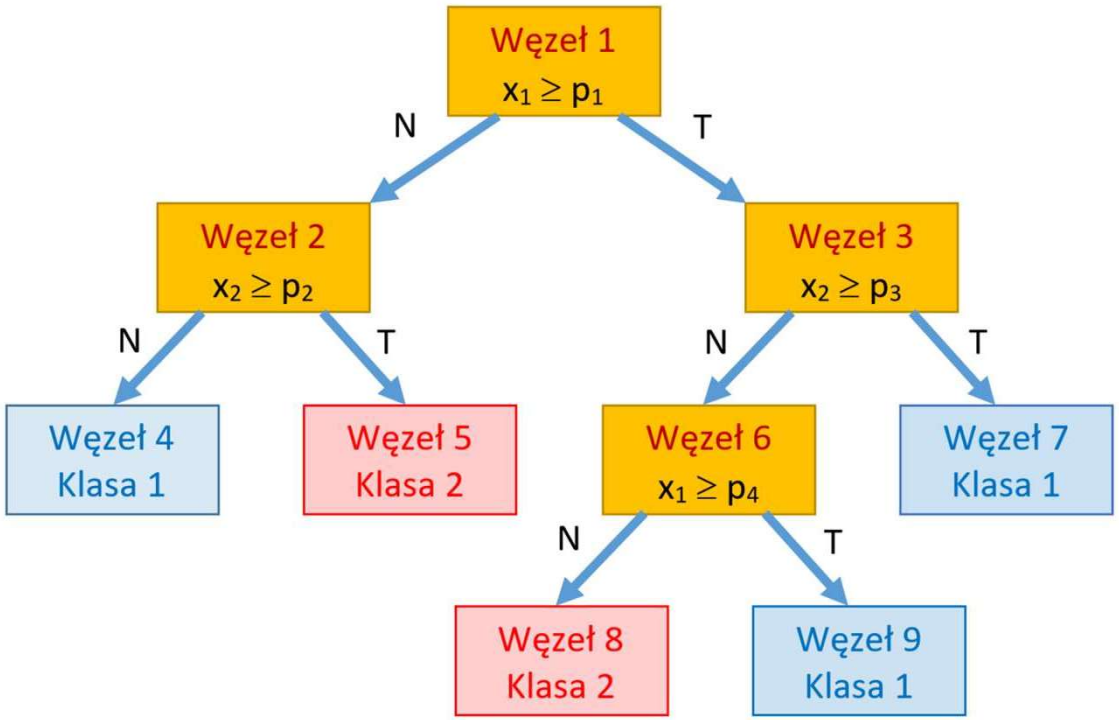
Decision trees – building the model



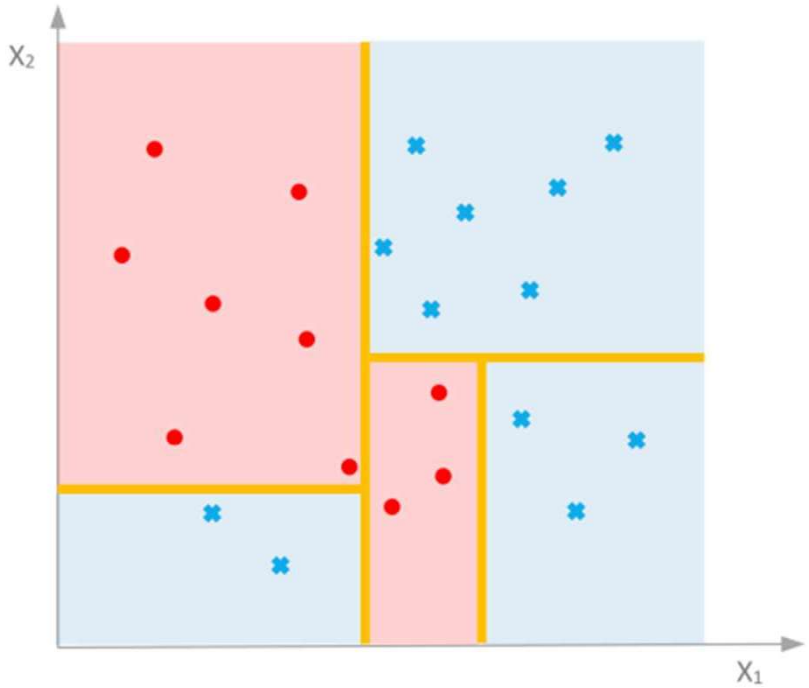


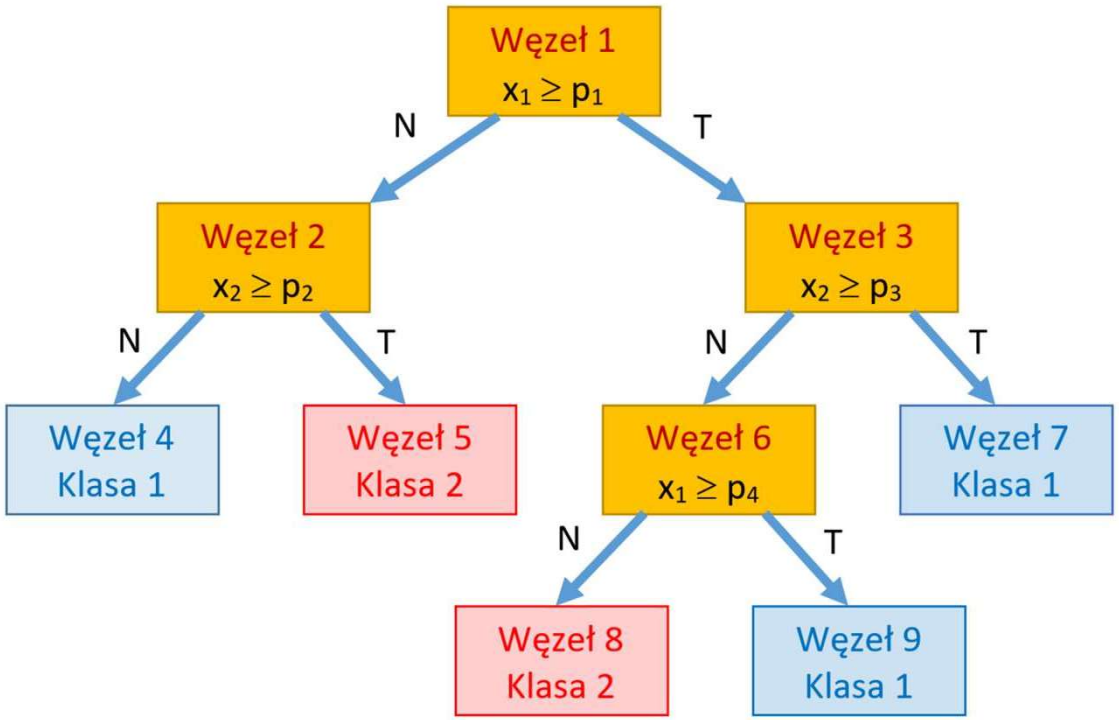
Decision trees – building the model





Decision trees – building the model





Advantages and disadvantages of Decision Trees

Advantages	Disadvantages
Easy to understand, interpret and visualize.	Decision Trees can be unstable and change greatly with slight changes in training data.
Applications with categorical values (sunny, cloudy, rainy) and numerical values (wind speed = 10 kmh) can be mapped.	In case of unbalanced training data (e.g. very often sunny weather) this so-called bias can also be present in the tree.
Non-linear relationships between variables do not affect the accuracy of the tree.	Trees can quickly become very complex and overfit the training data. As a result, they do not generalize as well to previously unseen data.
The number of decision-making levels is theoretically unlimited.	High training time
Several decision trees can be combined to form a so-called random forest.	

How to build a tree

- **CART-Classification and Regression Trees**
- **ID3-Iterative Dichotomiser 3**
- C4.5
- CHAID-Chi-squared Automatic Interaction Detection

Main question: which feature should be use first?

Exemplary problem – prediction of heart disease

Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
NO	NO	NO	NO
YES	YES	YES	YES
YES	YES	NO	NO
YES	NO	YES	YES
etc.	etc.	etc.	etc.

$P_{yes} = 0.5$
$P_{no} = 0.5$

$G = 1 - 0.5^2 - 0.5^2 = 0.5$

$P_{yes} = 1 \quad P_{no} = 0 \quad G = 0$

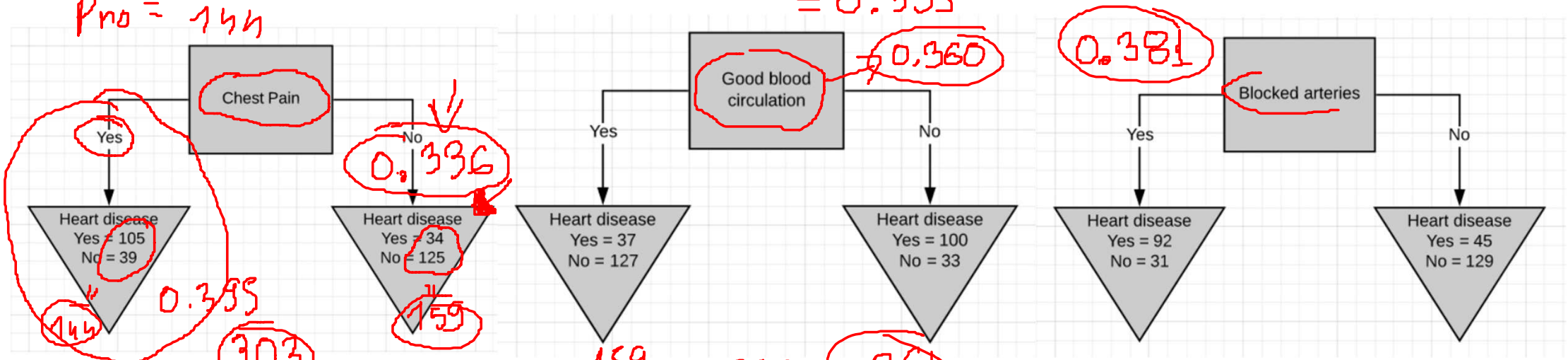
How different features answer to the question

$P_{yes} = \frac{105}{144}$

$P_{no} = \frac{39}{144}$

$G = 1 - P_{yes}^2 - P_{no}^2 = 1 - \left(\frac{105}{144}\right)^2 - \left(\frac{39}{144}\right)^2$

$= 0.395$



144

159

303

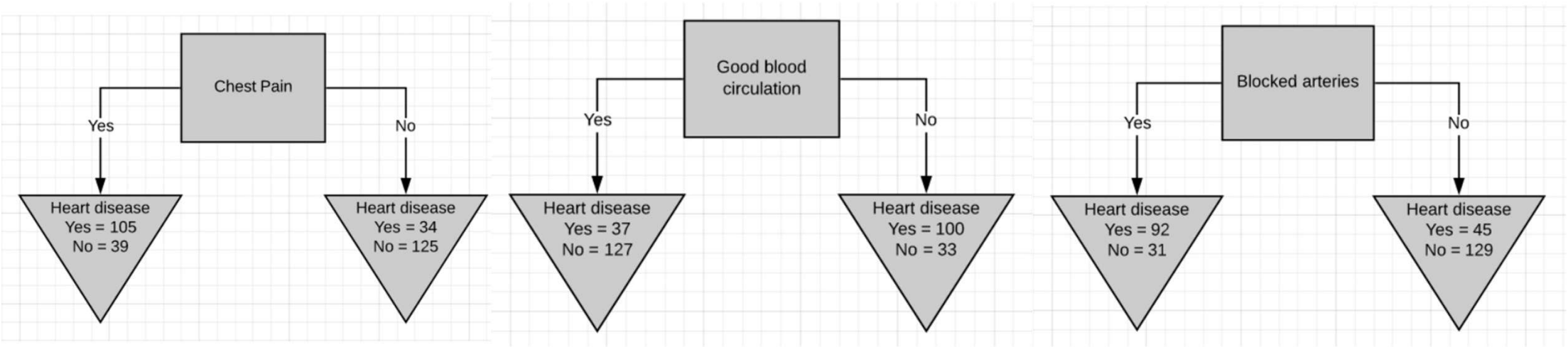
$\frac{144}{303} \cdot 0.395 + \frac{159}{303} \cdot 0.336 = 0.364$

Gini impurity score

$$\mathrm{I}_G(p) = \sum_{i=1}^{J}\left(p_i \sum_{k\neq i} p_k\right) = \sum_{i=1}^{J} p_i(1-p_i) = \sum_{i=1}^{J}(p_i - p_i^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} p_i^2 = 1 - \sum_{i=1}^{J} p_i^2$$

Squared probability for each possible answer – in case of binary trees – probability of first and second answer

How different features answer to the question

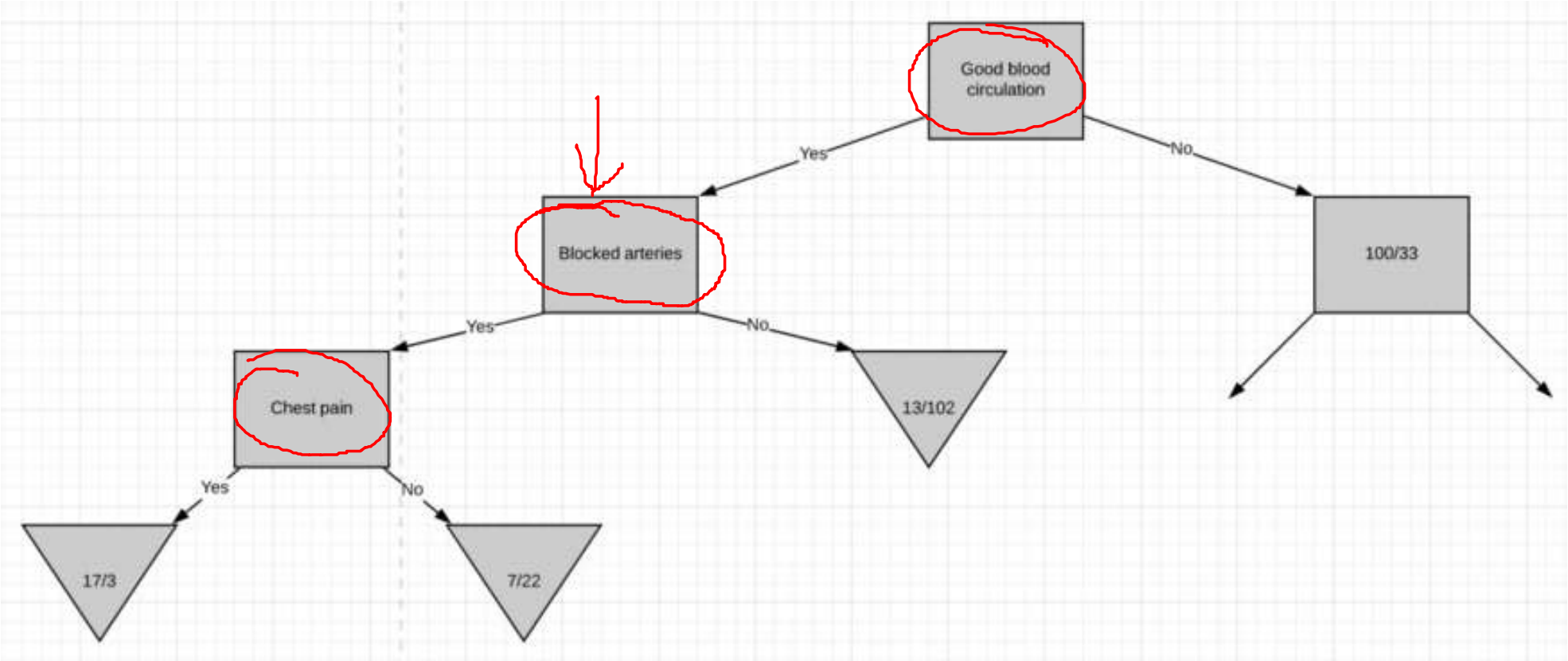


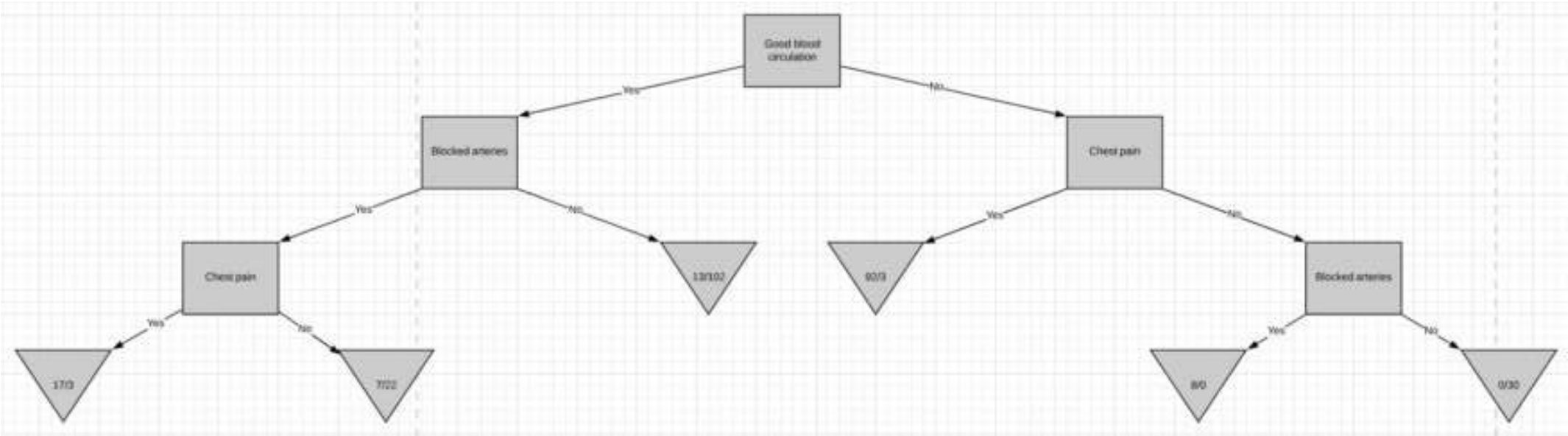
CART-Classification and Regression Trees

- We use Gini impurity score to grade diffent features – this describe how well they answer our main question
- We look for the feature with the least impurity
- Impurity = 0 means that the feature perfectly split the data

CART-Classification and Regression Trees

- We use Gini impurity score to grade diffent features – this describe how well they answer our main question
- Impurity = 0 means that the feature perfectly split the data
- Steps:
  - Calculate the Gini impurity scores.
  - If the node itself has the lowest score, then there is no point in separating the patients anymore and it becomes a leaf node.
  - If separating the data results in improvement then pick the separation with the lowest impurity value.


Good blood circulation
Yes
No
Blocked arteries
100/33
Yes
No
Chest pain
13/102
Yes
No
17/3
7/22


Good blood circulation
Yes
No
Blocked arteries
Chest pain
Yes
No
Yes
No
Chest pain
13/10/2
9/2/3
Blocked arteries
Yes
No
Yes
No
17/3
7/7/2
8/0
0/30

Entropy

In information theory, the **entropy** of a random variable is the average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes.

Entropy gives a number between 0 and 1, where means pure separation and 1 random separation

Entropy describes impurity of subset rather than impurity of the choice, so we use information gain to grade possible splits.

$$H(x) = \sum_{i=1}^{n} p(i) \log \frac{1}{p(i)} = -\sum_{i=1}^{n} p(i) \log p(i)$$

Examples

0 1

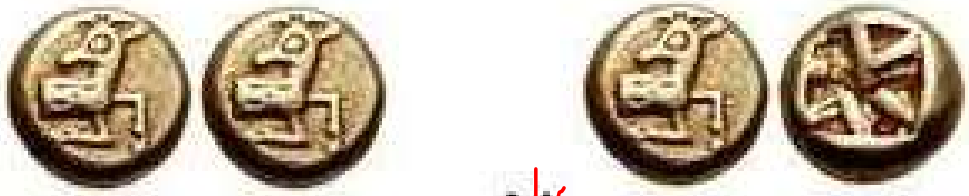
0 - 15 = 16 = 2^4

0000

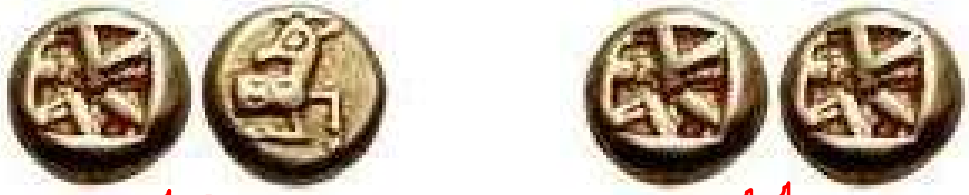
....

1111

00 01

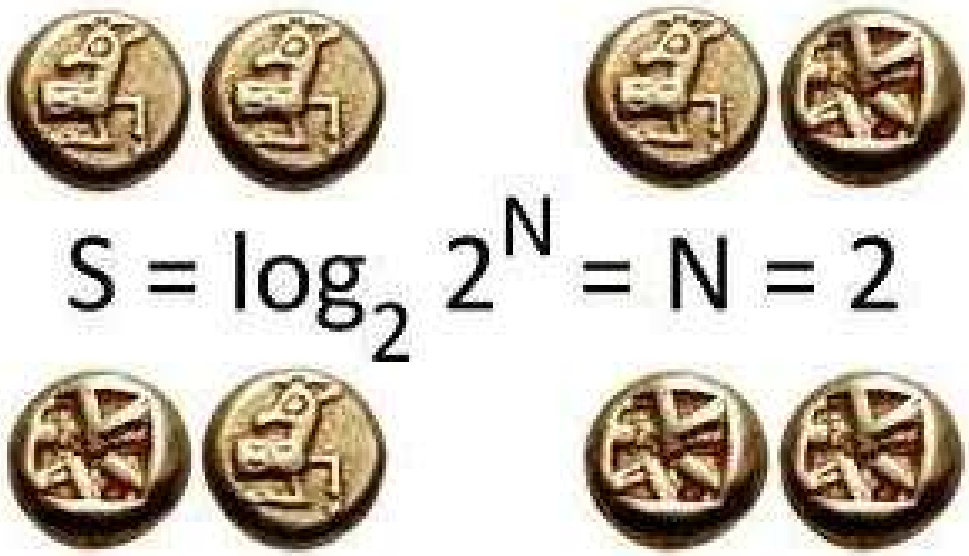


$$S = \log_2 2^N = N = 2^4$$



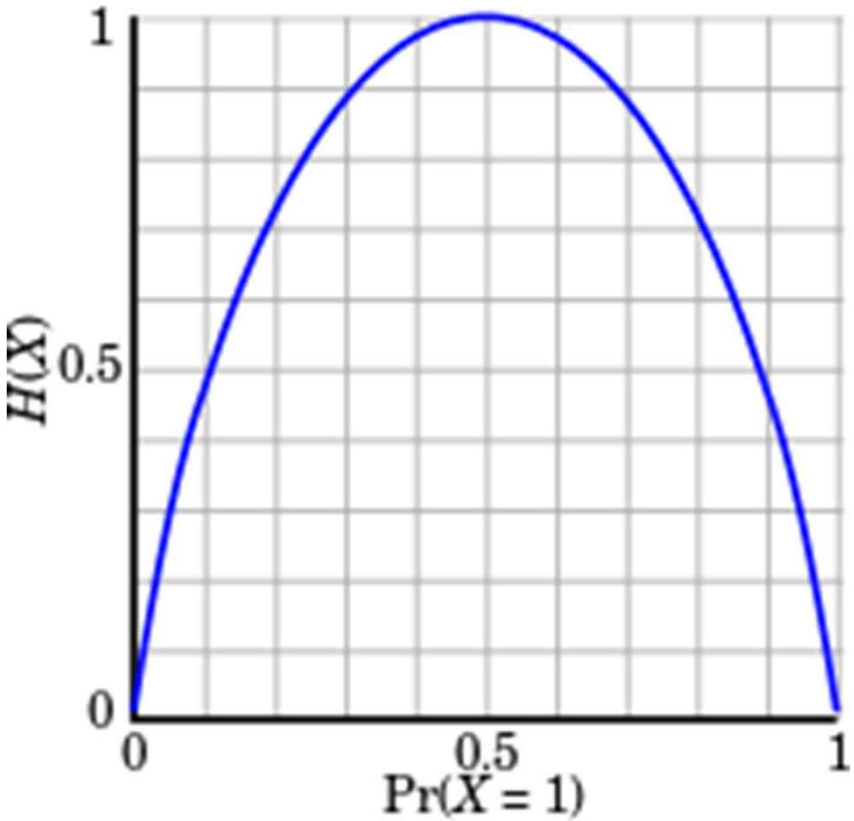
10 11

Examples



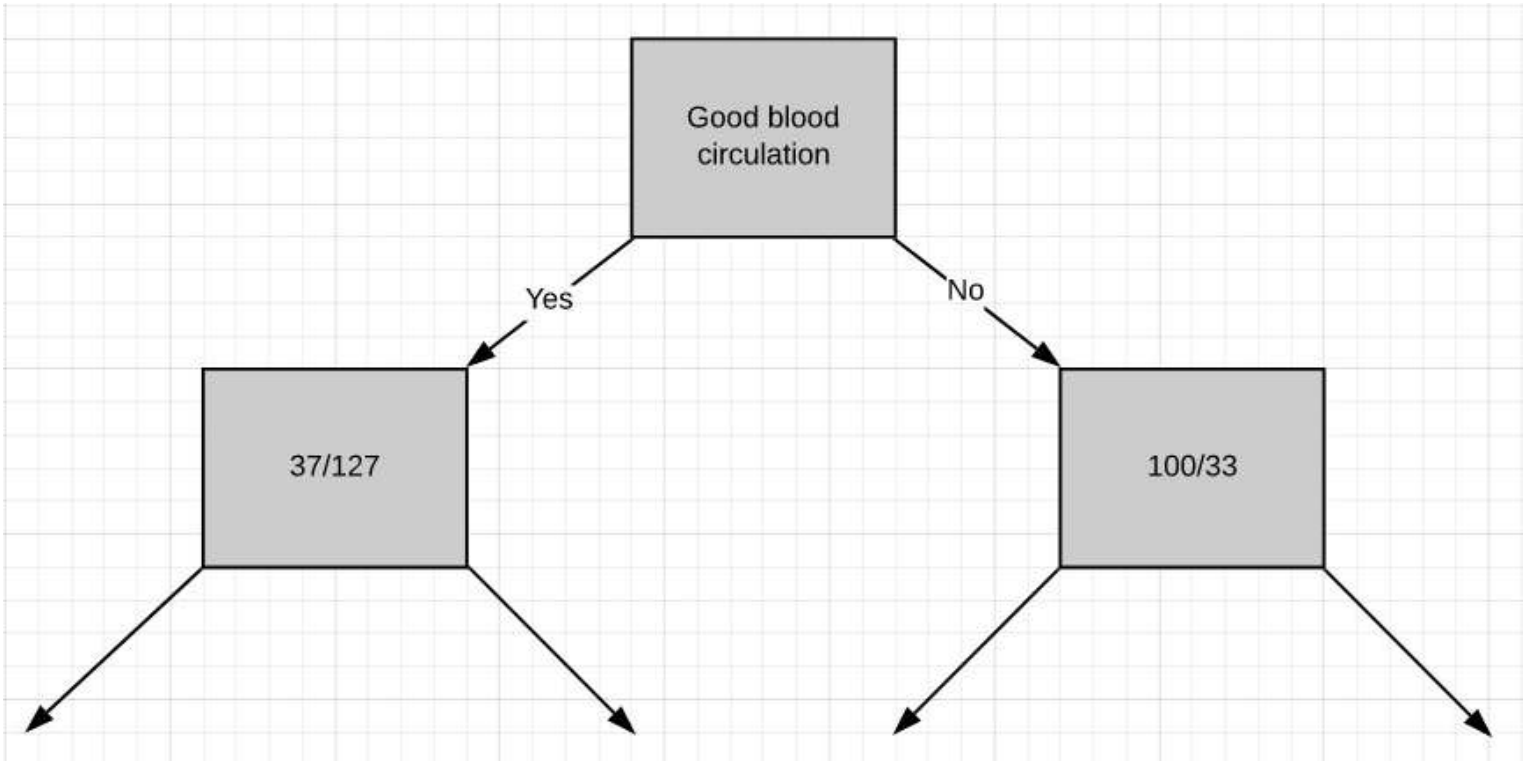
$$S = \log_2 2^N = N = 2$$

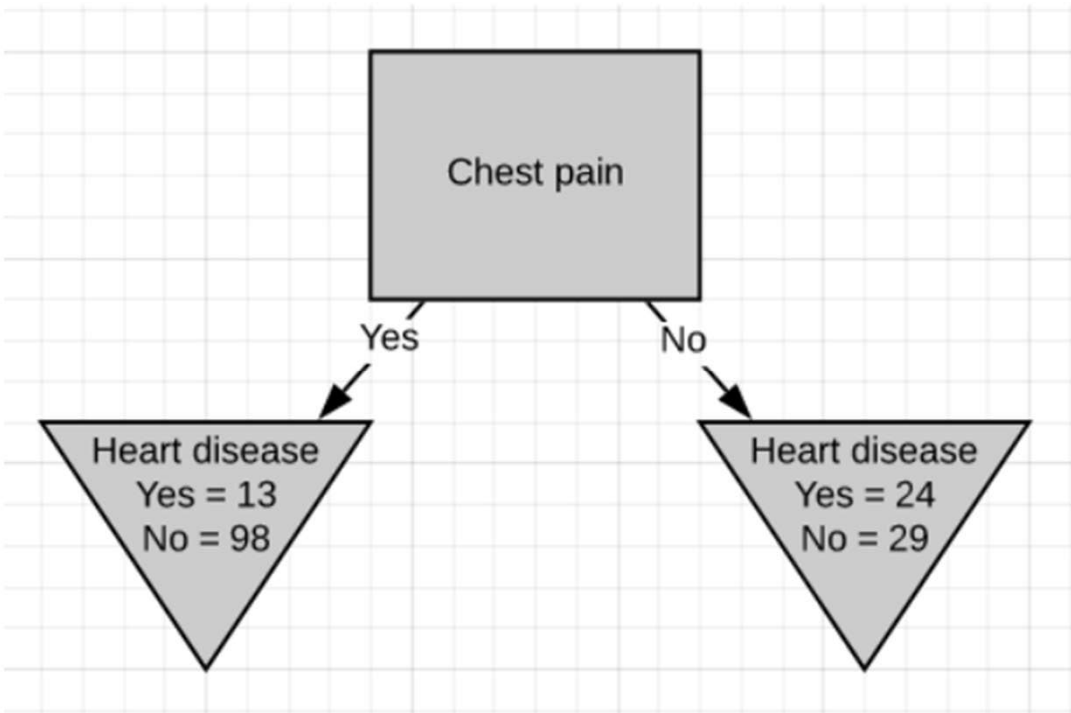
Examples



$$H(S) = -\frac{37}{164}\log_2\frac{37}{164} - \frac{127}{164}\log_2\frac{127}{164}$$

$$H(S) = 0.770$$





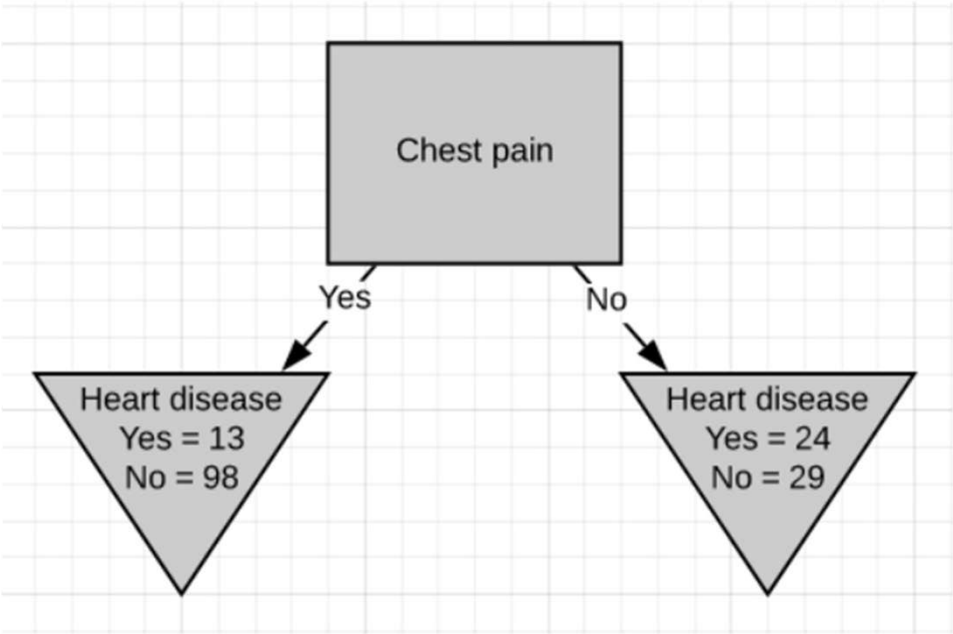
$$H(S) = -\frac{13}{111}\log_2\frac{13}{111} - \frac{98}{111}\log_2\frac{98}{111}$$

$$H(S) = 0.521$$

$$H(S) = -\frac{24}{53}\log_2\frac{24}{53} - \frac{29}{53}\log_2\frac{29}{53}$$

$$H(S) = 0.993$$

What is a gain in this situation



$$Gain(S, A) = 0.77 - \frac{111}{164}0.521 - \frac{53}{164}0.993$$

$$Gain(S, A) = 0.098$$

- https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
- https://scikit-learn.org/stable/modules/tree.html
- https://www.kaggle.com/prashant111/decision-tree-classifier-tutorial