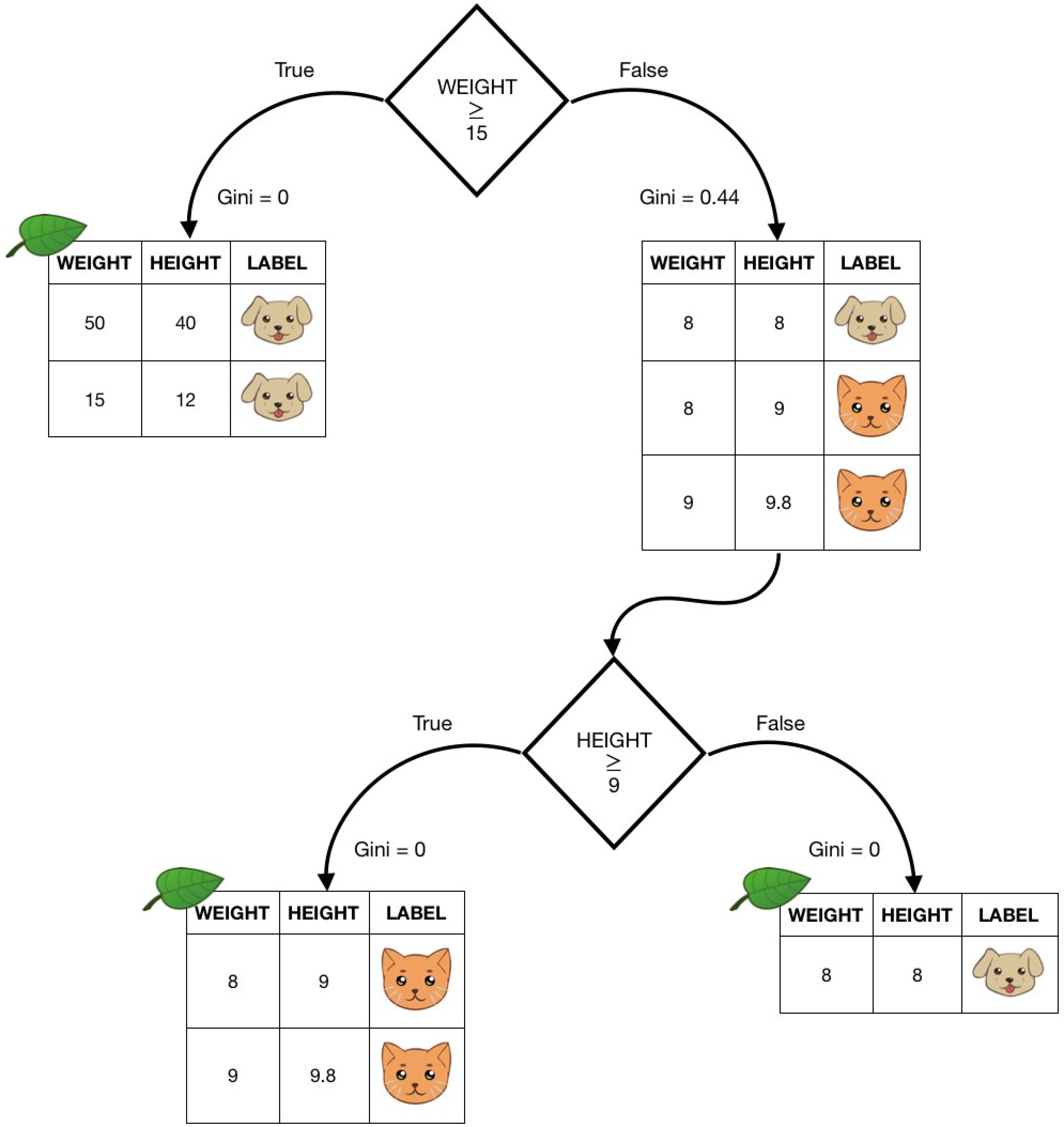
Martin Mundt, Dr. Iuliia Pliushch, Prof. Dr. Visvanathan Ramesh

Pattern Analysis & Machine Intelligence Praktikum: MLPR-19

Week 3: Random Forests

Decision trees and random forests

- **Decision tree** is a machine learning algorithm for classification and regression
- **Random forests** is an **ensemble** learning algorithm which uses **multiple** decision trees for classification and regression

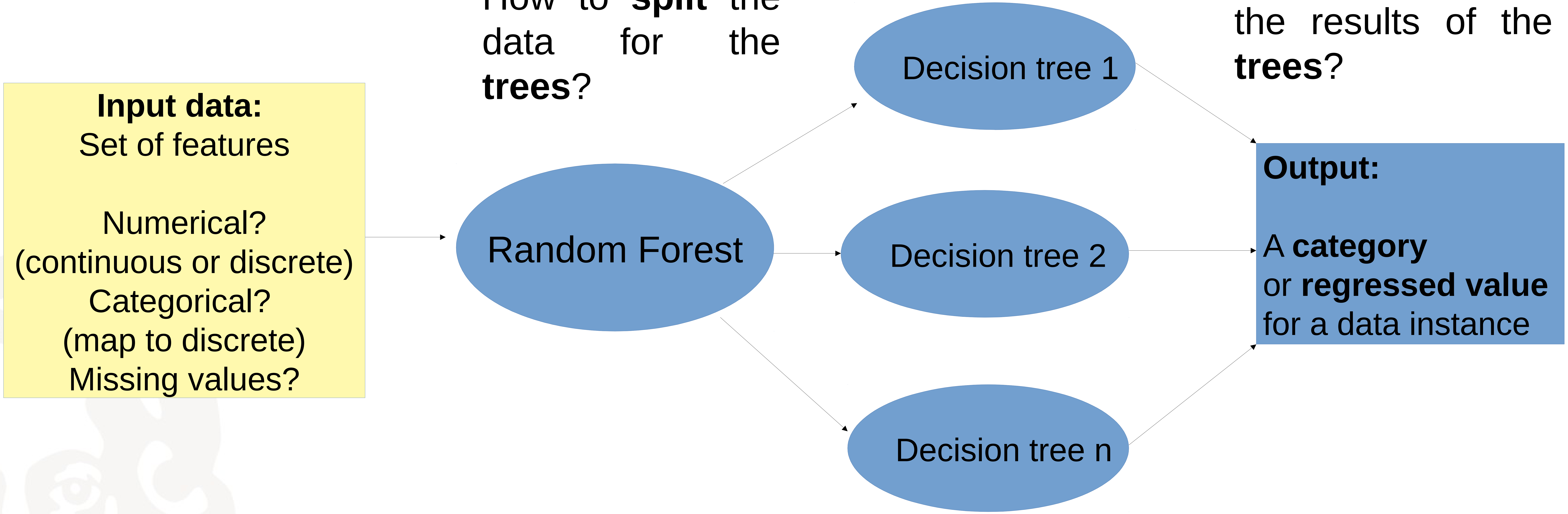


https://www.ke.tu-darmstadt.de/lehre/ws-18-19/mldm/dt.pdf

https://towardsdatascience.com/decision-tree-an-algorithm-that-works-like-the-human-brain-8bc0652f1fc6

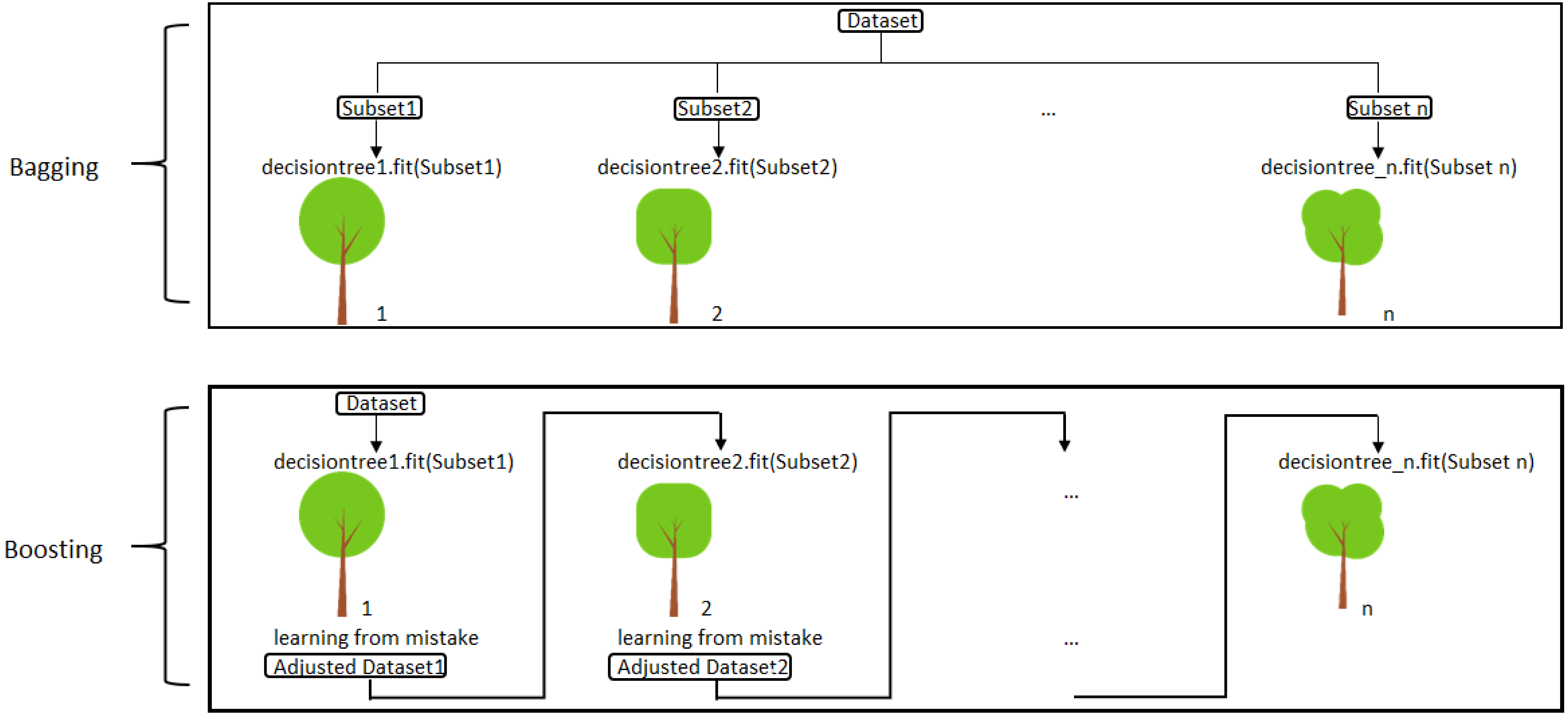
Decision trees and random forests

- How to **split** the data for the **trees**?
- How to **combine** the results of the **trees**?



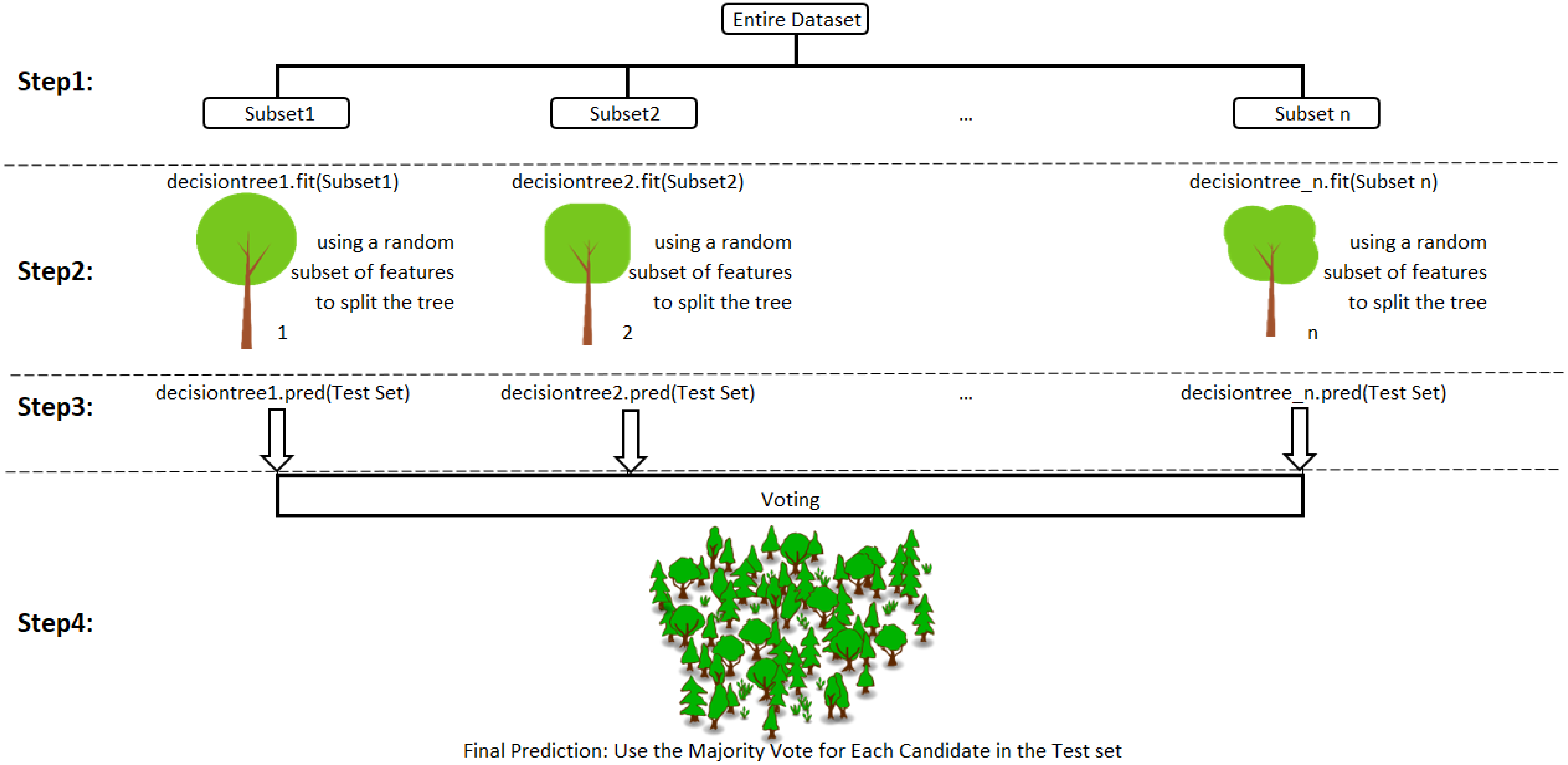
- **Pre- (while growing) and Postpruning** of trees as a means to avoid **overfitting**

Ensemble methods



https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725

Random forest (bagging)



https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725

Decision tree algorithms

- **ID3** (developed in 1986 by Ross Quinlan):
  - **categorical** features and targets
  - splitting criterion: **Information Gain**
- **C.5** (Quinlan) – commercial version of C4.5
- **C4.5** (Quinlan, 1993):
  - partitions the **continuous** features into a **discrete** set of intervals
  - suports missing values
  - splitting criterion: **Gain Ratio**
- **CART** (Classification and Regression trees):
  - similar to C4.5
  - supports **numerical target** variables (regression)
  - splitting criterion: **Gini-Index** for Classification**,** **Sum-of-Squares** for Regression

https://www.ke.tu-darmstadt.de/lehre/ws-18-19/mldm/dt.pdf https://scikit-learn.org/stable/modules/tree.html

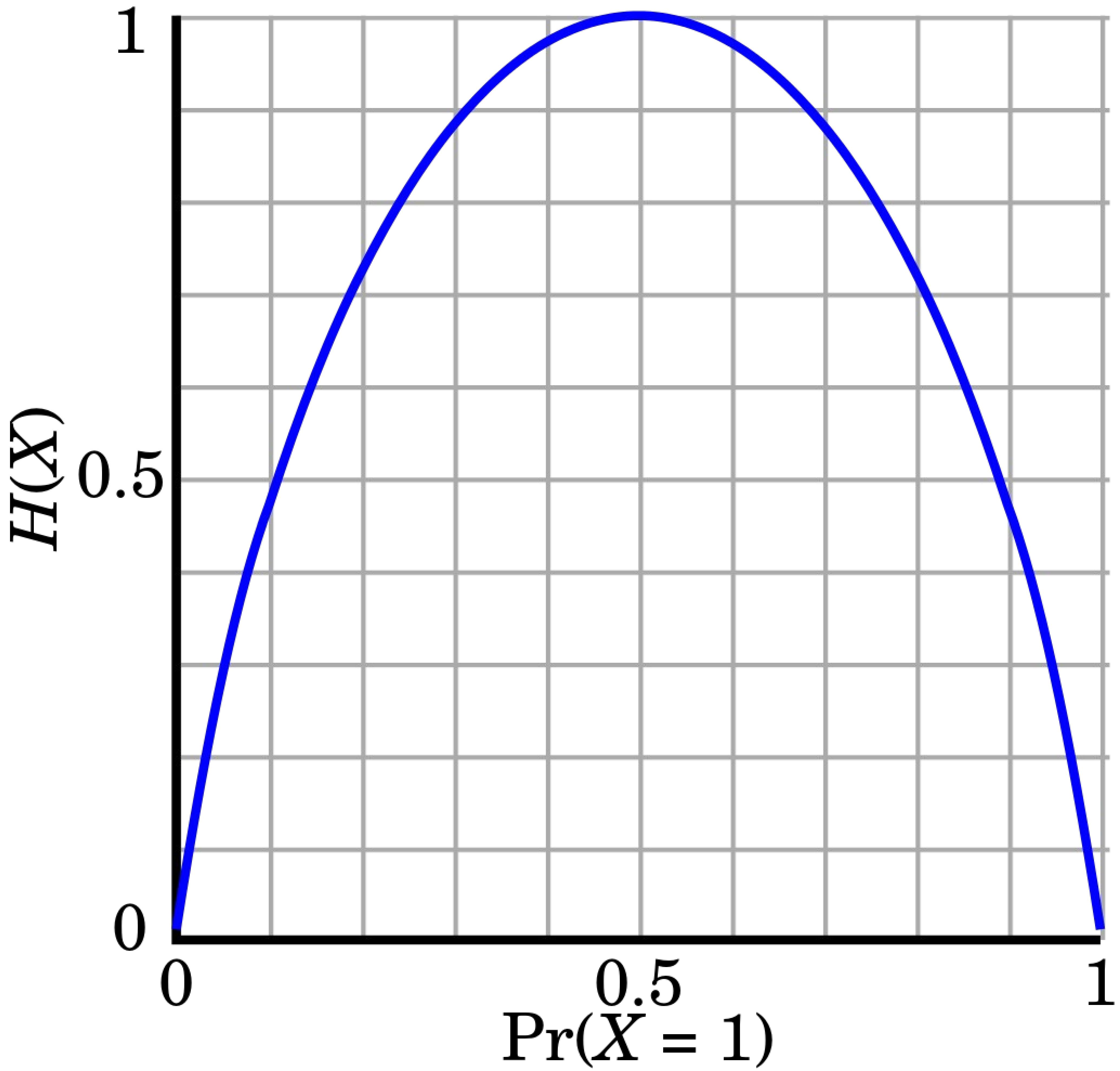
Splitting criteria: Entropy

- **Binary:**

$$-p_0 * \log_2(p_0) - p_1 * \log_2(p_1)$$

- **Multiclass:**

$$-\sum_{i \in Classes} p_i * \log_2(p_i)$$



By Brona and Alessio DamatoNewer version by Rubber Duck (☮ • ✍) - original work by Brona, published on Commons at Image:Binary entropy plot.png. Converted to SVG by Alessio Damato, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1984868

https://www.ke.tu-darmstadt.de/lehre/ws-18-19/mldm/dt.pdf

Splitting criteria

$$H(parent) = -\frac{2}{5} * \log_2\left(\frac{2}{5}\right) - \frac{3}{5} * \log_2\left(\frac{3}{5}\right) = 0.97$$

- **Entropy Gain:**

$$Gain(S, A) = H(S) - \sum_i \frac{|S_i|}{|S|} H(S_i)$$

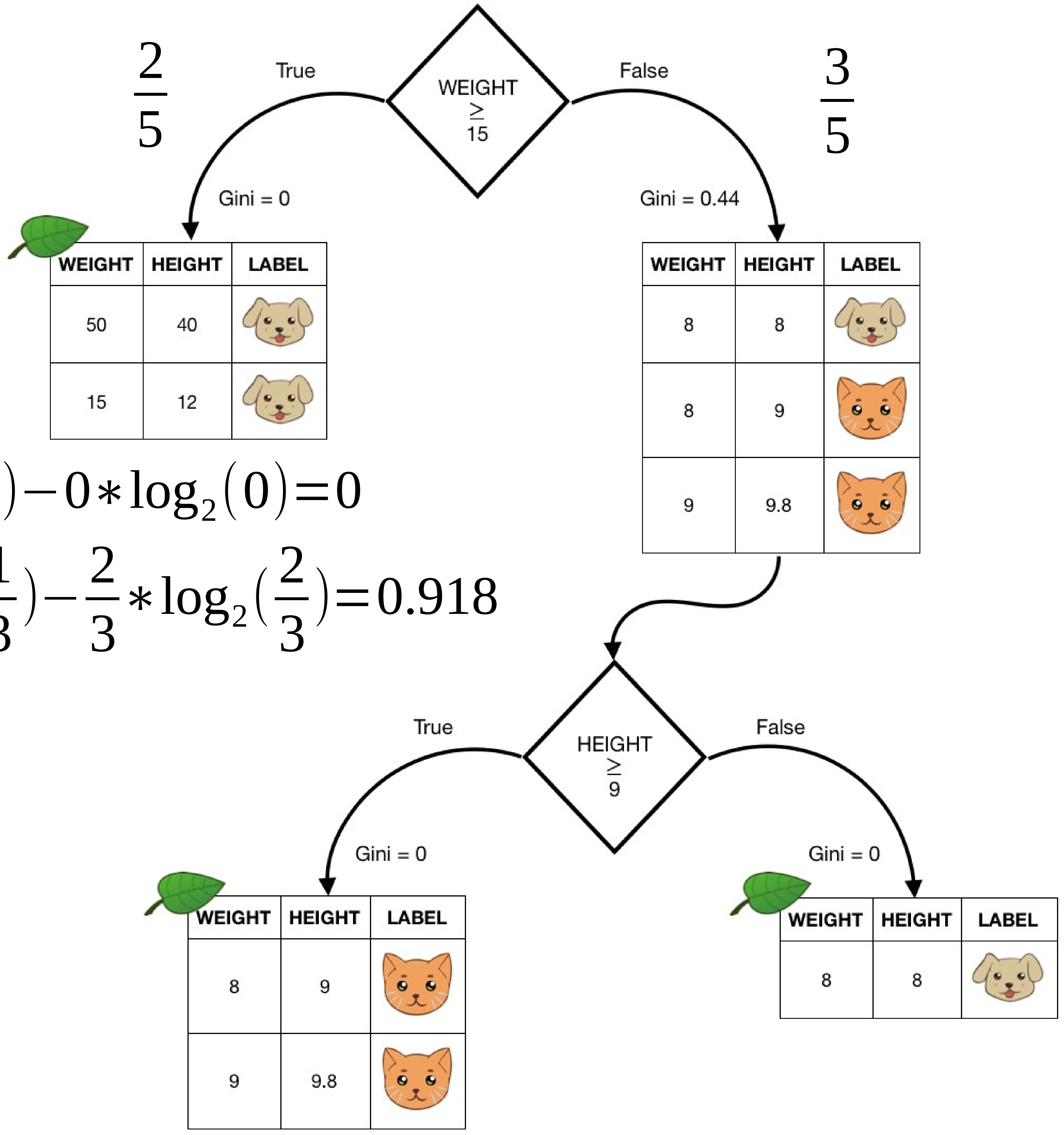
- **Intrinsic Information:**

$$H(leftchild) = -\frac{2}{2} * \log_2\left(\frac{2}{2}\right) - 0 * \log_2(0) = 0$$

$$H(rightchild) = -\frac{1}{3} * \log_2\left(\frac{1}{3}\right) - \frac{2}{3} * \log_2\left(\frac{2}{3}\right) = 0.918$$

$$IntI(S, A) = -\sum_i \frac{|S_i|}{|S|} \log_2\left(\frac{|S_i|}{|S|}\right)$$

- **Gain Ratio:** $\dfrac{Gain(S, A)}{IntI(S, A)}$



https://www.ke.tu-darmstadt.de/lehre/ws-18-19/mldm/dt.pdf

https://towardsdatascience.com/decision-tree-an-algorithm-that-works-like-the-human-brain-8bc0652f1fc6

Splitting criteria (CART)

- **Gini (impurity measure)**
  – for classification

$$Gini(S) = 1 - \sum_{i \in Classes} p_i^2$$

$$Gini(S, A) = \sum_i \frac{|S_i|}{|S|} Gini(S_i)$$

- **MSE (Mean Squared Error)**
  – for regression

$$MSE(S) = \frac{1}{N} \sum_{i \in Ndata} (y_i - \mu_y)^2$$

$$Var(X) = E[(X-\mu)^2] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

https://www.ke.tu-darmstadt.de/lehre/ws-18-19/mldm/dt.pdf

https://en.wikipedia.org/wiki/Variance

https://web.stanford.edu/class/stats202/content/lec19.pdf

Decision tree

- Input: Set of features, class to predict

- 1. Create a (root) node
- 2. If termination criteria are met, make it a **leaf**
- 2. Select the best **feature** to split the data according to **criterion (loop over selected features)**
- 3. Split the **data** accordingly
- 4. Create subtrees for each **data subset (RECURSION!)**

Titanic dataset