A supervised learning algorithm to solve ML problems: decision tree

What is decision tree algorithm?

Decision tree is a type of supervised learning algorithm that can be used to solve classification and regression problems. The algorithm makes decisions to divide the data into different categories step by step.

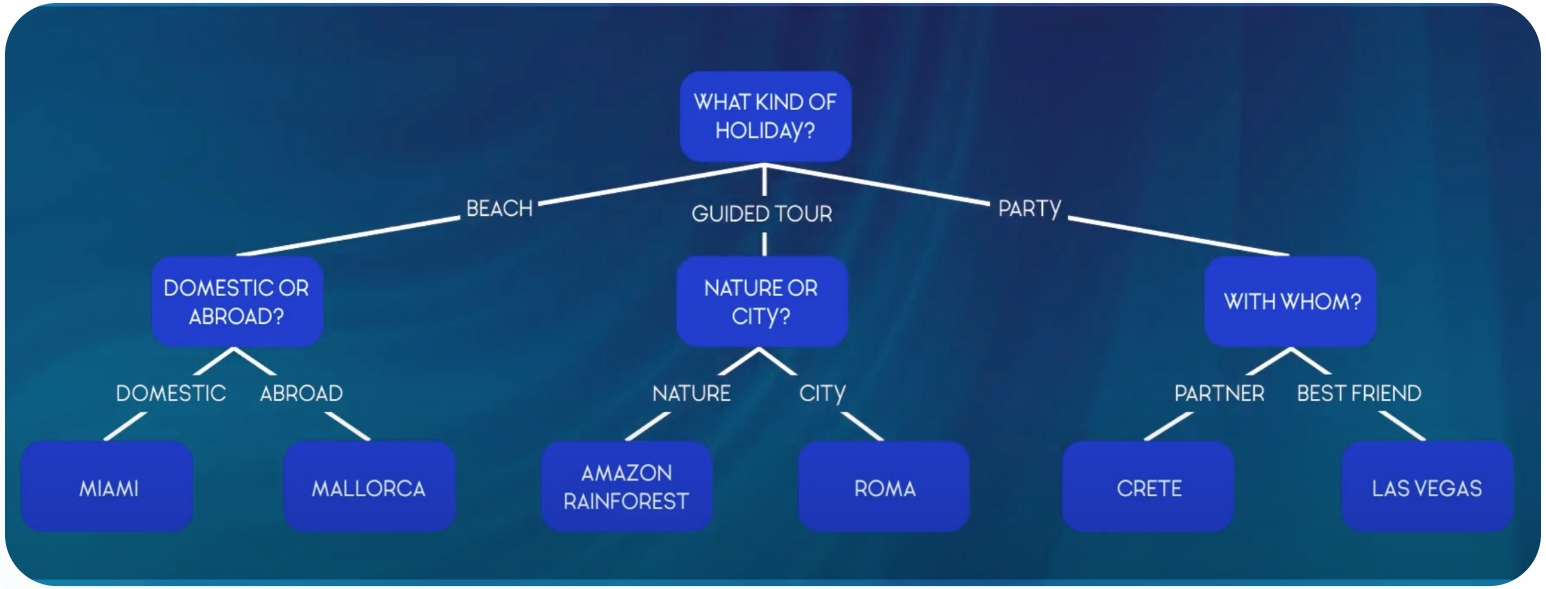
The intuition is the same as **making decisions in our daily lives**. E.g., when we have to figure out what to wear to work, where to go on holiday, etc.

How is the decision tree concept used in ML?

Decision tree algorithm **is a series of if-else statements** that can be used to **predict a result based on data**.

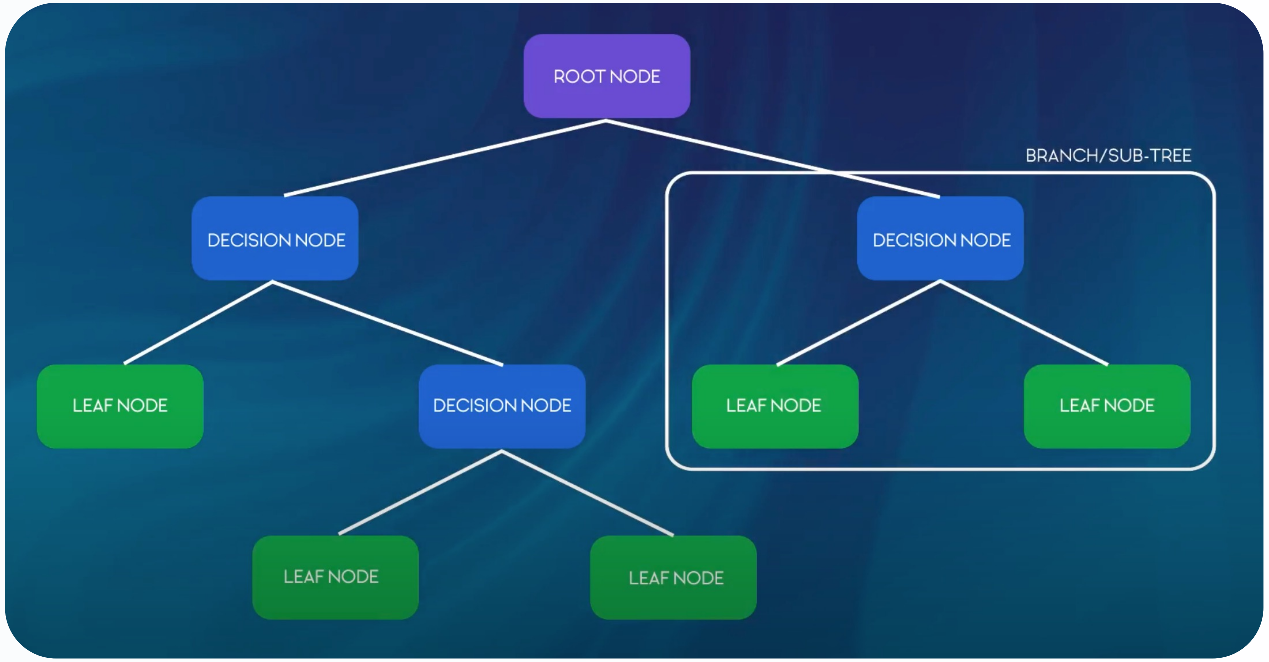
Simply put, it…

- …randomly asks questions,
- …checks if the dataset is divided correctly,
- …changes the questions accordingly.



A decision tree is also described as a series of so-called ‘nodes’

- **Root or top node -** the starting point of the tree
- Then we have **a set of decision nodes** until we reach **the leaf node**
- **Leaf node -** the final category or prediction/result



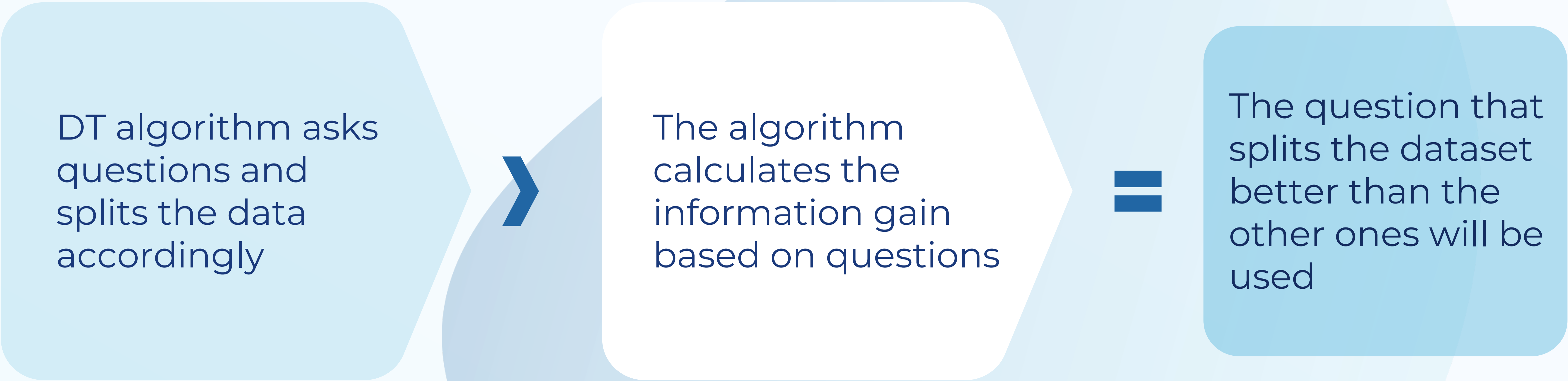
Where to use decision tree algorithms

Decision tree algorithms can be used for solving:

- **Classification problems:** e.g., predicting whether a computer’s price is low, medium, or high
- **Regression problems:** e.g., predicting the selling price of a house.

▶ Generally, decision trees are more suitable for **classification problems** because the logic of decision based algorithms is **categorical**. Predicting the output of continuous values, which is the case in regression problems, is **harder based on decisions**.

How to use decision tree algorithms



DT algorithm asks questions and splits the data accordingly

The algorithm calculates the information gain based on questions

The question that splits the dataset better than the other ones will be used

The decision tree algorithm has some advantages & limitations

Advantages

- Takes less effort to prepare data
- No need to normalize and scale it
- Missing values in the data don’t significantly affect the process of creating a decision tree

Disadvantages

- Takes longer to train
- Unstable - slight changes in the data can result in an entirely different tree being constructed
- Tends to overfit for complicated data

To make the decision tree algorithm less complex: pruning

What is pruning?

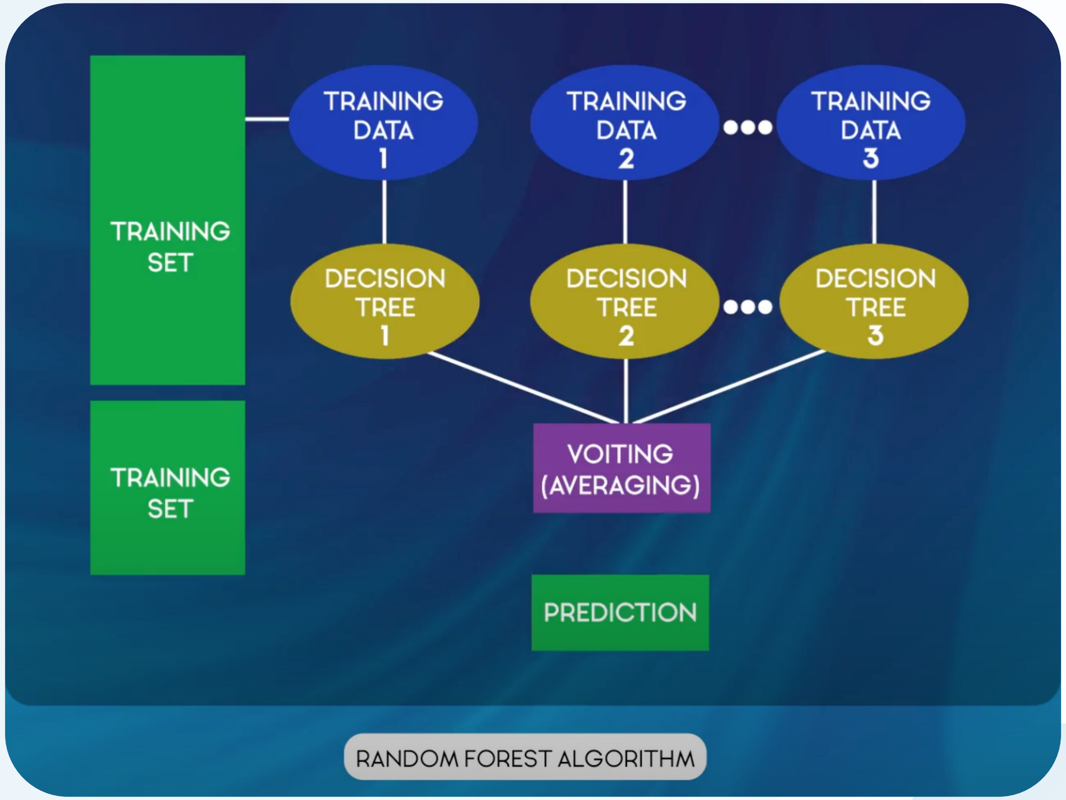
Pruning is removing some unnecessary features based on the information gain in order to get less complex decision tree. Pruning helps to **prevent overfitting** and **high variance** so that the model **generalizes well to unseen data**.

Random forests - an improved version of decision trees

AI BUSINESS SCHOOL

How does random forest algorithm work?

The random forest algorithm uses many trees instead of using only one. It runs a test dataset on each tree, gets predictions from each and based on the majority votes, that means the category that appears the most, we get a final output.



RANDOM FOREST ALGORITHM