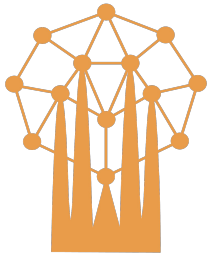
AI Saturdays

Saturdays.AI
Barcelona

Today

9:30h **Decision Trees + Random Forest**

11:30h **Coffee break**



Saturdays.AI
Barcelona

Decision Trees

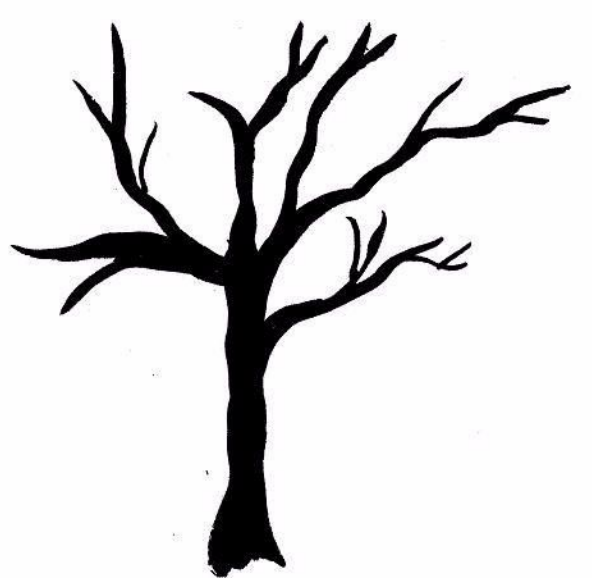
by Saturdays AI

Get ready for the future AI!

Decision Tree Algorithms

Saturdays.AI Barcelona

A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values or classes (homogenous).

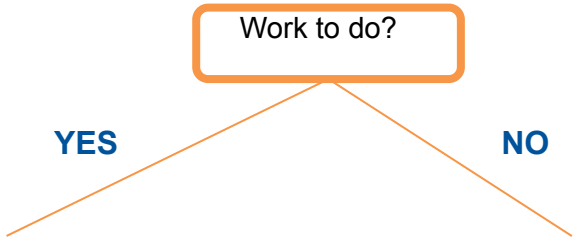


Decide a plan with friends

Saturdays.AI
Barcelona

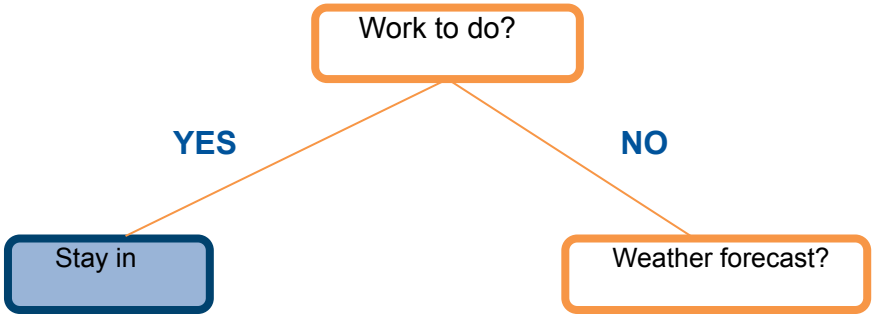
Work to do?

Decide a plan with friends



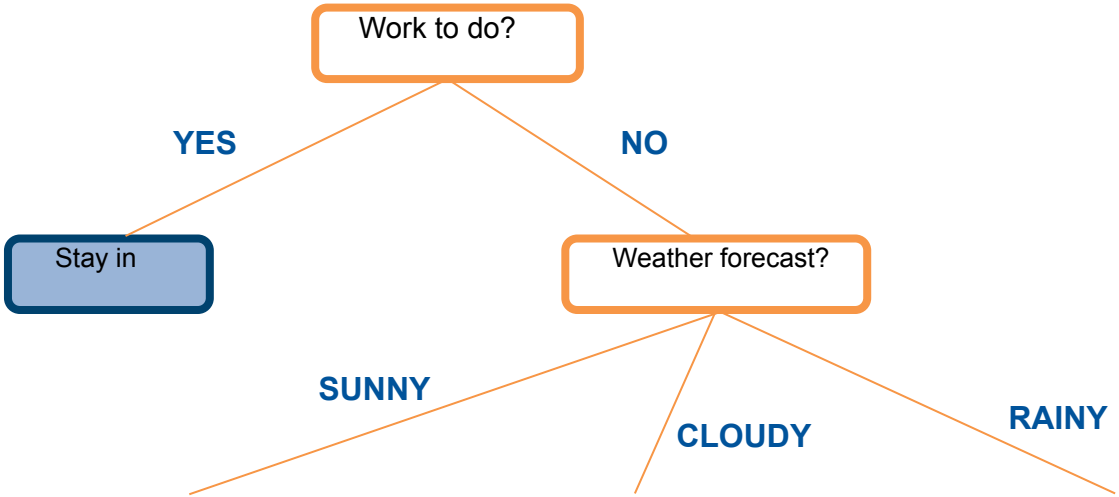
Saturdays.AI
Barcelona

Decide a plan with friends

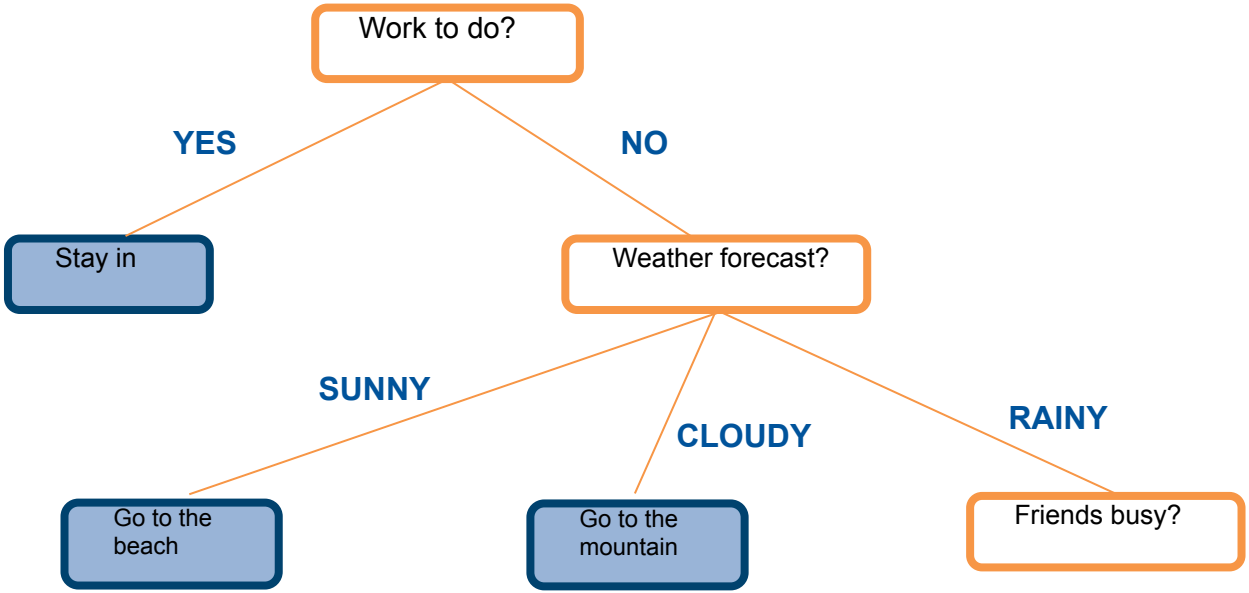


Saturdays.AI
Barcelona

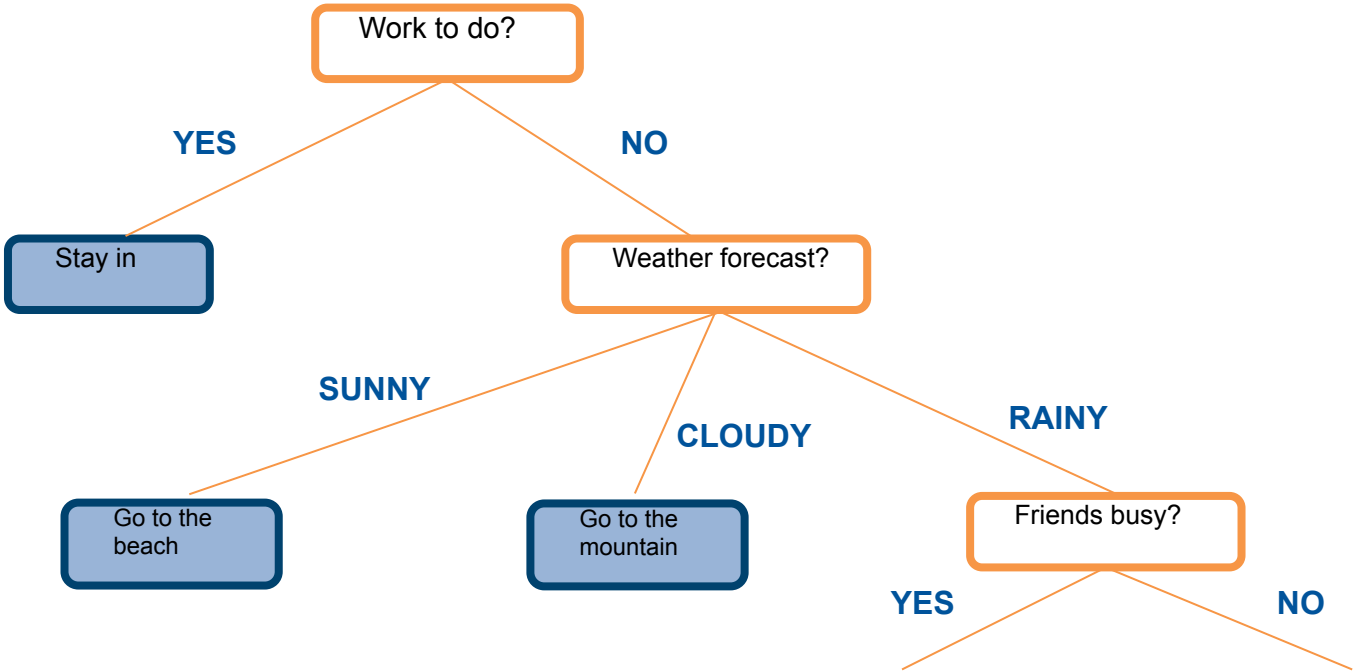
Decide a plan with friends



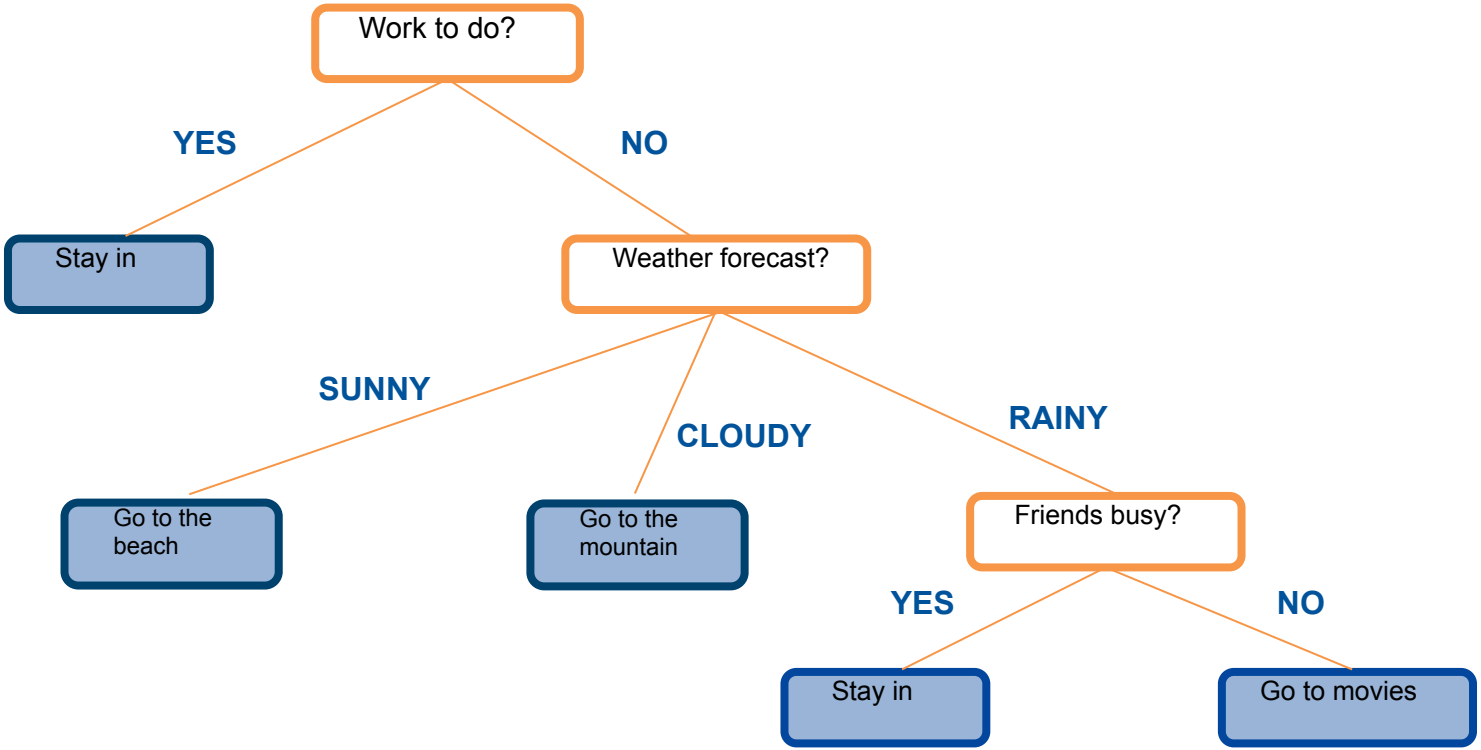
Decide a plan with friends



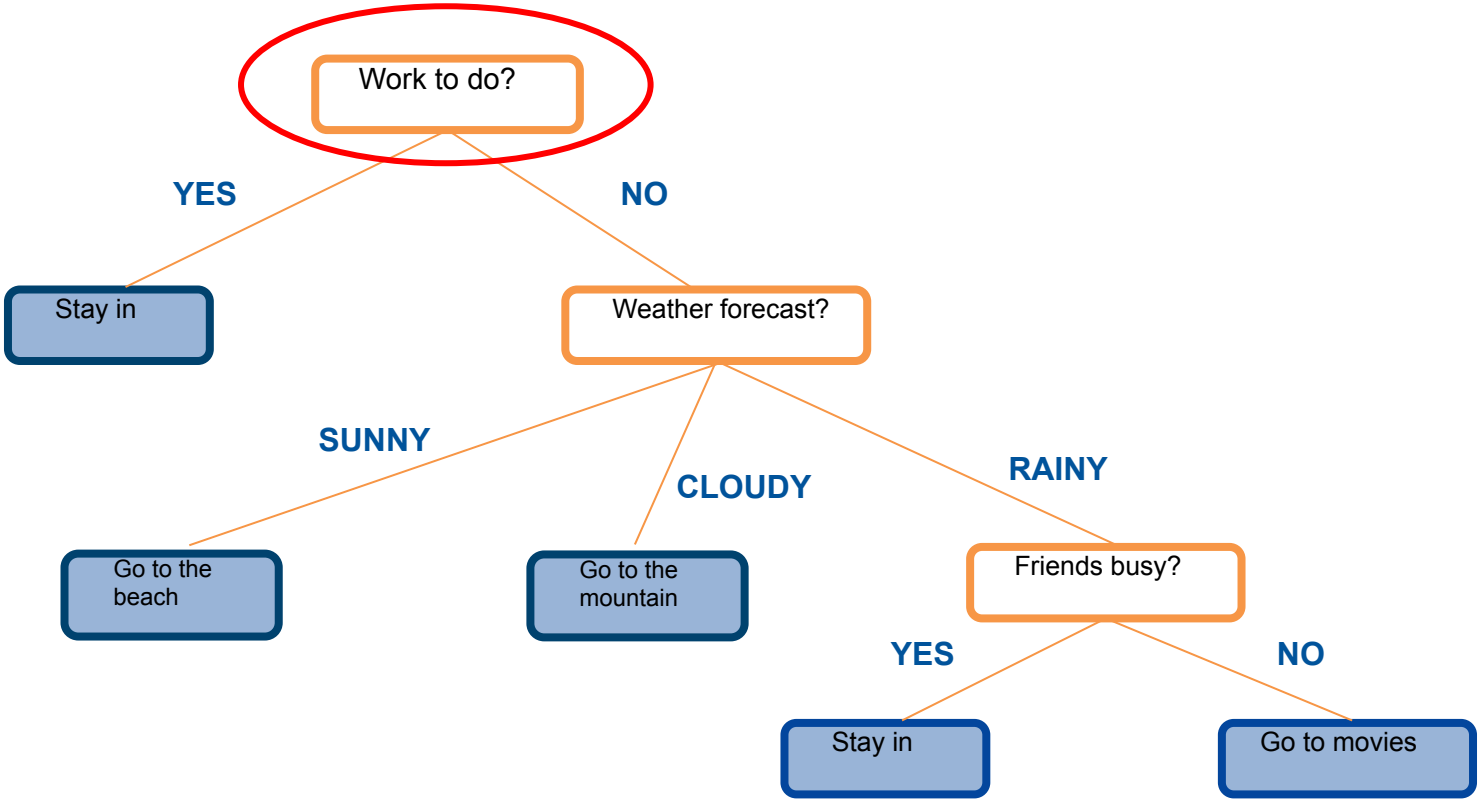
Decide a plan with friends



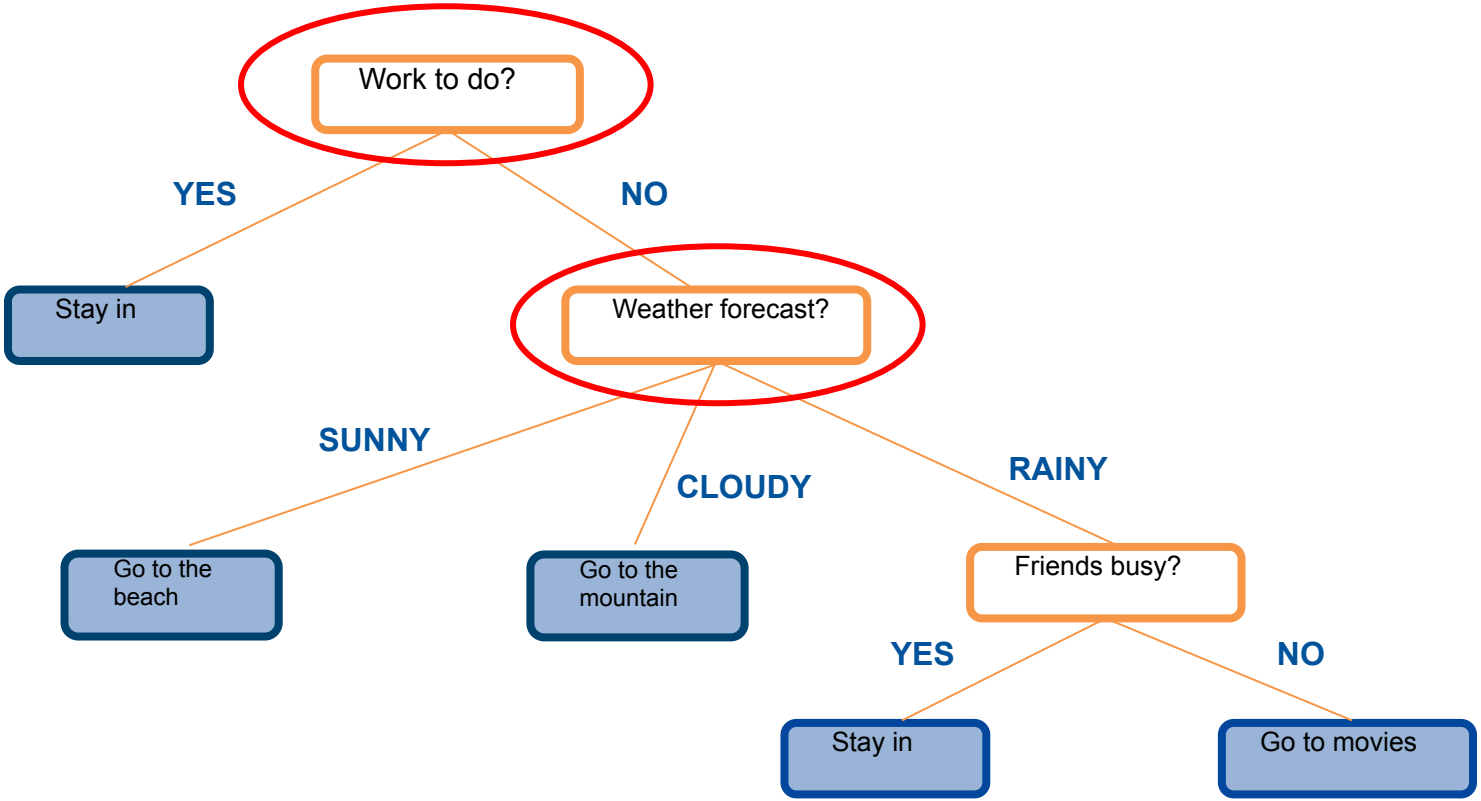
Decide a plan with friends



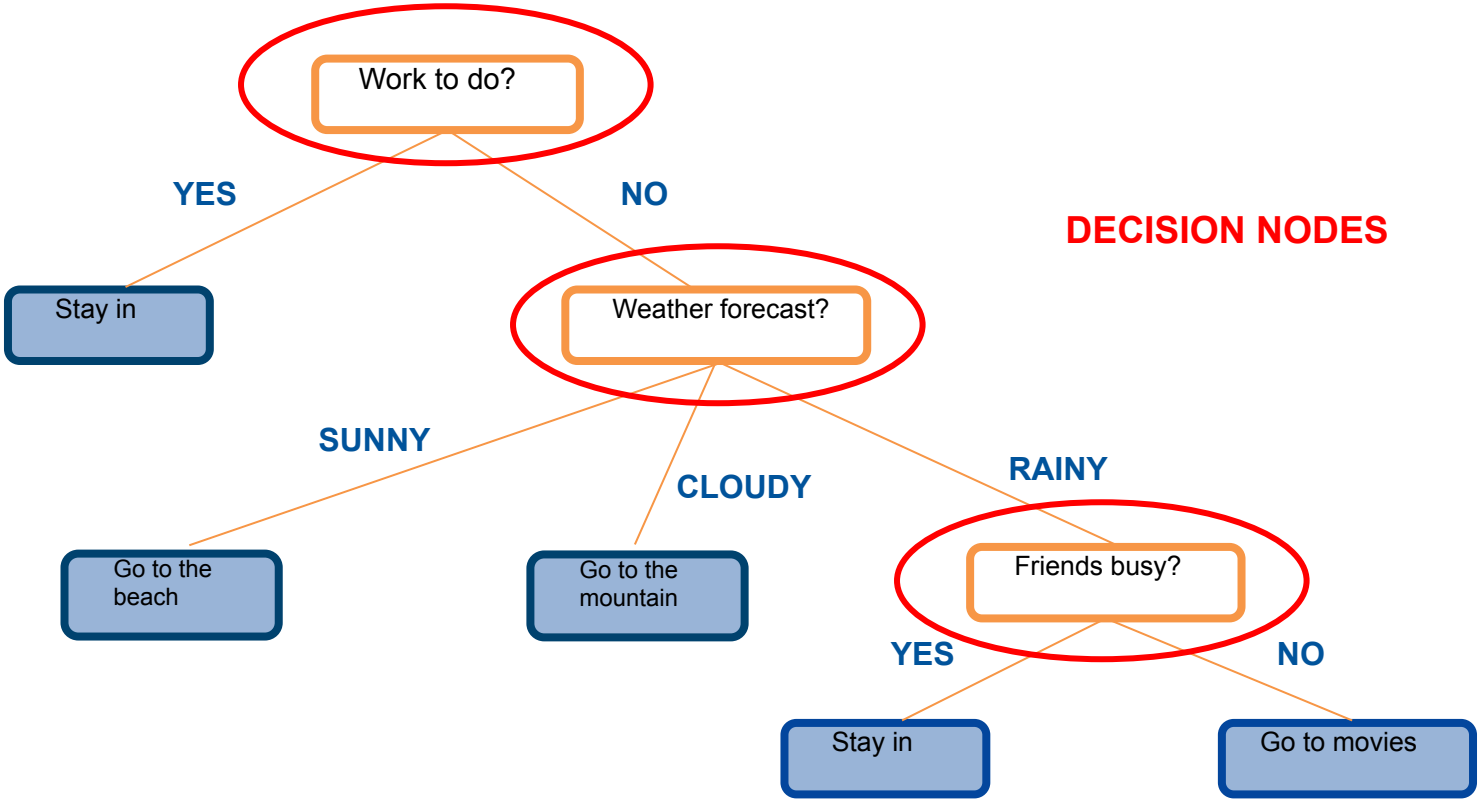
Decide a plan with friends



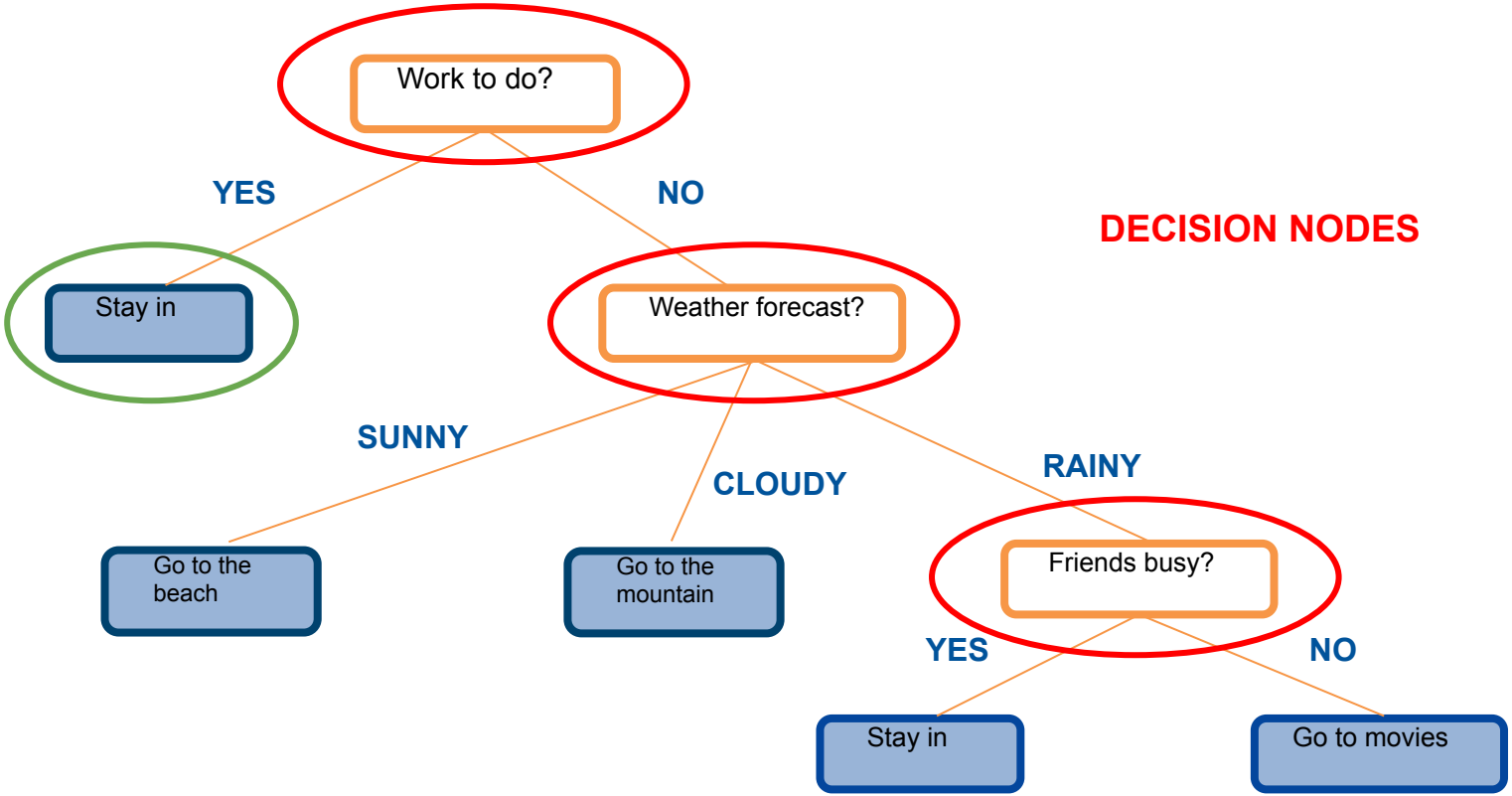
Decide a plan with friends



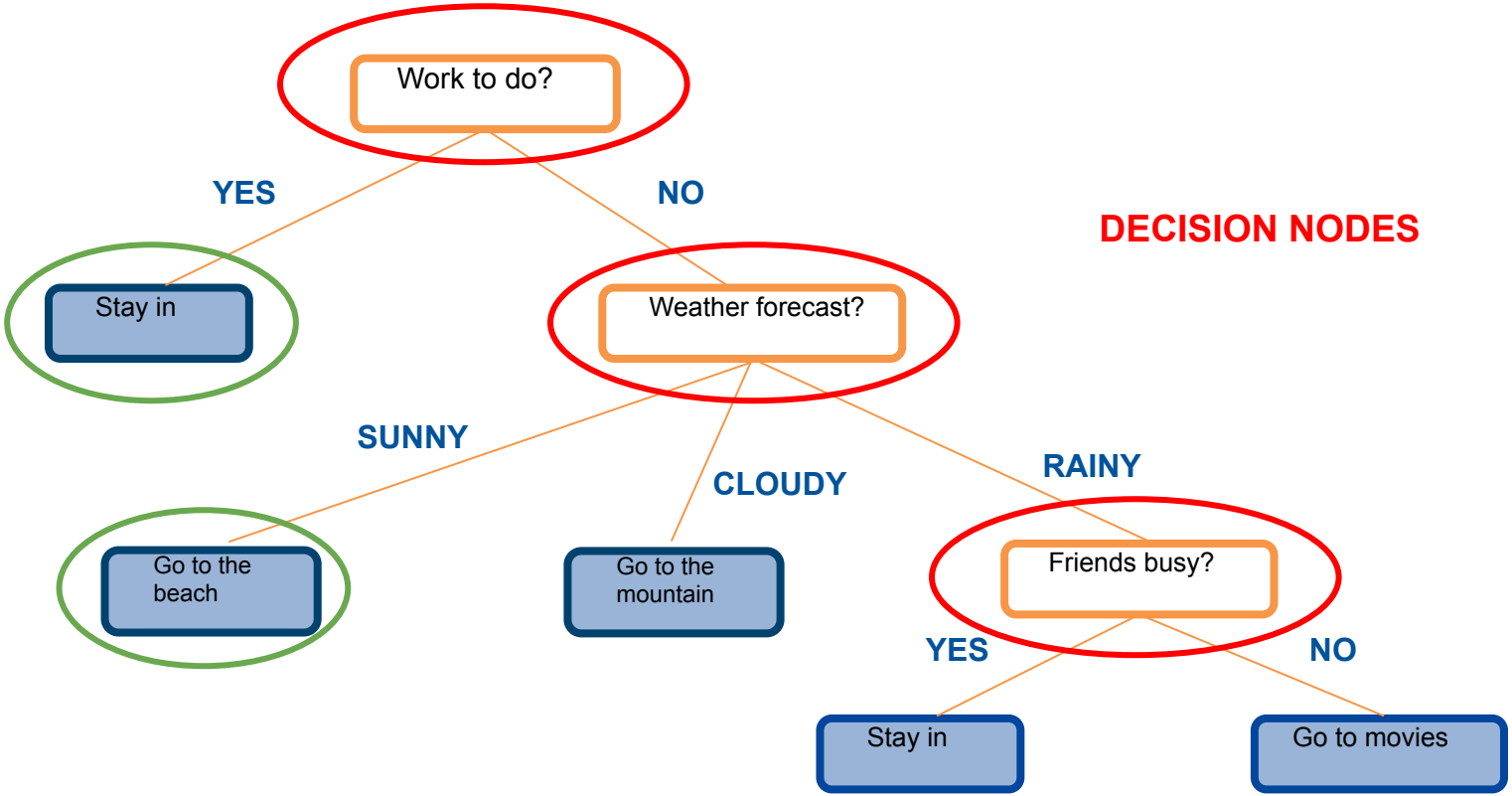
Decide a plan with friends



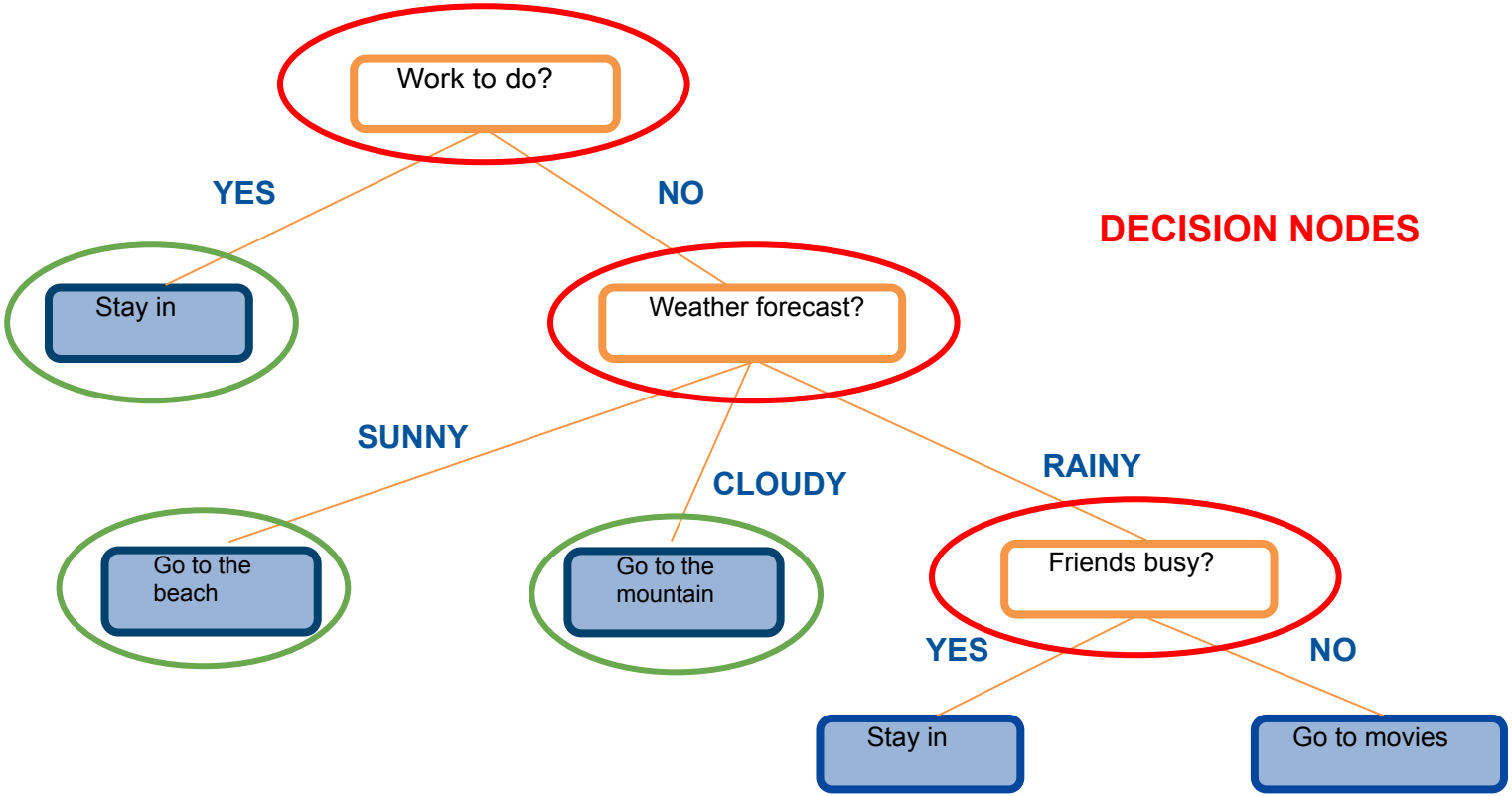
Decide a plan with friends



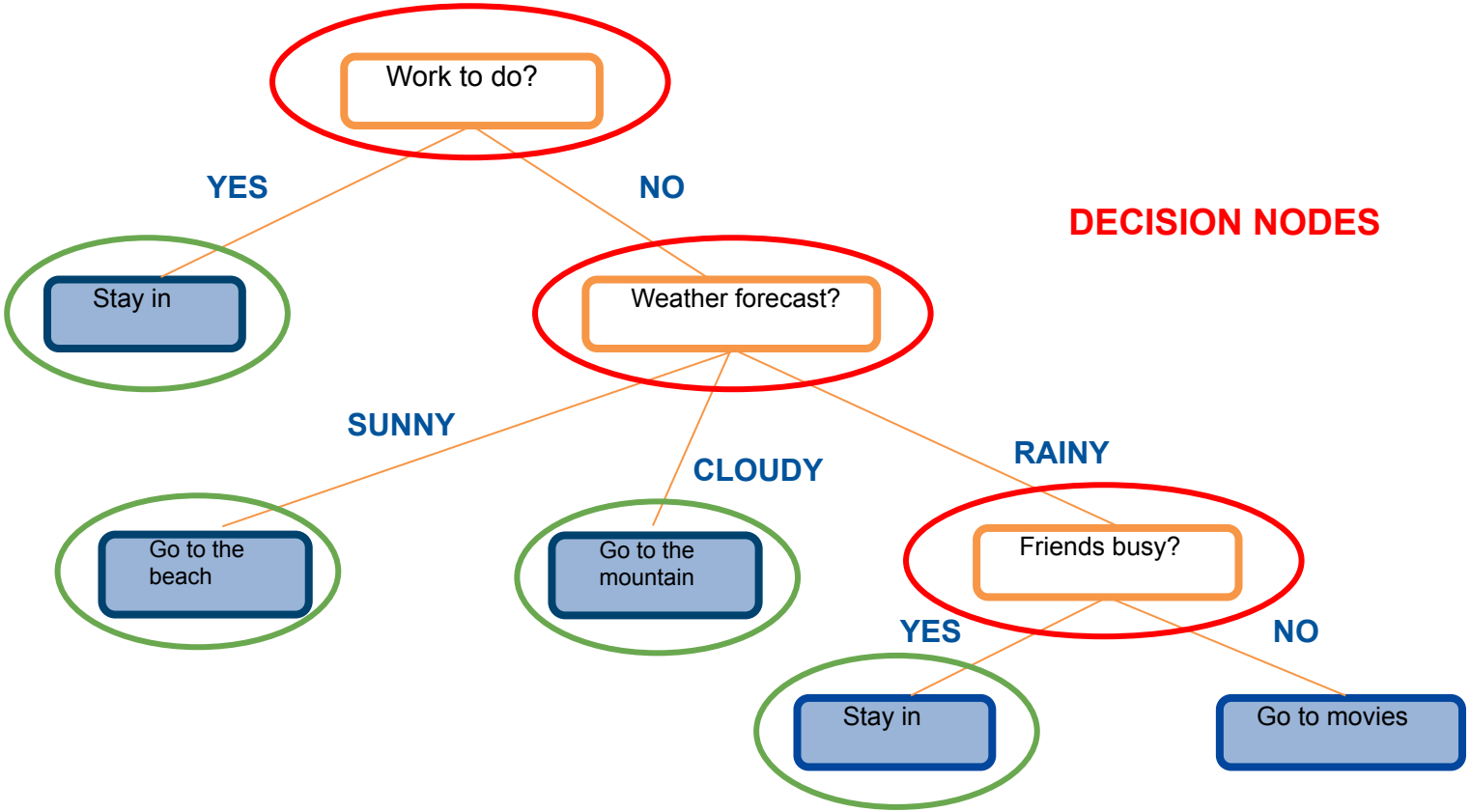
Decide a plan with friends



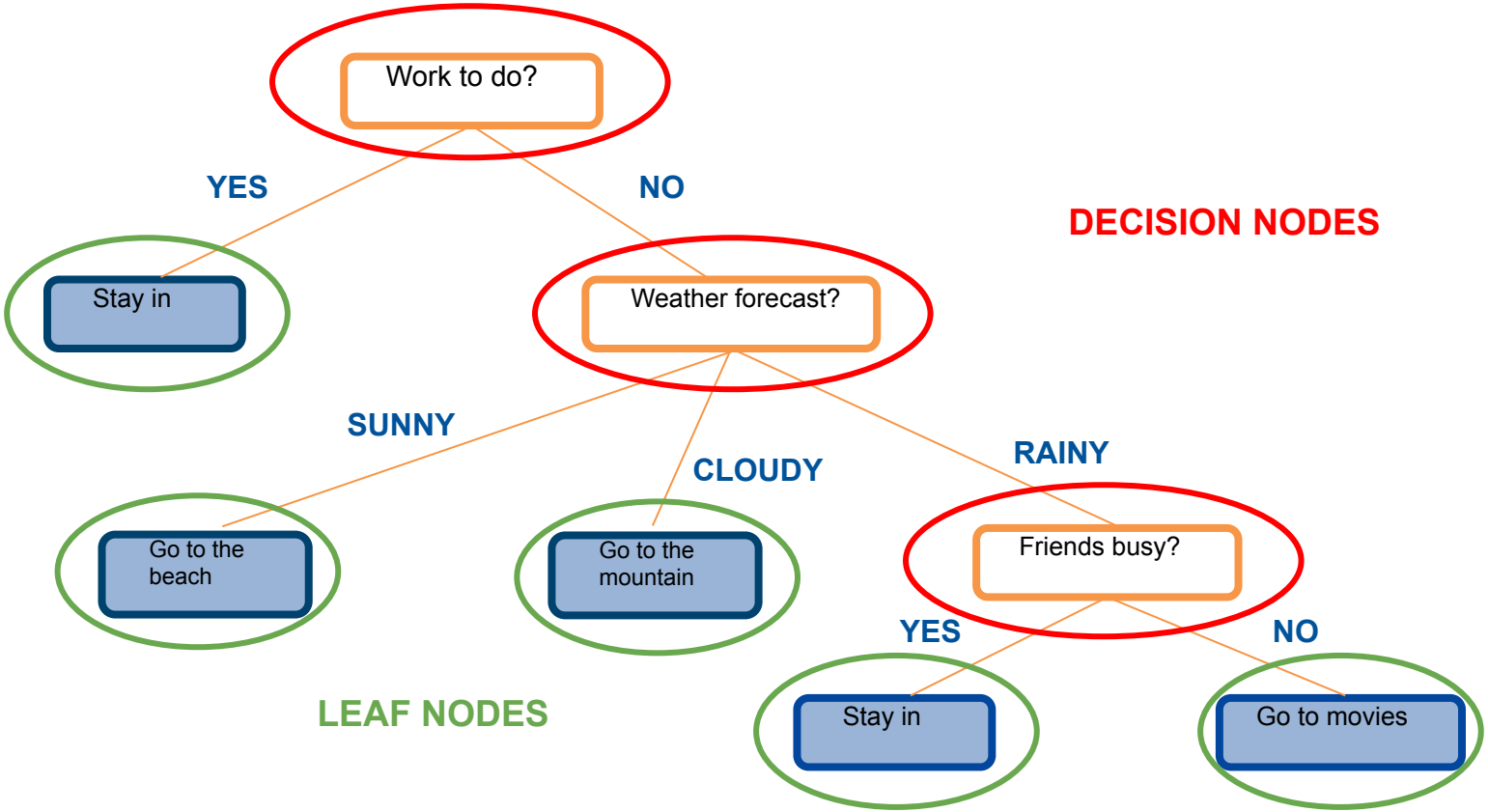
Decide a plan with friends



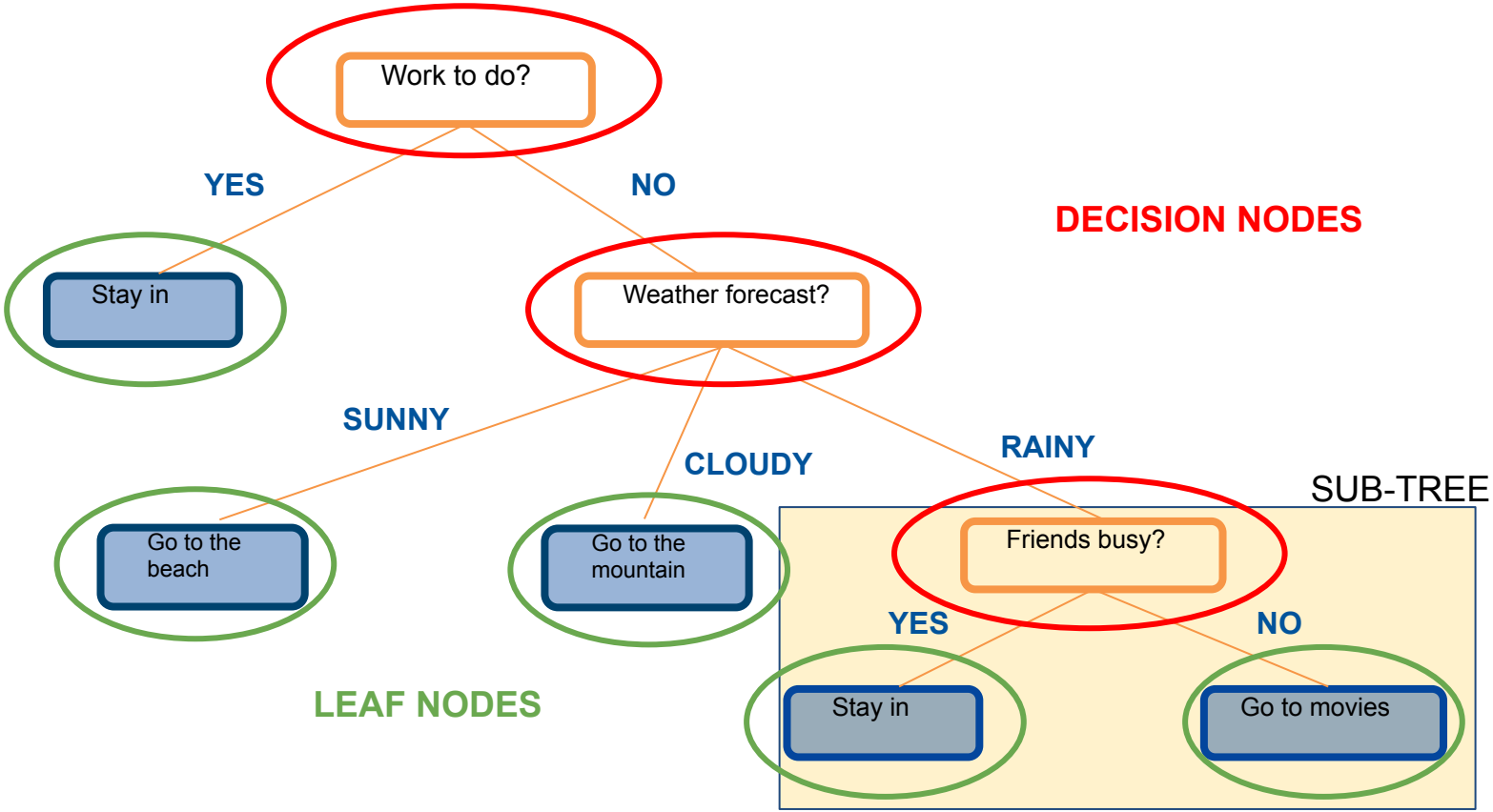
Decide a plan with friends



Decide a plan with friends

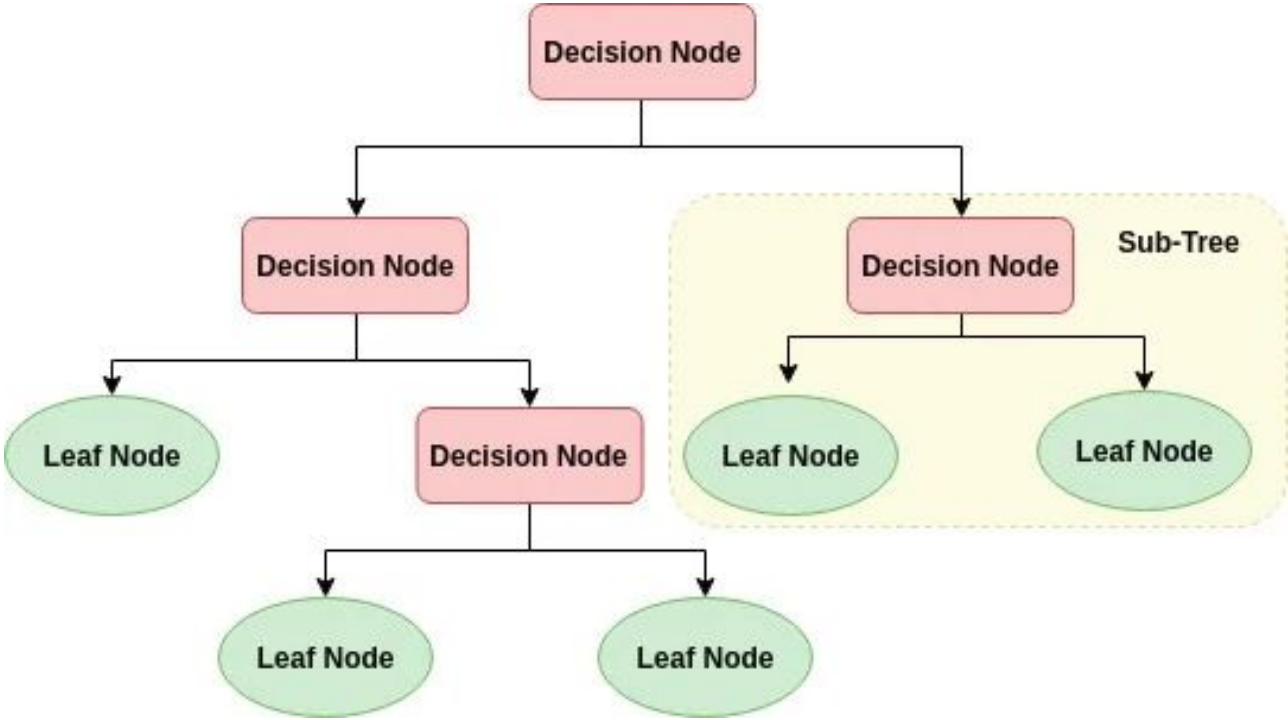


Decide a plan with friends

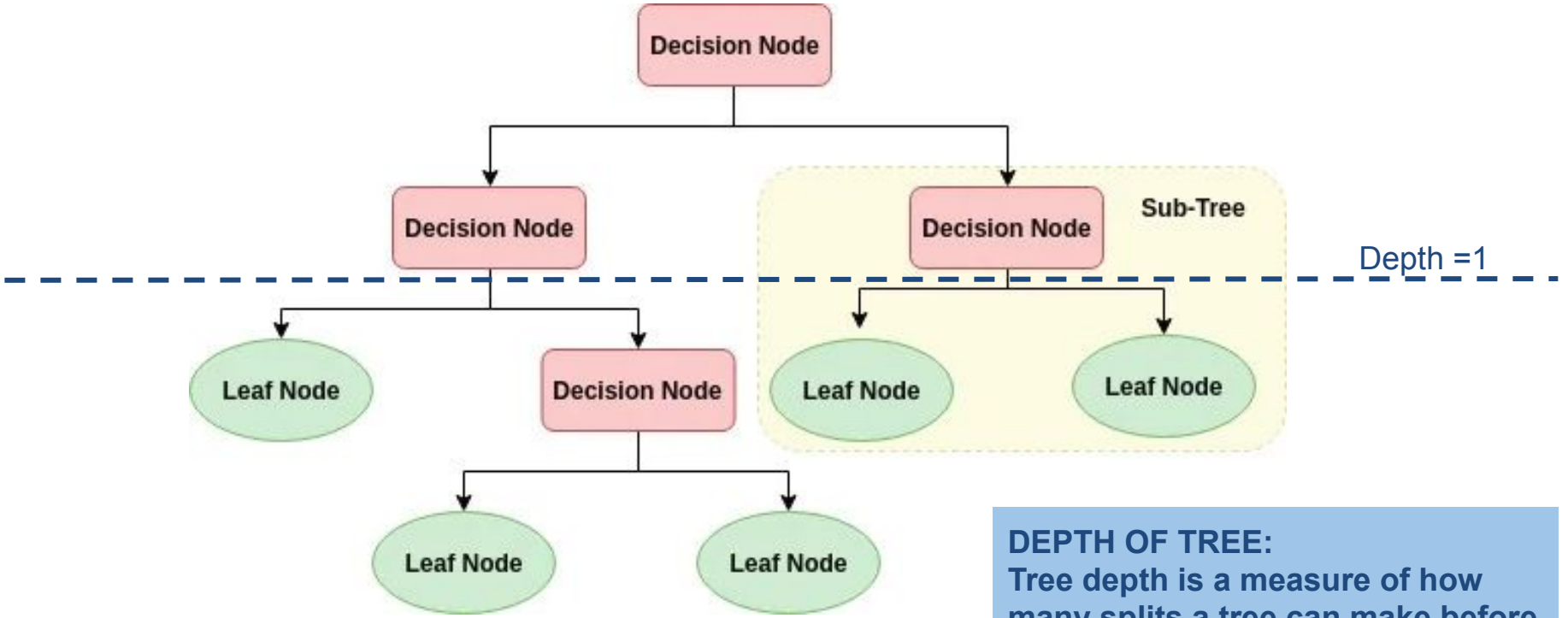


Decision Tree Algorithms

Saturdays.AI
Barcelona

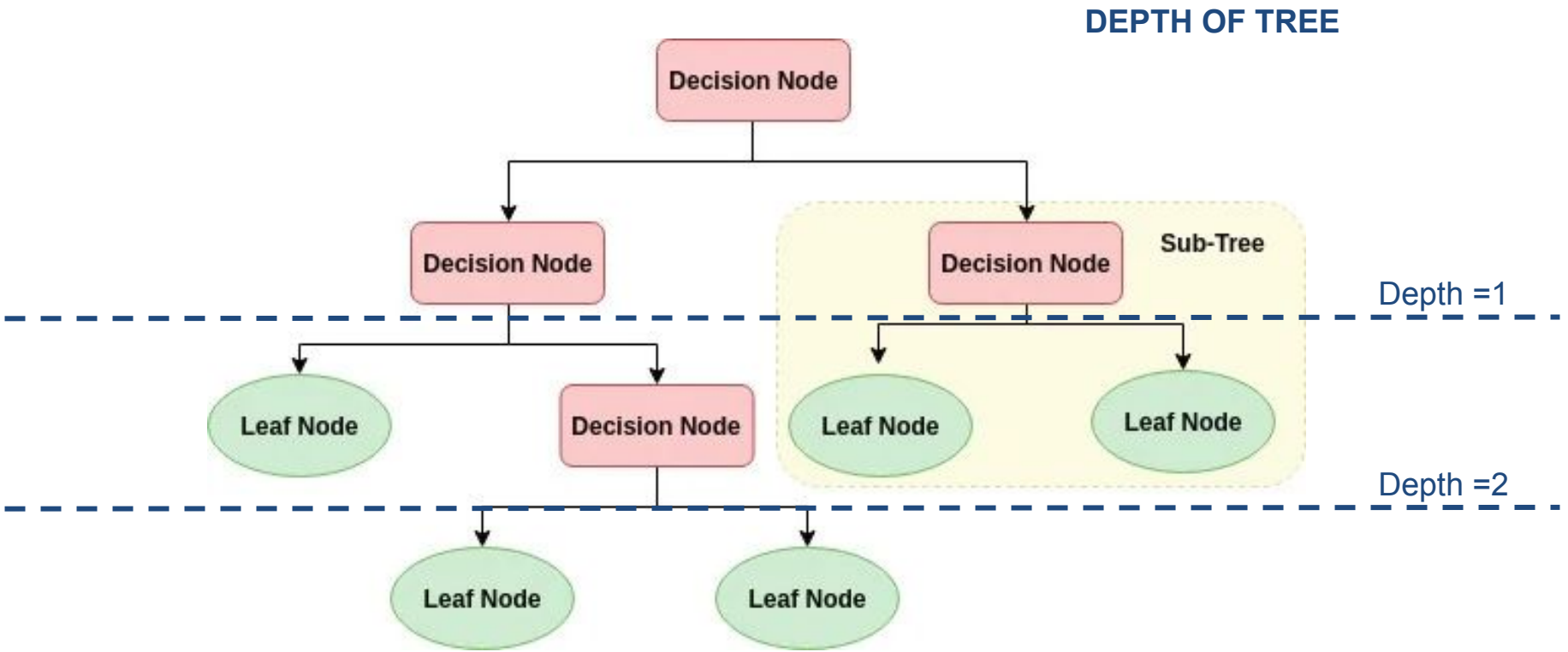


Decision Tree Algorithms

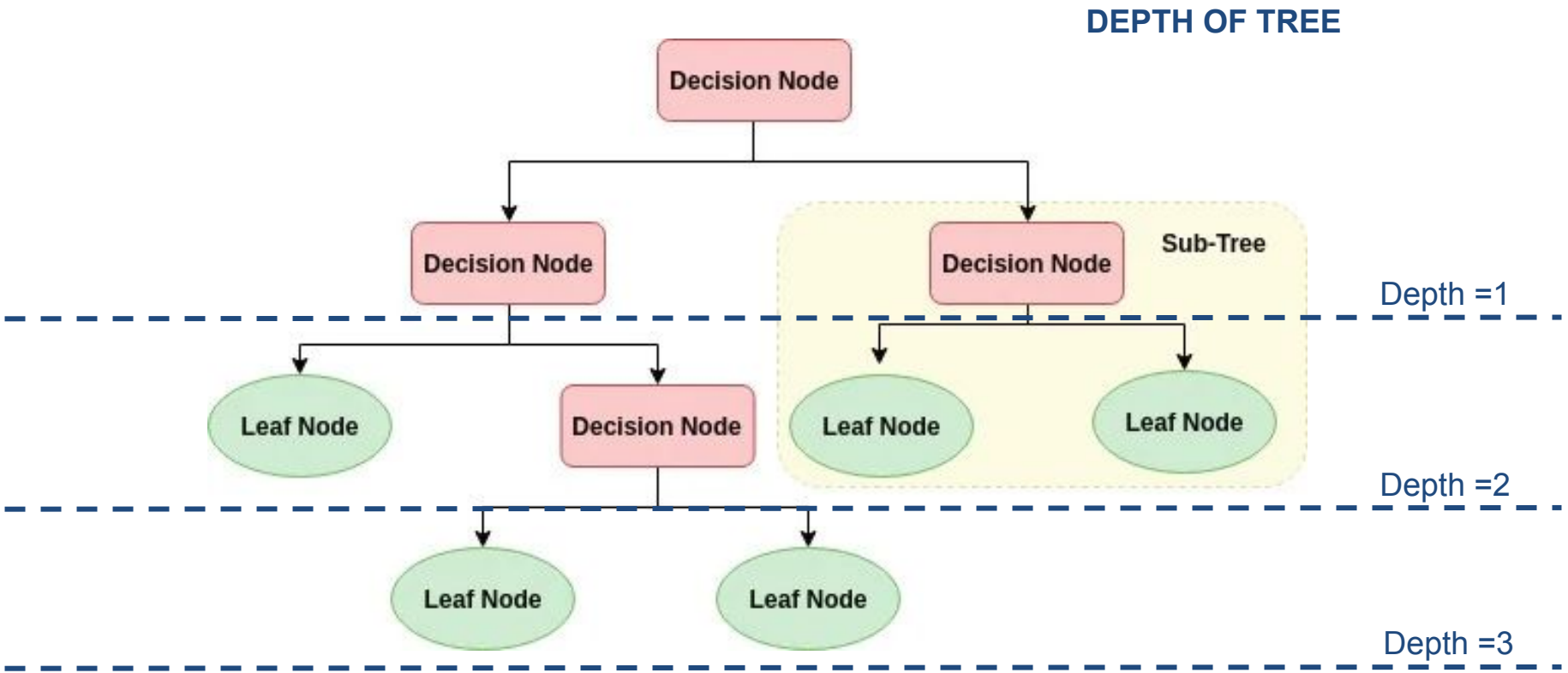


DEPTH OF TREE:
Tree depth is a measure of how many splits a tree can make before coming to a prediction.

Decision Tree Algorithms

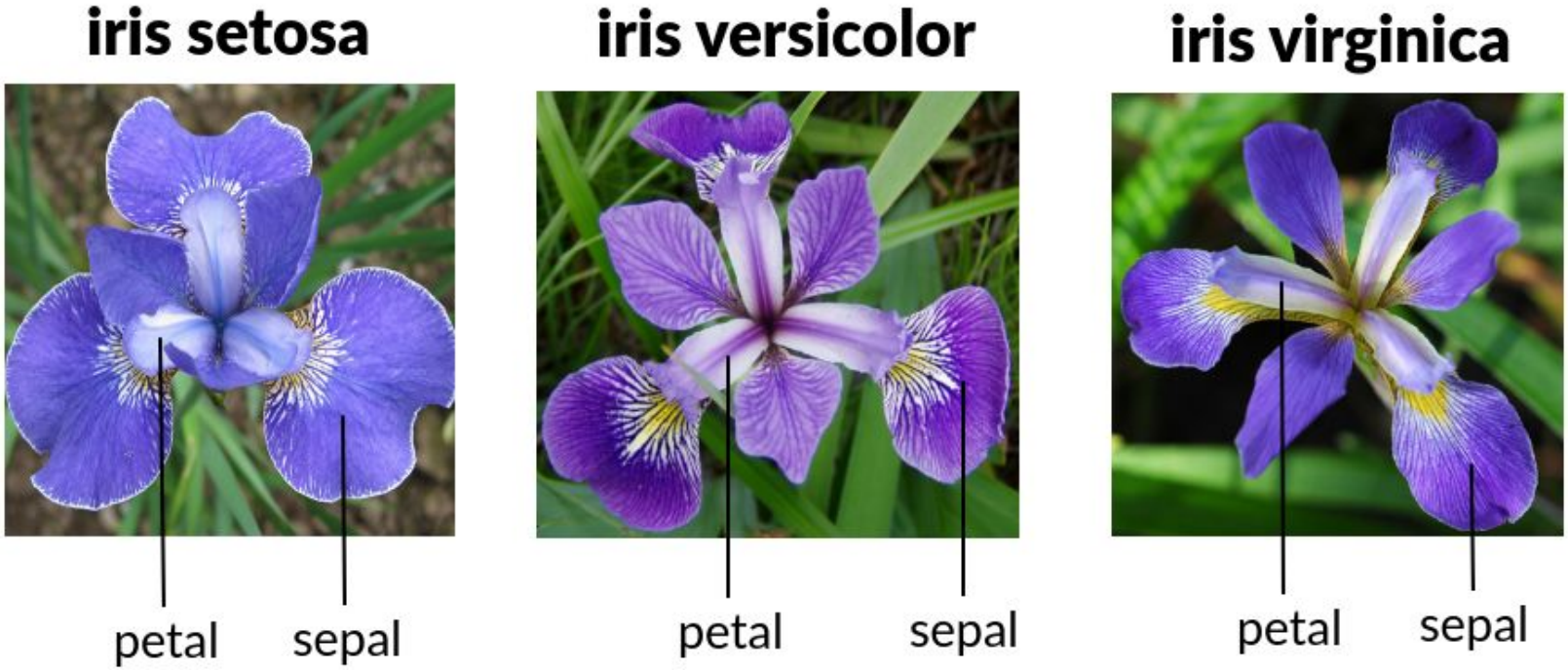


Decision Tree Algorithms

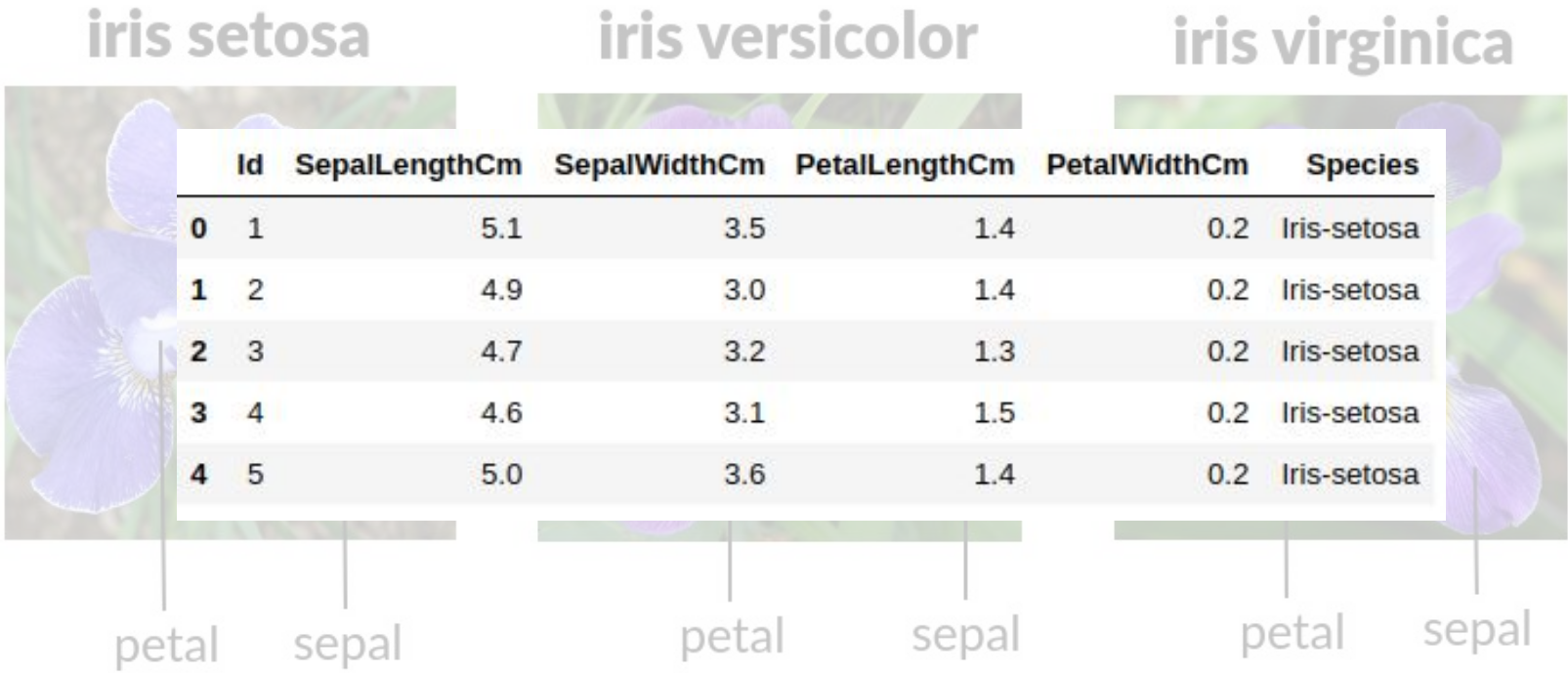


Decision Tree Algorithms

Saturdays.AI
Barcelona

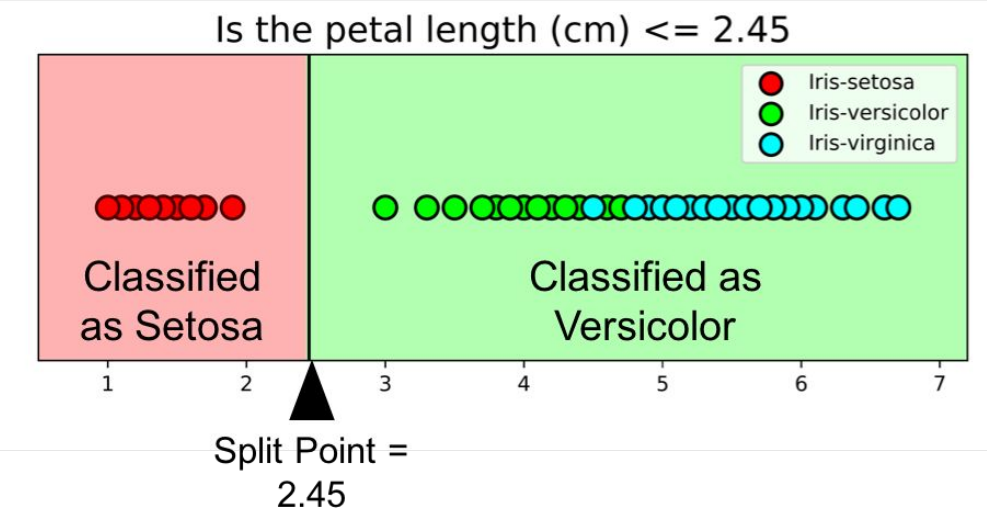


Decision Tree Algorithms



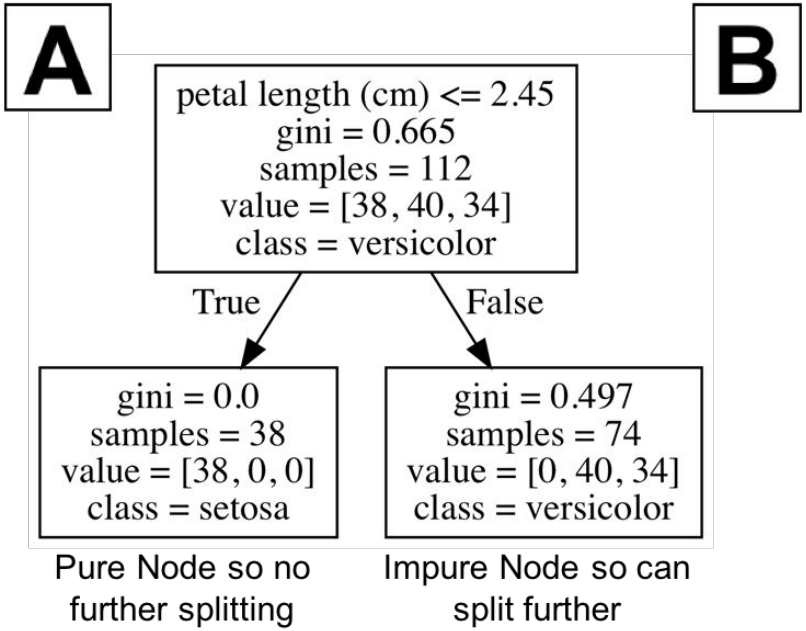
	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

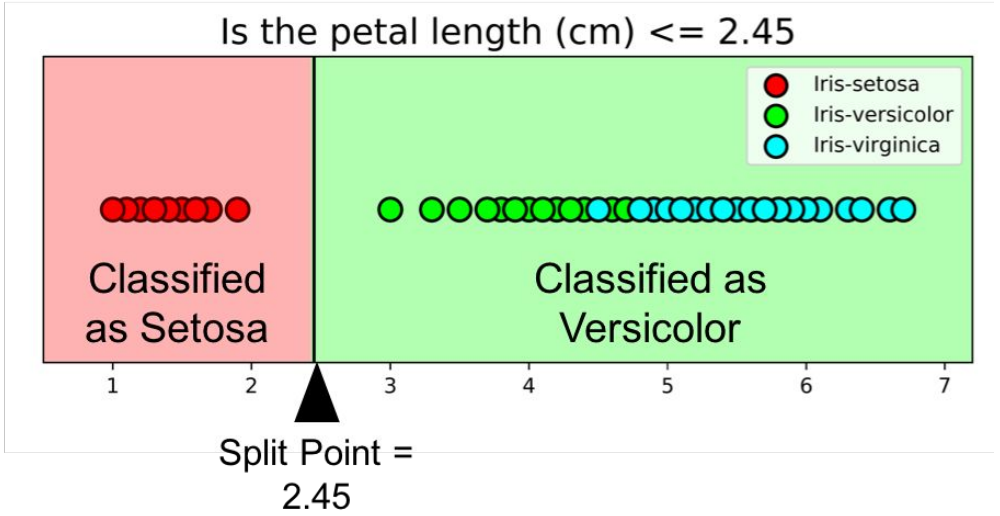
Decision Tree Algorithms



Decision Tree Algorithms

Saturdays.AI Barcelona





Decision Tree Algorithms

Saturdays.AI Barcelona

GINI IMPURITY:

The **Gini impurity** can be computed by summing the probability of an item with label being chosen times the probability of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

Further information about gini index: learnbymarketing.com/481/decision-tree-flavors-gini-info-gain

Decision Tree Algorithms

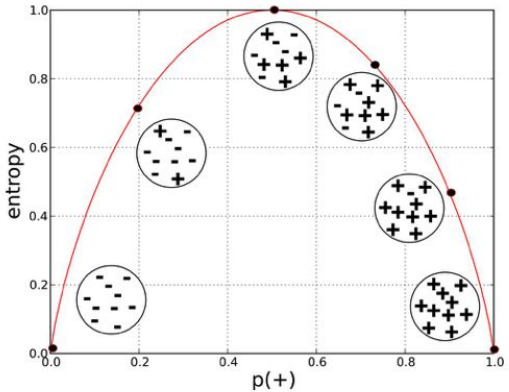
Saturdays.AI Barcelona

GINI IMPURITY:

The **Gini impurity** can be computed by summing the probability of an item with label being chosen times the probability of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

ENTROPY:

Entropy to calculate the homogeneity (or impurity) of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one.



Further information about gini index: learnbymarketing.com/481/decision-tree-flavors-gini-info-gain

Decision Tree Algorithms

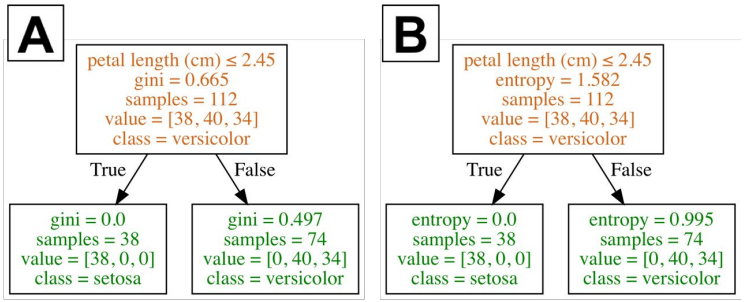
Saturdays.AI Barcelona

GINI IMPURITY:

The **Gini impurity** can be computed by summing the probability of an item with label being chosen times the probability of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

ENTROPY:

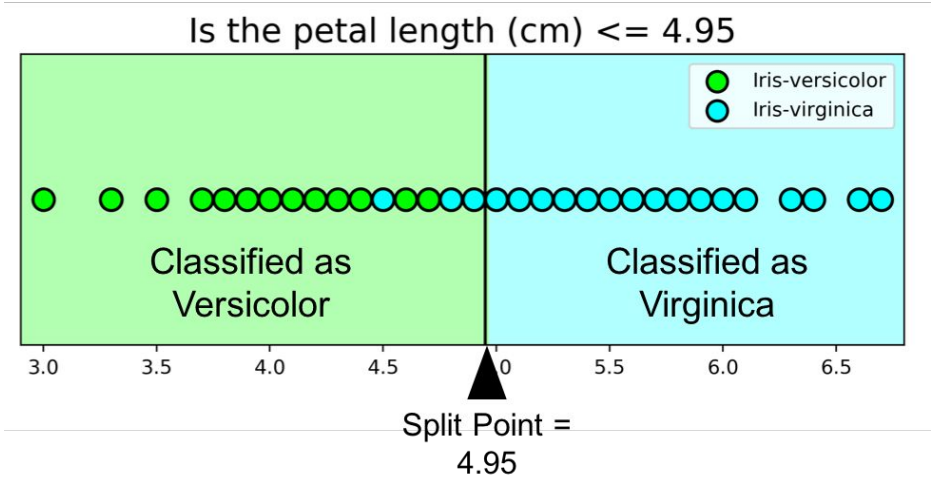
Entropy to calculate the homogeneity (or impurity) of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one.



criterion= 'gini' criterion= 'entropy'

Further information about gini index: learnbymarketing.com/481/decision-tree-flavors-gini-info-gain

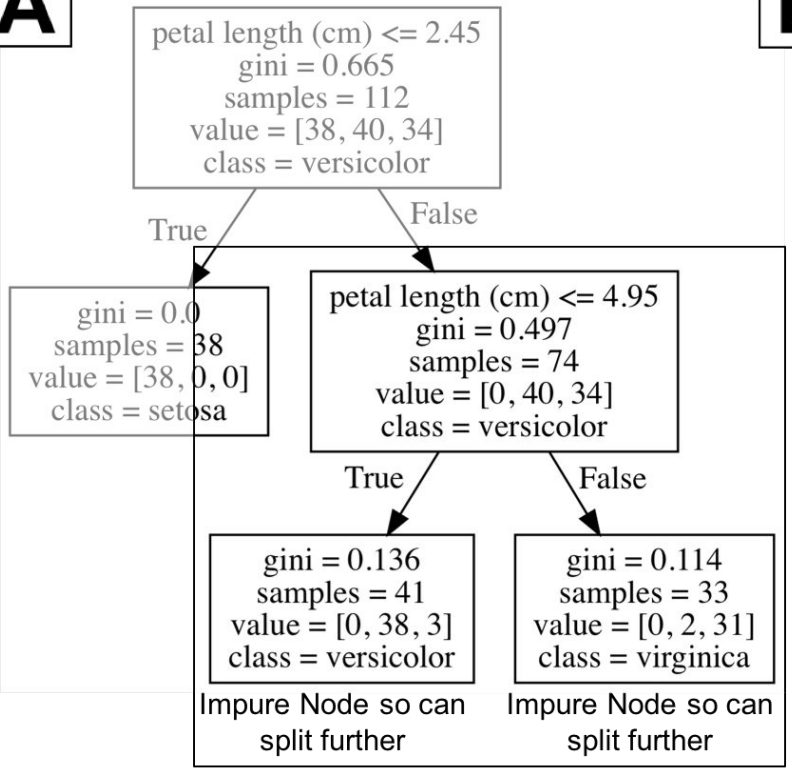
Decision Tree Algorithms



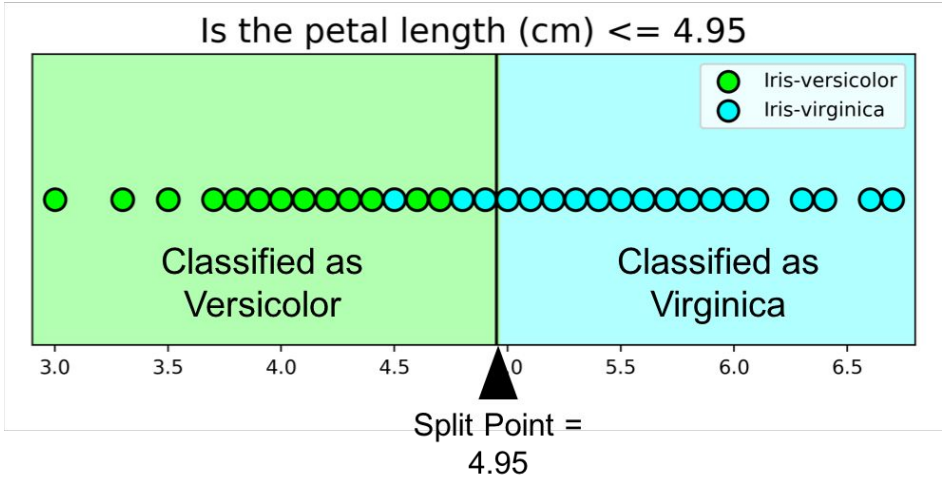
Decision Tree Algorithms

Saturdays.AI
Barcelona

A

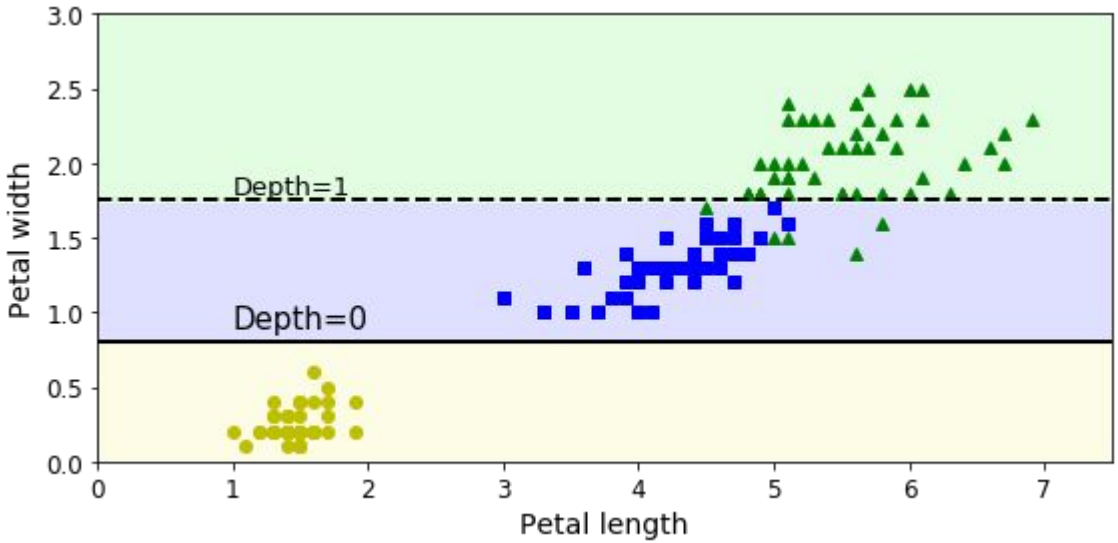


B

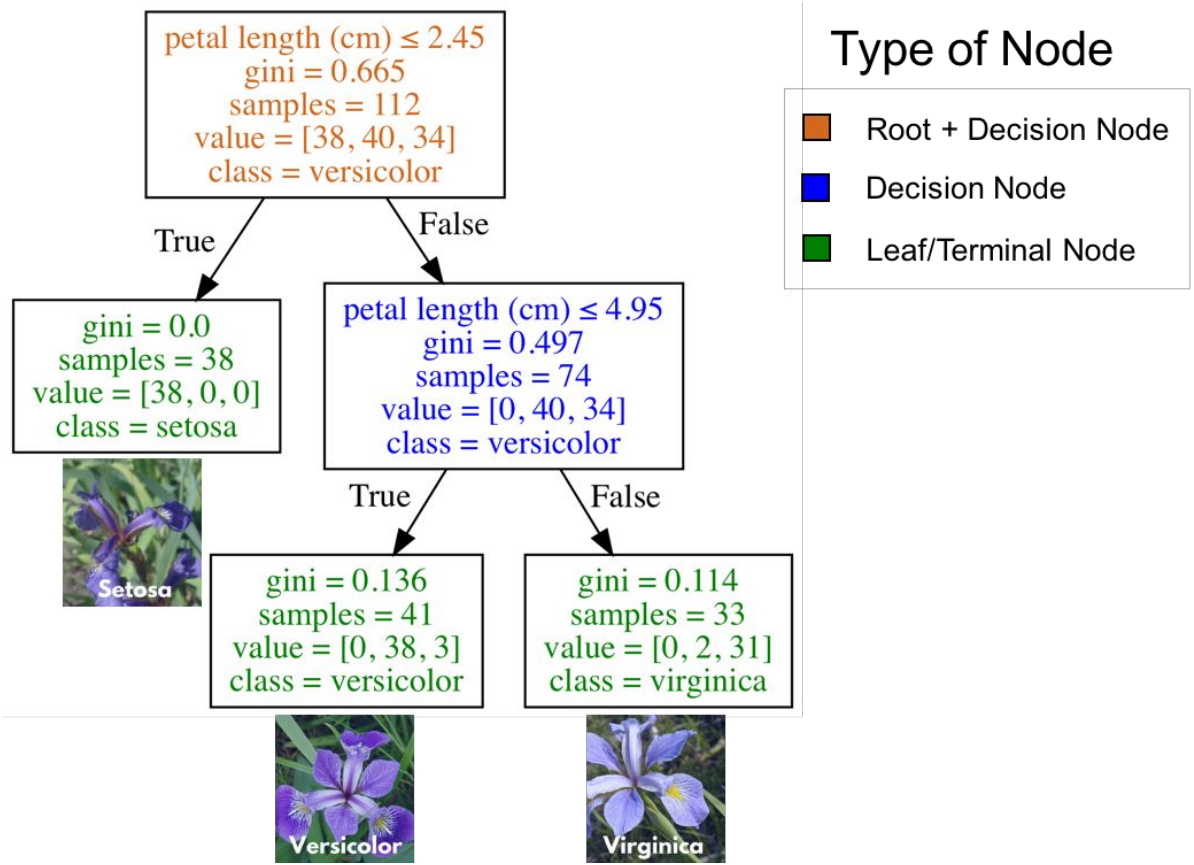


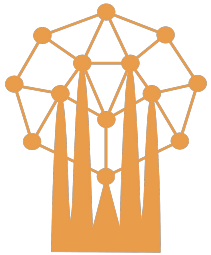
Decision Tree Algorithms

Saturdays.AI
Barcelona



Decision Tree Algorithms





Saturdays.AI
Barcelona

Random Forest

by Saturdays AI

Get ready for the future AI!

Random Forest...but first OVERFITTING

Saturdays.AI Barcelona

RANDOM FOREST:

The **random forest** is a model made up of many decision trees. Rather than just simply averaging the prediction of trees (which we could call a “forest”), this model uses **two key concepts** that gives it the name *random*:

1. **Random sampling of training data** points when building trees

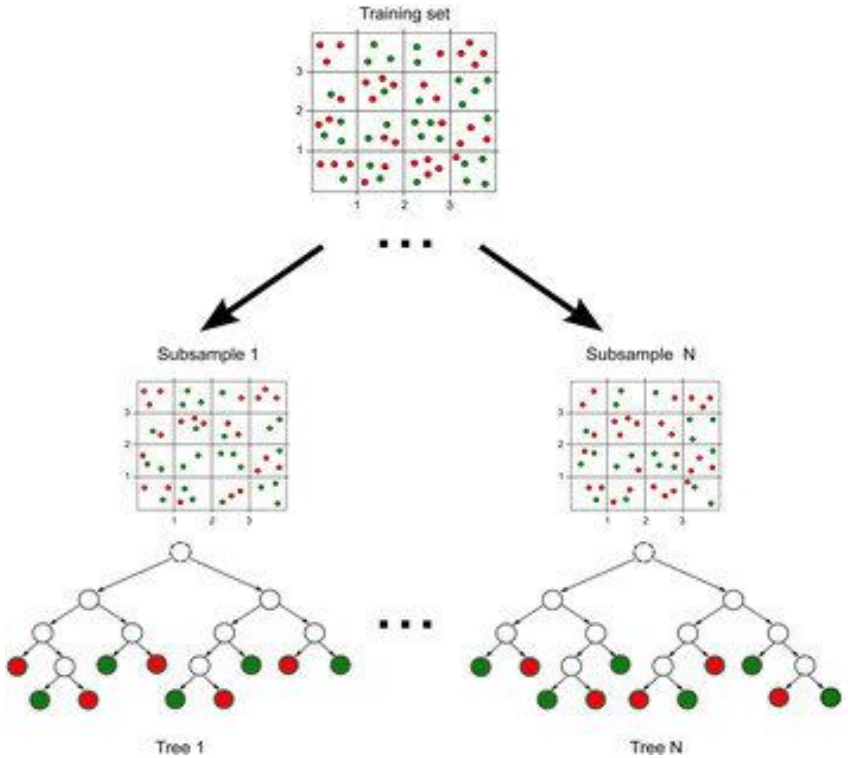
2. **Random subsets of features** considered when splitting nodes

Lectures: berkeley.edu - Random forest - Key Concepts

Random Forest: Random sampling of training data

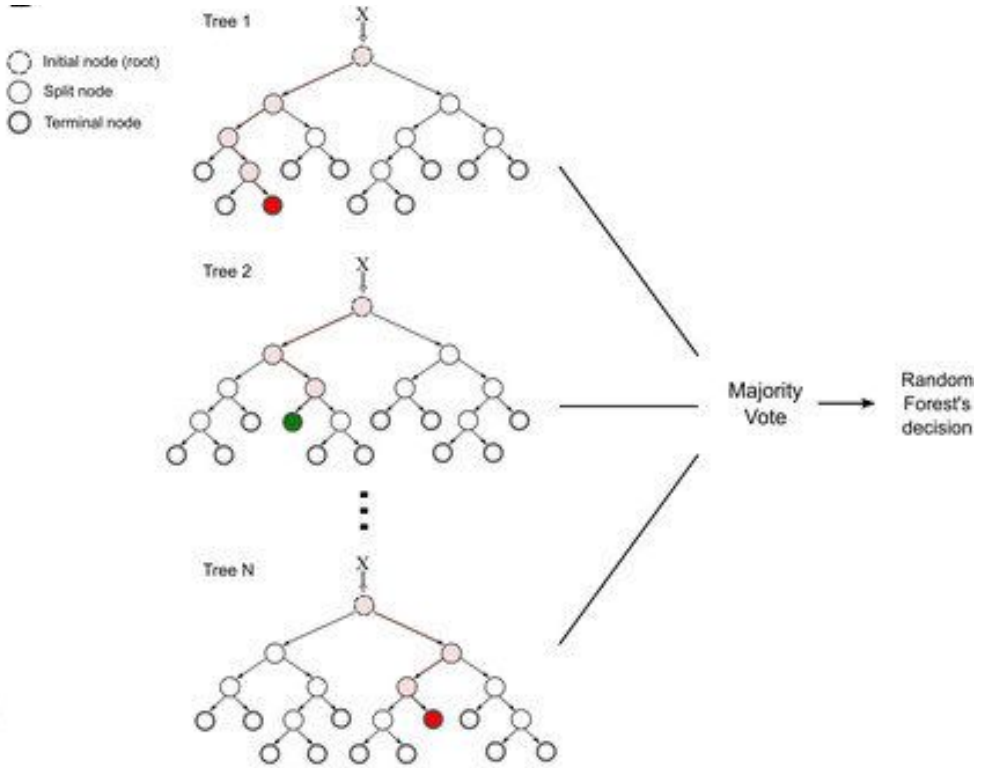
Saturdays.AI Barcelona

Each tree in a random forest learns from a random sample of the data points.



Random Forest: Random sampling of training data

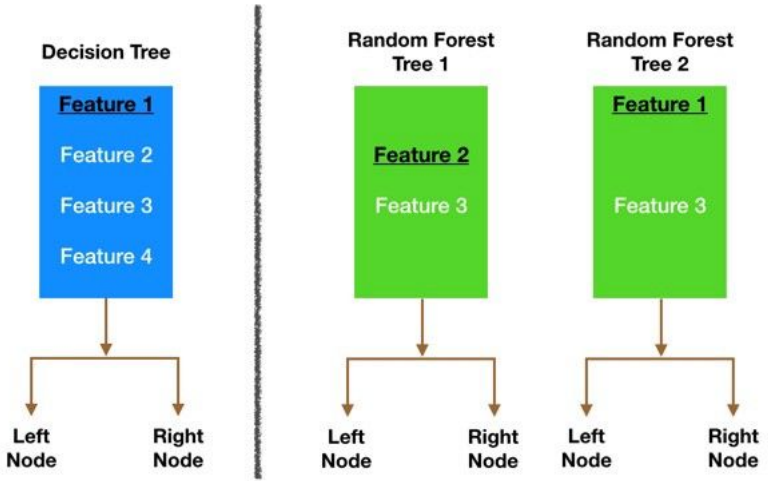
At test time, predictions are made by averaging the predictions of each decision tree.



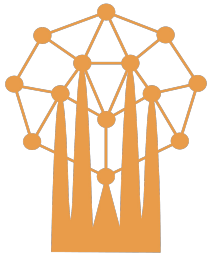
Random Forest: Random subsets of features

Saturdays.AI Barcelona

Only a subset of all the features are considered for splitting each node in each decision tree.
Generally this is set to sqrt(n_features) for classification meaning that if there are 16 features, at each node in each tree, only 4 random features will be considered for splitting the node.



Lectures: Does random forest select a subset of features for every tree or every node?



Saturdays.AI
Barcelona

Gradient Boosting Machines

by Saturdays AI

Get ready for the future AI!

Gradient Boosting Machines

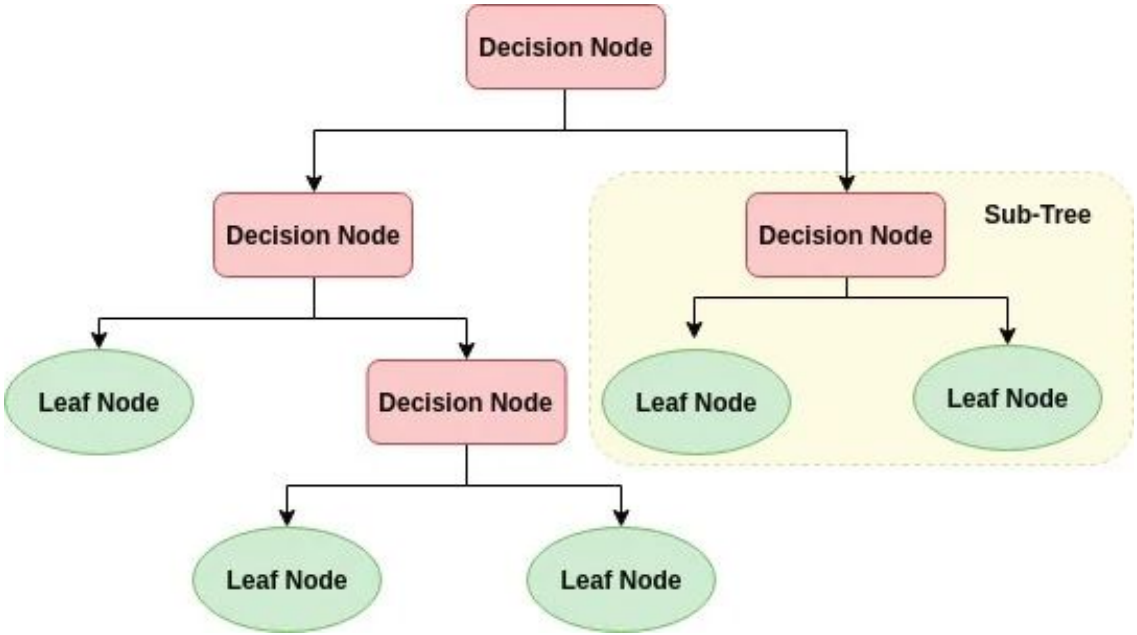
Like Random Forests but….

Not so random….

Gradient Boosting Machines

Saturdays.AI Barcelona

Train 1 tree as in Random Forest

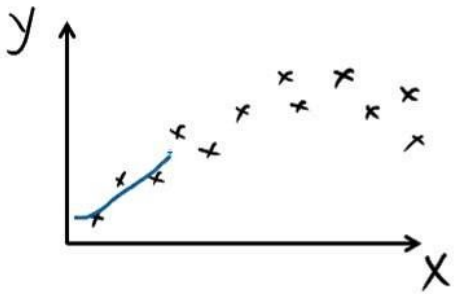


Gradient Boosting Machines

Saturdays.AI
Barcelona

Calculate **cost function**

- Describes how well the current response surface $h(\boldsymbol{x})$ fits the available data (on a given data set): $J(y_i, h(\boldsymbol{x}_i))$



- Smaller values of the cost function correspond to a better fit
- Machine learning goal: construct $h(\boldsymbol{x})$ such that J is minimized
- In regression, $h(\boldsymbol{x})$ is usually directly interpretable as predicted response

SALFORD SYSTEMS

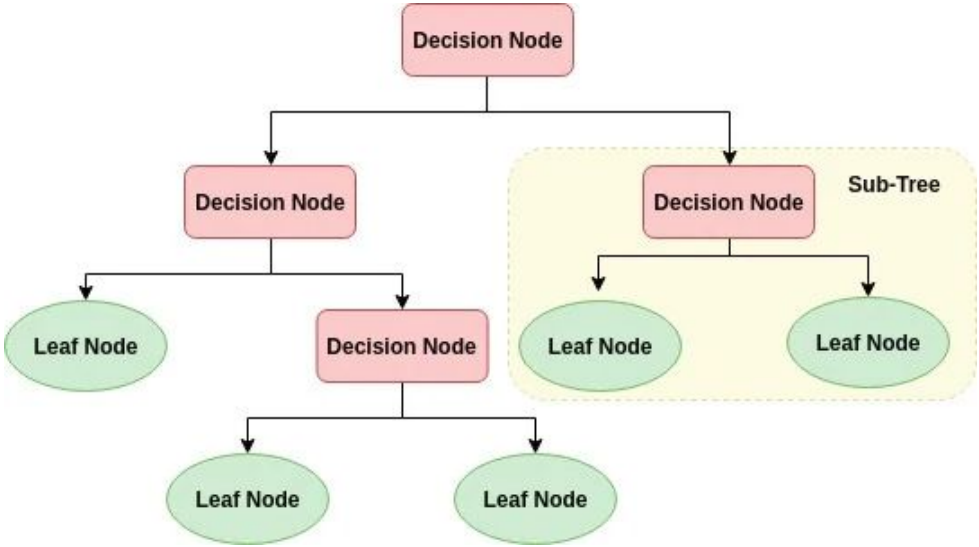
Gradient Boosting Machines

Saturdays.AI Barcelona

For which data points is the tree performing worst?

Give more importance to these data points when making the next tree

Train 2nd tree considering importance



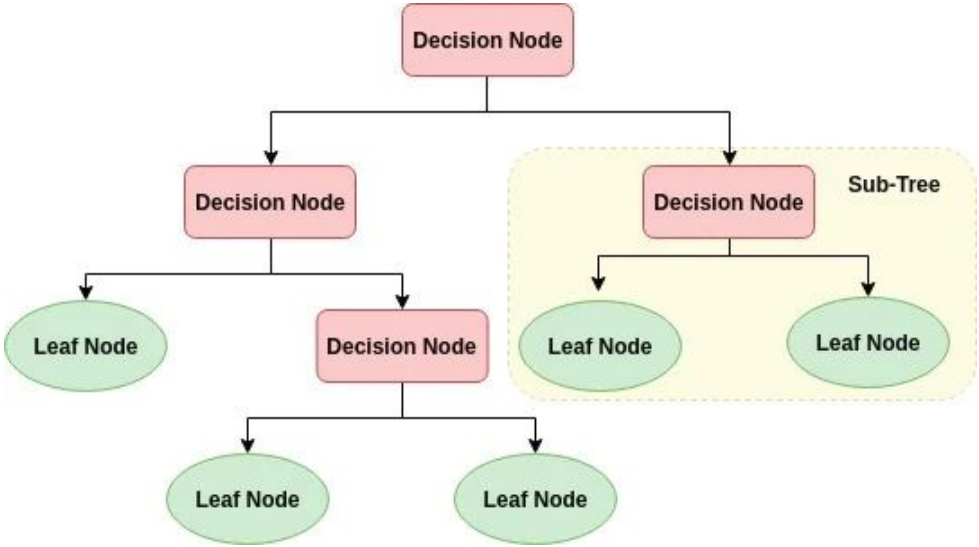
Gradient Boosting Machines

Saturdays.AI Barcelona

For which data points is the combination of the two trees performing worst?

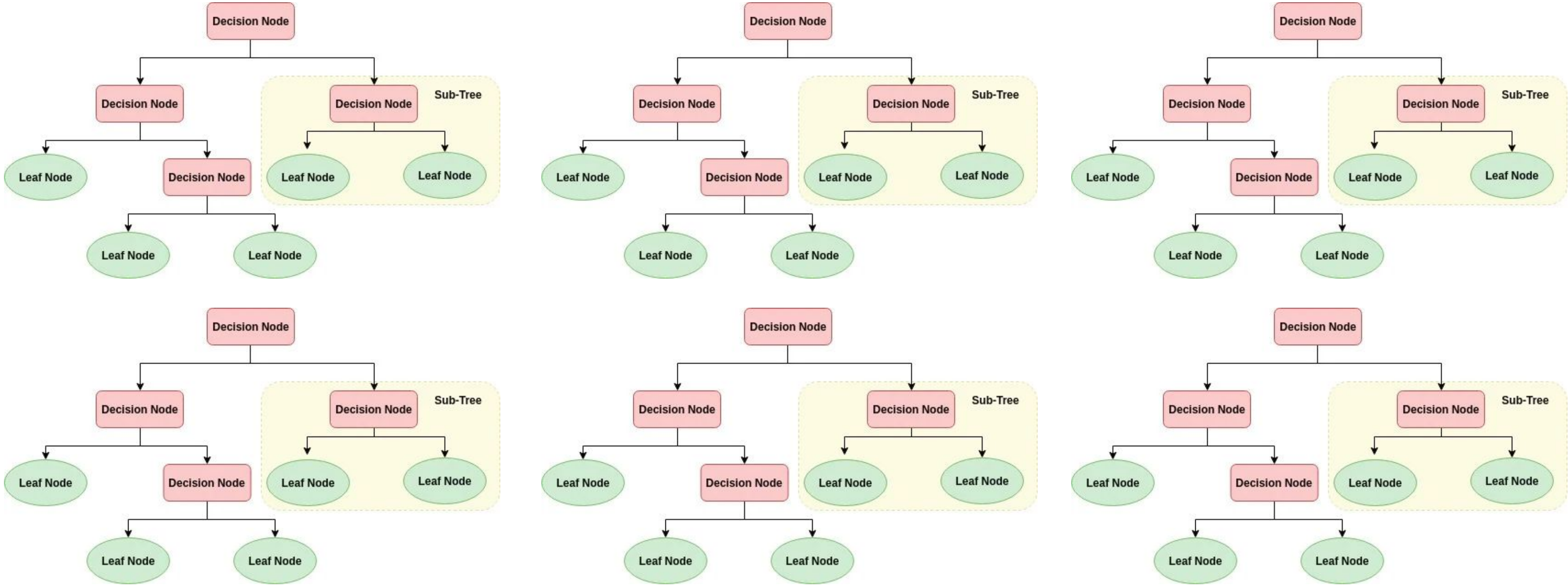
Give more importance to these data points when making the next tree

Train 3rd tree considering importance



Gradient Boosting Machines

Repeat with many more trees



Gradient Boosting Machines

Available algorithms:

XGBoost

CatBoost

AdaBoost

LightGBM

....

XGBoost

Saturdays.AI
Barcelona

The fastest kid in town

Usually better results than plain random forest

CatBoost

Saturdays.AI Barcelona

Last one to join the party

Great with categorical features

Great quality without parameter tuning

Reduce time spent on parameter tuning, because CatBoost provides great results with default parameters

Categorical features support

Improve your training results with CatBoost that allows you to use non-numeric factors, instead of having to pre-process your data or spend time and effort turning it to numbers.

Fast and scalable GPU version

Train your model on a fast implementation of gradient-boosting algorithm for GPU. Use a multi-card configuration for large datasets.

Fast prediction

Apply your trained model quickly and efficiently even to latency-critical tasks using CatBoost's model applier