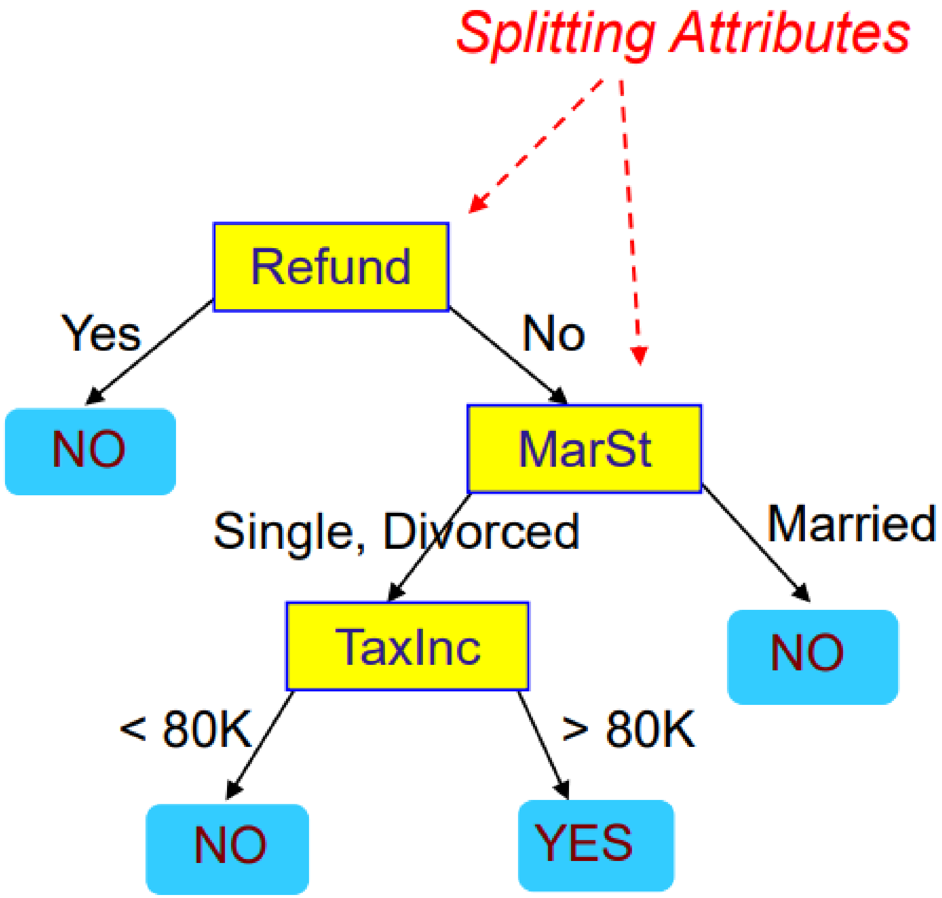
Árboles de decisión y ensambladores

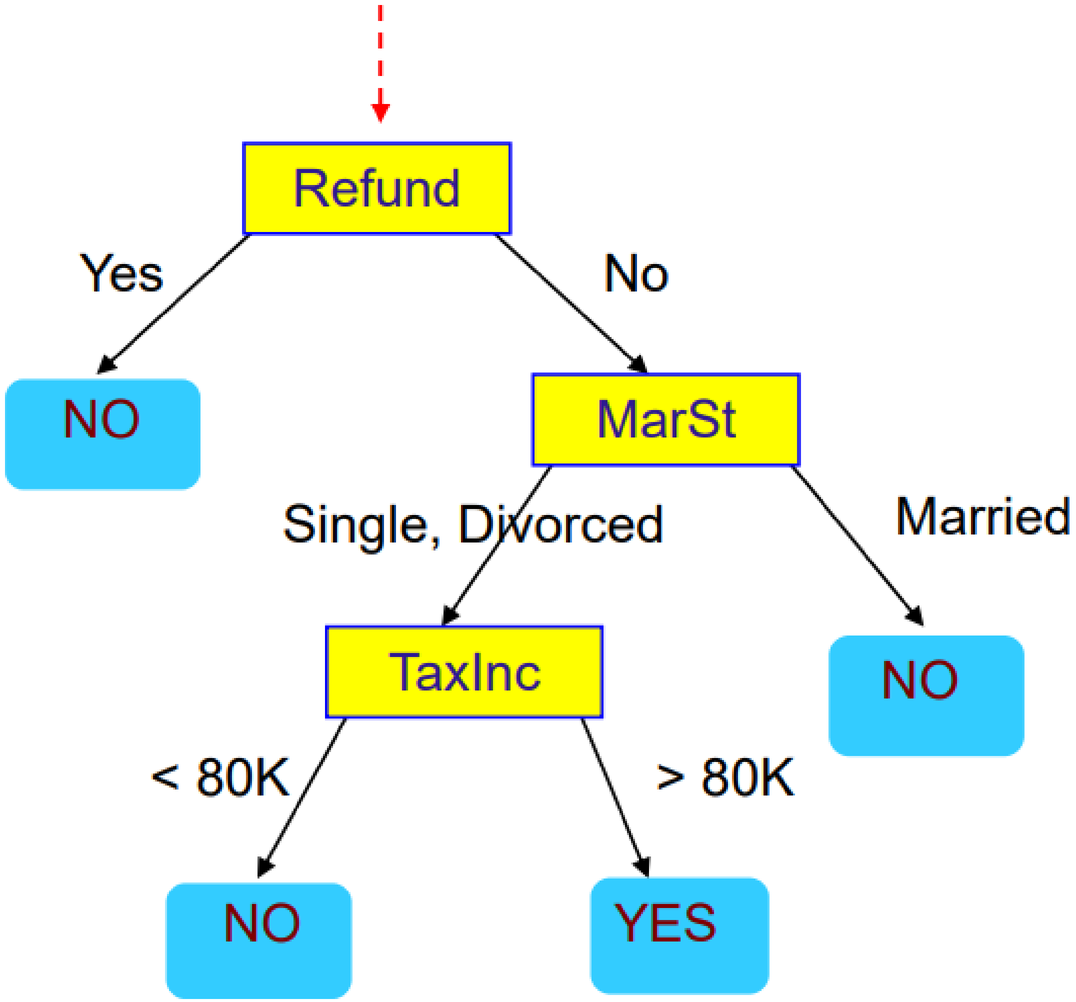
Classification trees

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



How to use

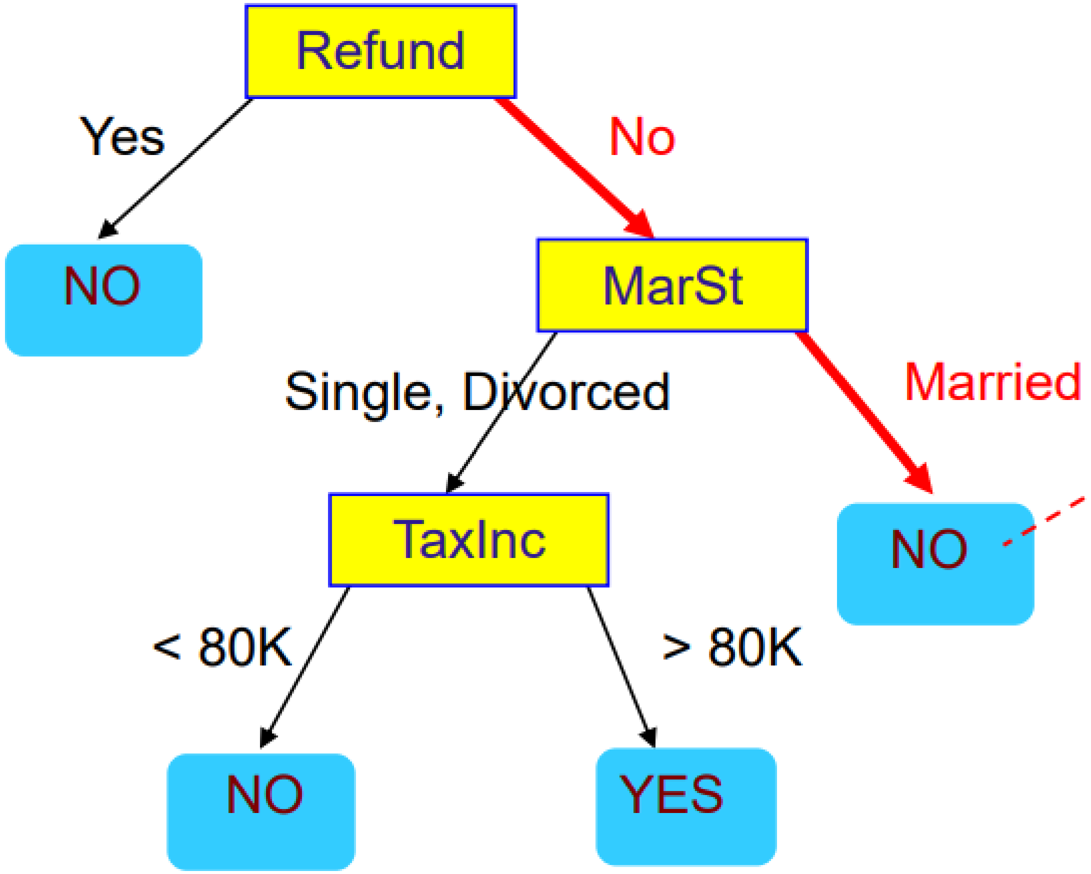


Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Aplicando el modelo a los datos de evaluación

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Test Data



Assign Cheat to “No”

Impurity Criterion

Gini Index

$$I_G = 1 - \sum_{j=1}^{c} p_j^2$$

p_j: proportion of the samples that belongs to class c for a particular node

Entropy

$$I_H = - \sum_{j=1}^{c} p_j log_2(p_j)$$

p_j: proportion of the samples that belongs to class c for a particular node.

*This is the the definition of entropy for all non-empty classes ($p \neq 0$). The entropy is 0 if all samples at a node belong to the same class.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Play Golf	
Yes	No
9	5

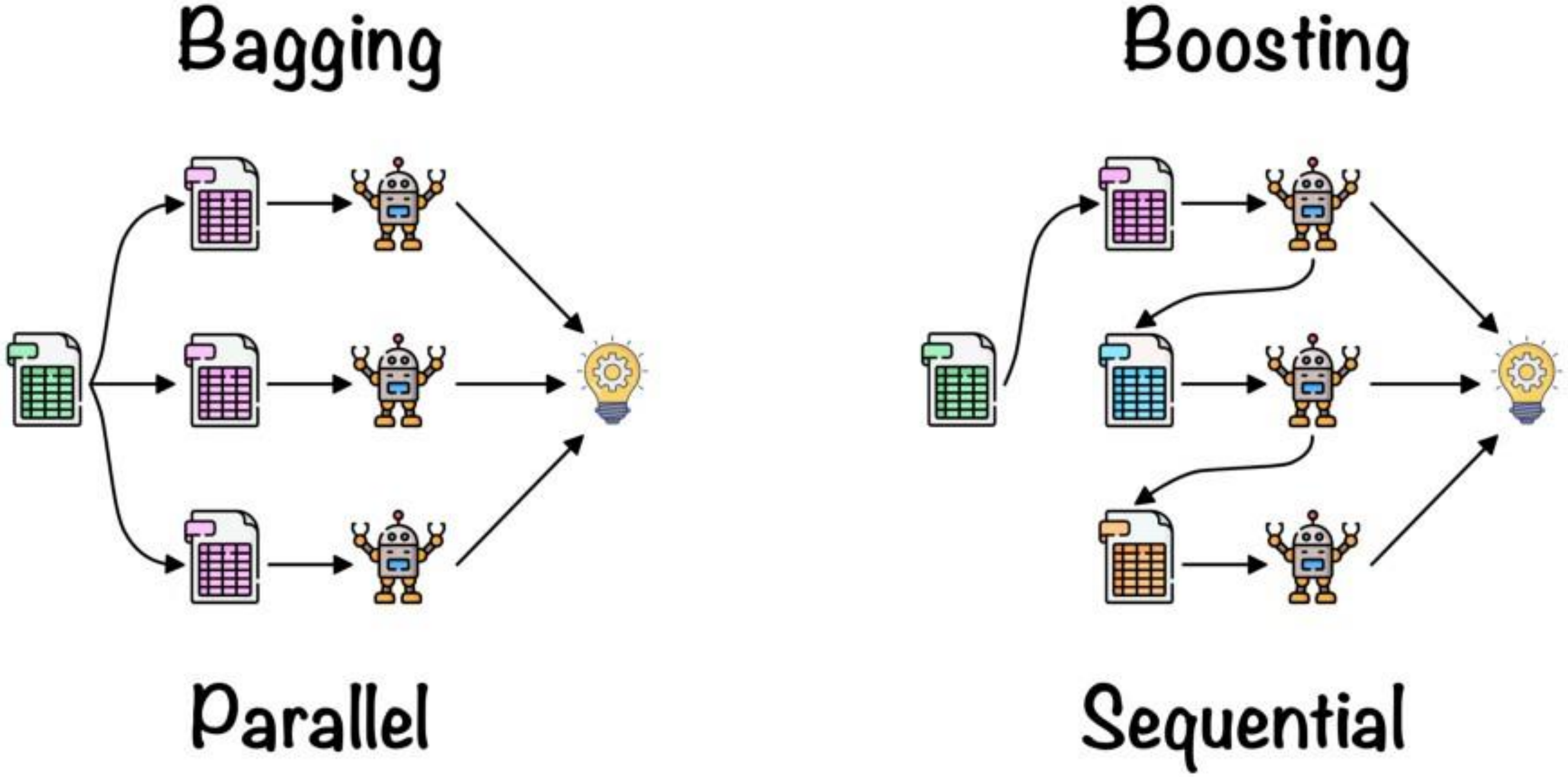
Entropy(PlayGolf) = Entropy (5,9)

= Entropy (0.36, 0.64)

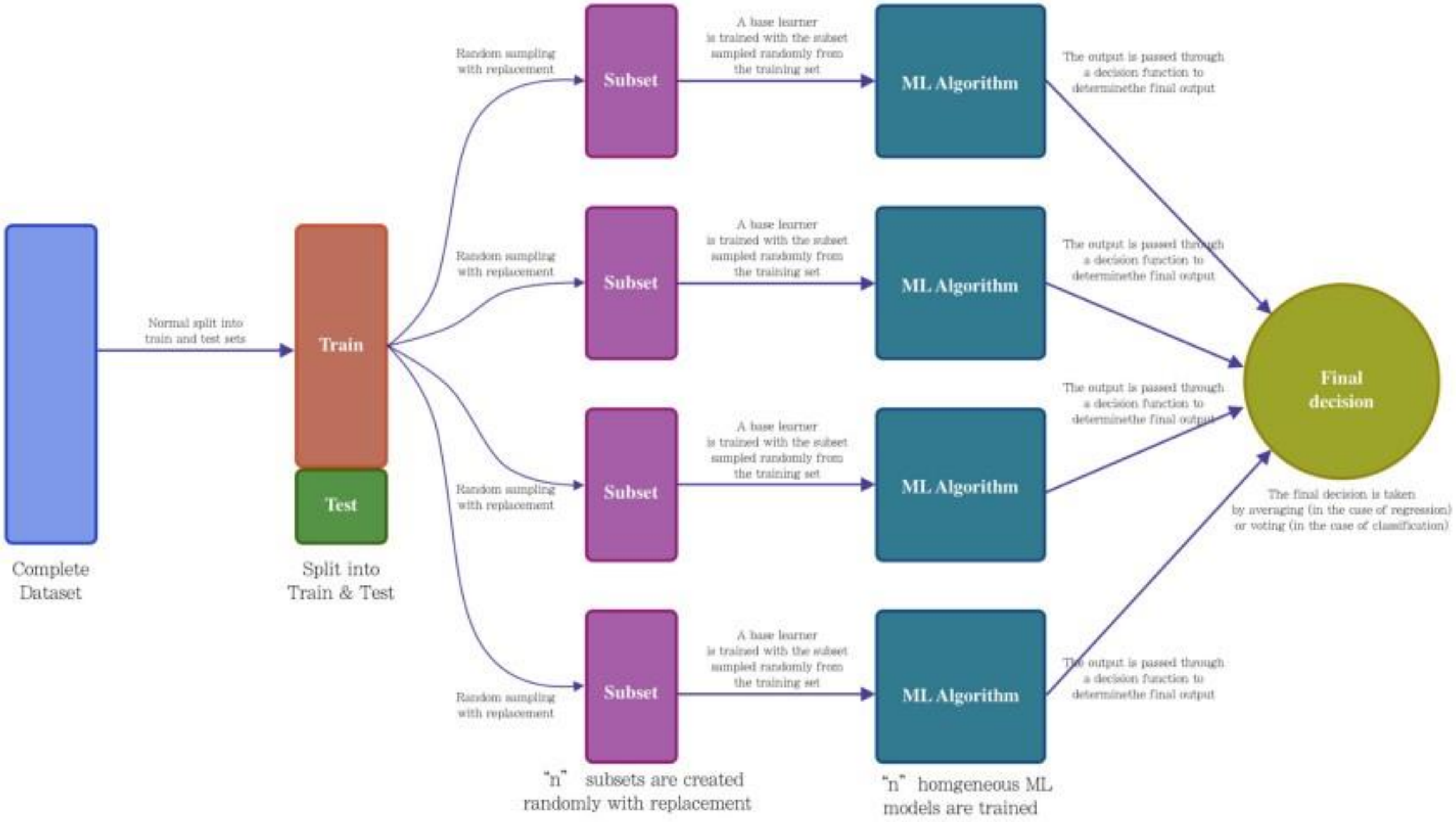
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)

= 0.94

Bagging y boosting



Bagging



Complete Dataset

Normal split into train and test sets

Train

Test

Split into Train & Test

Random sampling with replacement

Random sampling with replacement

Random sampling with replacement

Random sampling with replacement

Subset

Subset

Subset

Subset

"n" subsets are created randomly with replacement

A base learner is trained with the subset sampled randomly from the training set

A base learner is trained with the subset sampled randomly from the training set

A base learner is trained with the subset sampled randomly from the training set

A base learner is trained with the subset sampled randomly from the training set

ML Algorithm

ML Algorithm

ML Algorithm

ML Algorithm

"n" homgeneous ML models are trained

The output is passed through a decision function to determinethe final output

The output is passed through a decision function to determinethe final output

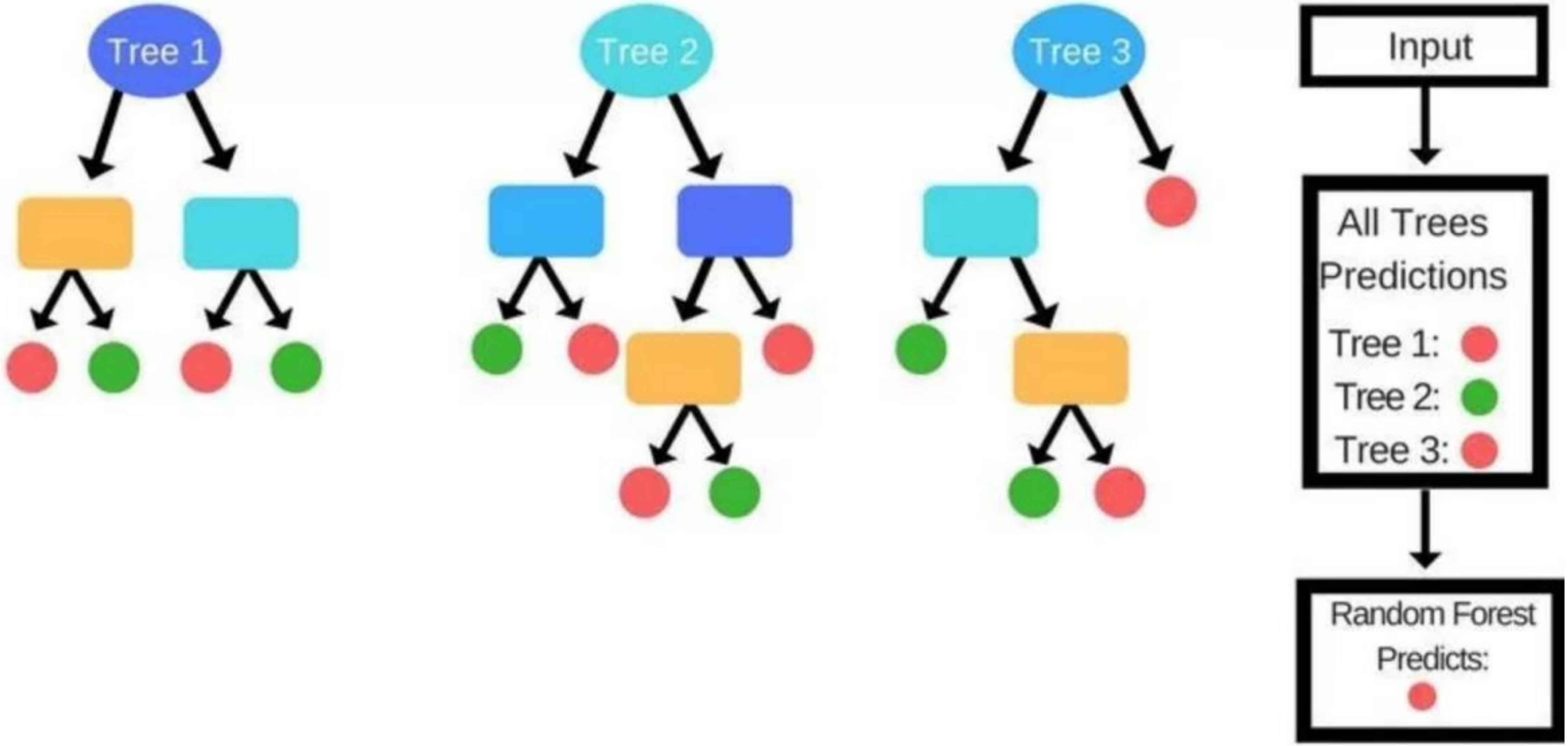
The output is passed through a decision function to determinethe final output

The output is passed through a decision function to determinethe final output

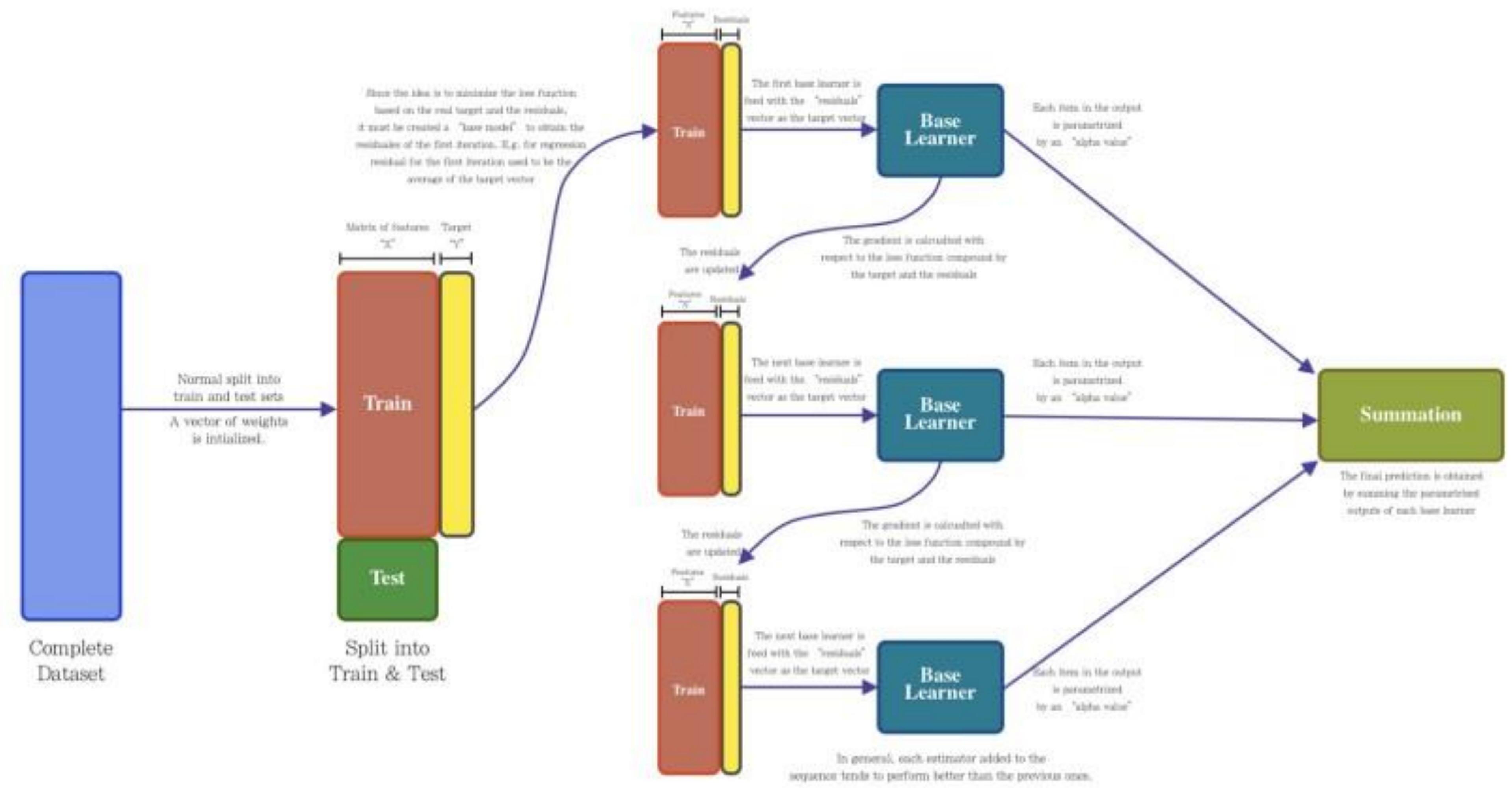
Final decision

The final decision is taken by averaging (in the case of regression) or voting (in the case of classification)

Random Forest



Boosting

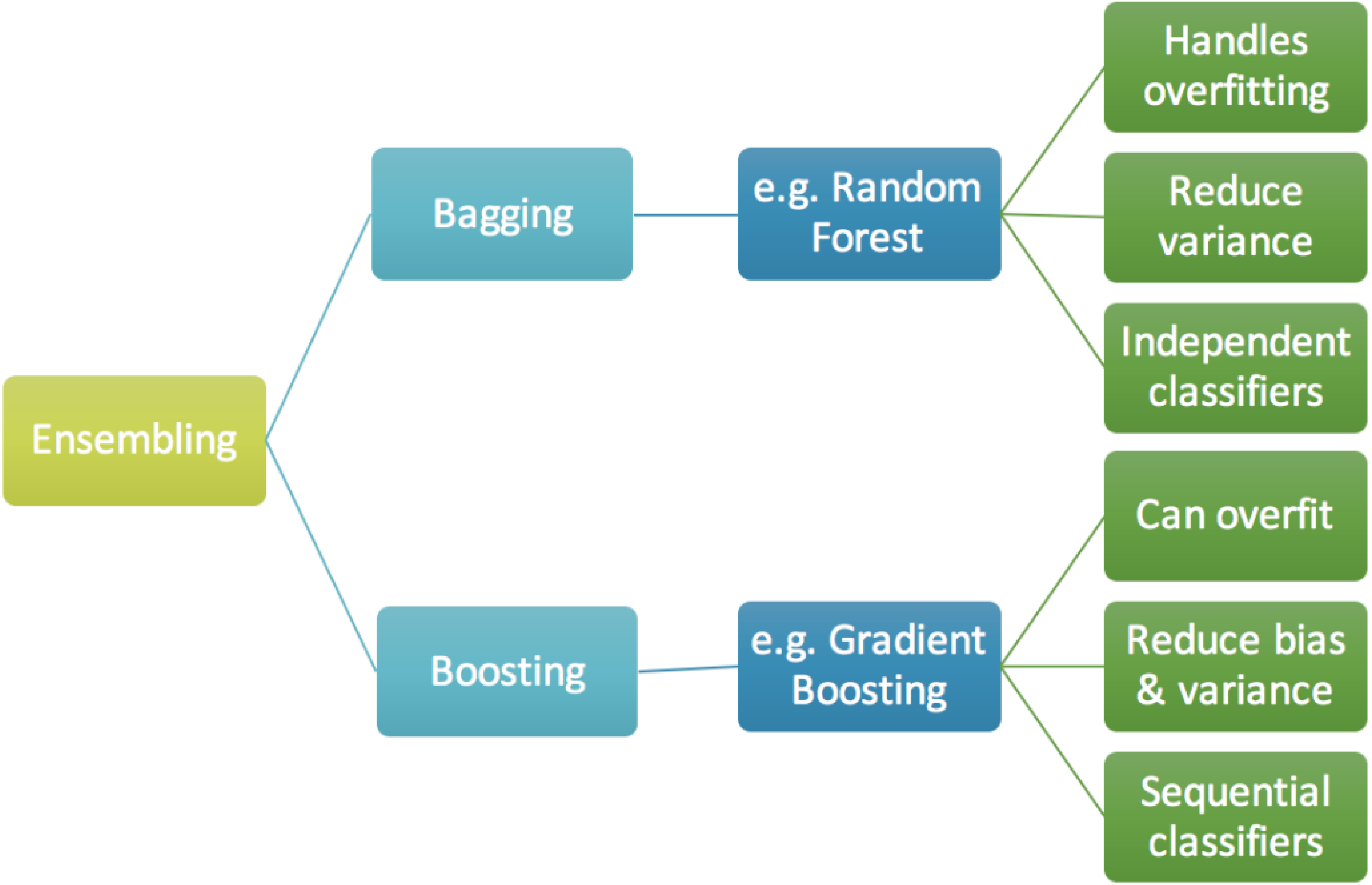


Complete Dataset
Normal split into train and test sets
A vector of weights is intialized.
Matrix of features "X"
Target "Y"
Train
Test
Split into Train & Test
Since the idea is to minimize the loss function based on the real target and the residuals, it must be created a "base model" to obtain the residuales of the first iteration. E.g. for regression residual for the first iteration used to be the average of the target vector
Features "X"
Residuals
Train
The first base learner is feed with the "residuals" vector as the target vector
Base Learner
Each item in the output is parametrized by an "alpha value"
The gradient is calculated with respect to the loss function compound by the target and the residuals
The residuals are updated
Features "X"
Residuals
Train
The next base learner is feed with the "residuals" vector as the target vector
Base Learner
Each item in the output is parametrized by an "alpha value"
The gradient is calculated with respect to the loss function compound by the target and the residuals
The residuals are updated
Features "X"
Residuals
Train
The next base learner is feed with the "residuals" vector as the target vector
Base Learner
Each item in the output is parametrized by an "alpha value"
In general, each estimator added to the sequence tends to perform better than the previous ones.
Summation
The final prediction is obtained by summing the parametrized outputs of each base learner

Random Forest vs Gradient Boosting

Random Forest	Gradient Boosting
Easier to tune Harder to overfit Show very low variance Easier to parallelize	Better accuracy with less trees if the data is noisy exhibits higher variance

Ensembling summary



Ensembling

Bagging

e.g. Random Forest

Handles overfitting

Reduce variance

Independent classifiers

Boosting

e.g. Gradient Boosting

Can overfit

Reduce bias & variance

Sequential classifiers