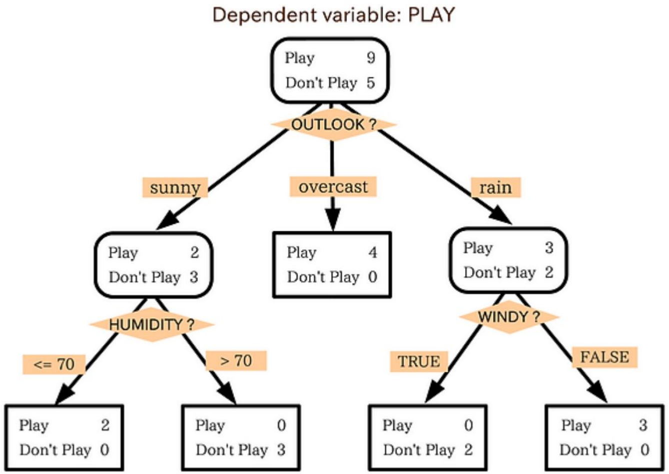
Decision tree and K-nearest neighbors (KNN) algorithm

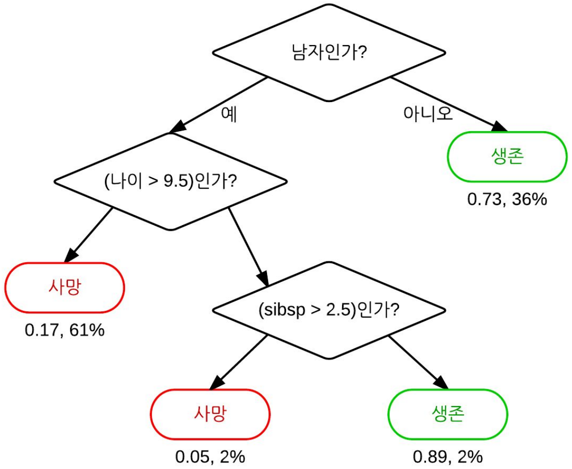
Decision tree

What is the decision tree?

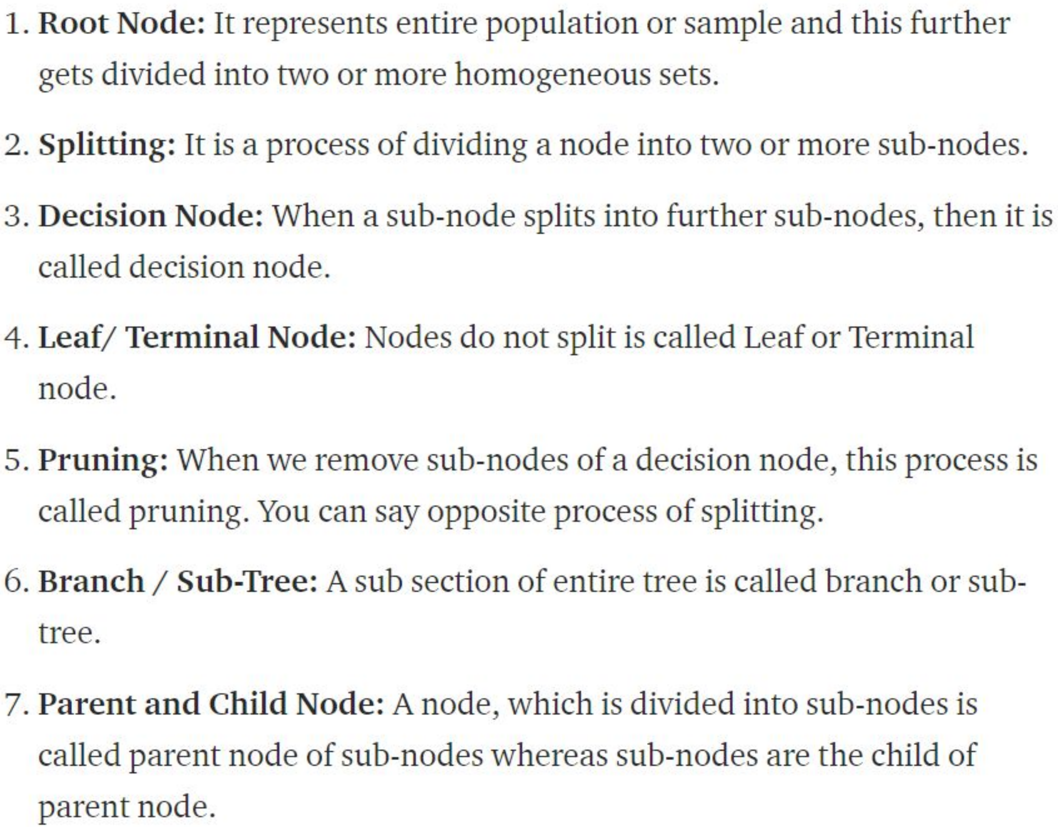
Decision tree: visually and explicitly represent decisions and decision making.

task: classification and regression

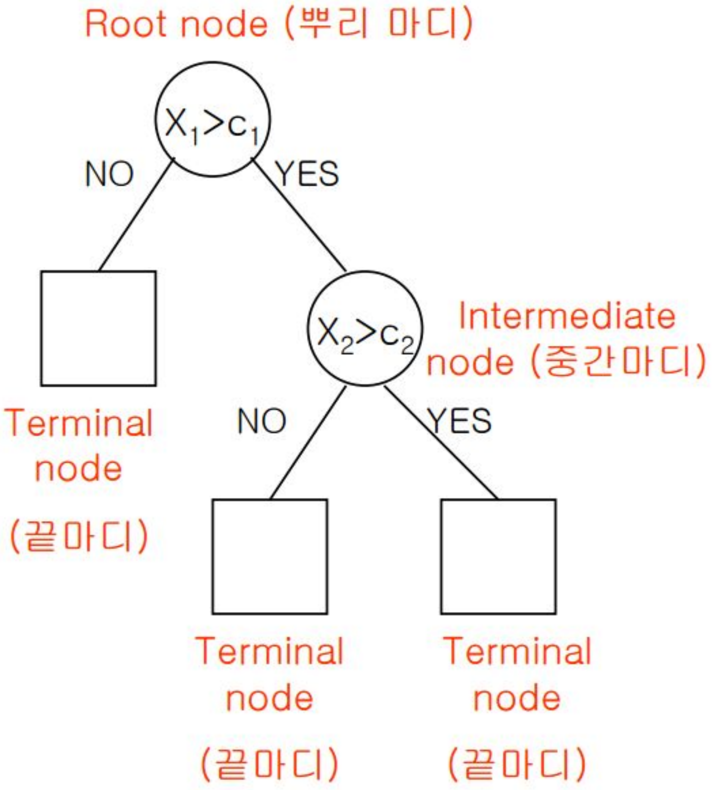




Common term



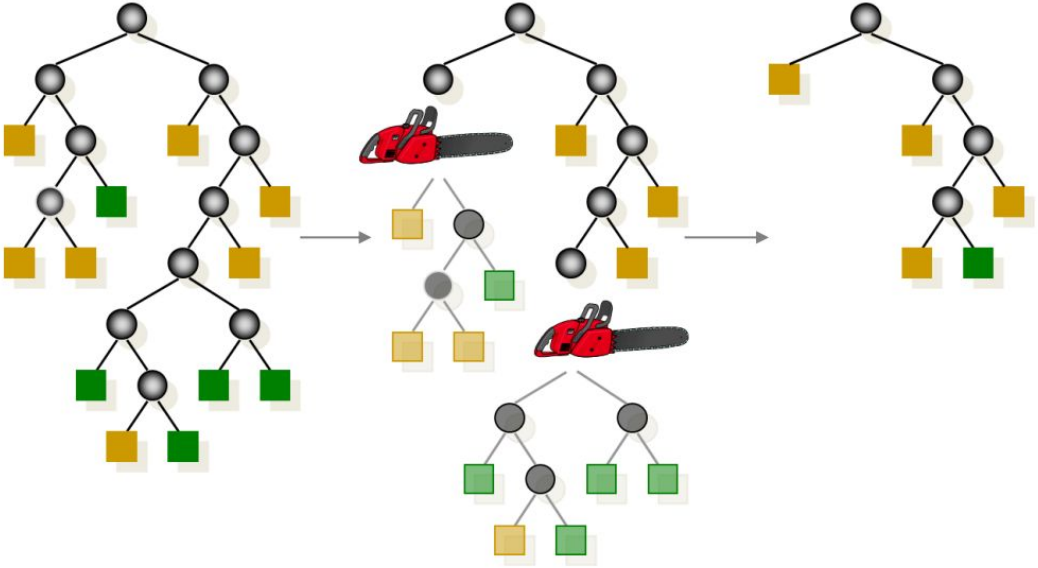
1. **Root Node:** It represents entire population or sample and this further gets divided into two or more homogeneous sets.
2. **Splitting:** It is a process of dividing a node into two or more sub-nodes.
3. **Decision Node:** When a sub-node splits into further sub-nodes, then it is called decision node.
4. **Leaf/ Terminal Node:** Nodes do not split is called Leaf or Terminal node.
5. **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.
6. **Branch / Sub-Tree:** A sub section of entire tree is called branch or sub-tree.
7. **Parent and Child Node:** A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node.



How to avoid overfitting

Pruning is a technique in machine learning and search algorithms that reduces the size of **decision trees** by removing sections of the **tree** that provide little power to classify instances.

Pruning: Removing side issue intermedia node

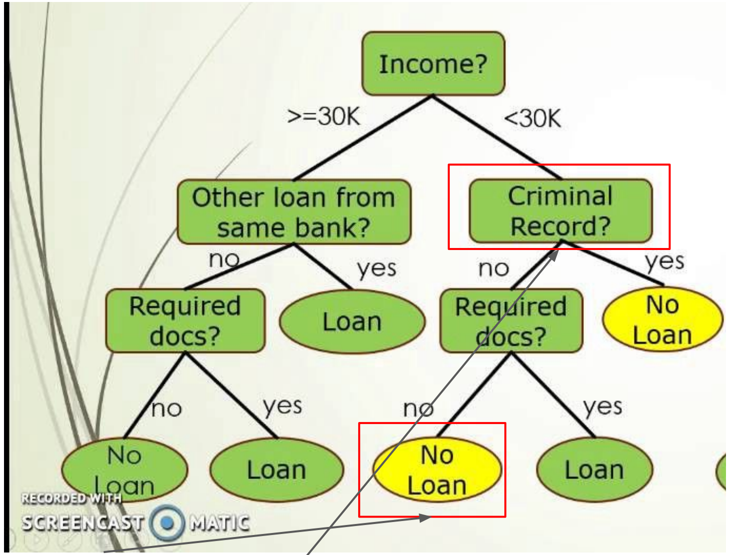


Detail

Splitting (training step)

1. Calculate cost each intermediate node using cost function (such as gini impurity and cross entropy)
2. Select low cost terminal node using greedy algorithm
3. go to 1 (recursive algorithm)
    a. Set a minimum number of training inputs to use on each leaf. (terminate condition)

Pruning (tree performance optimization)

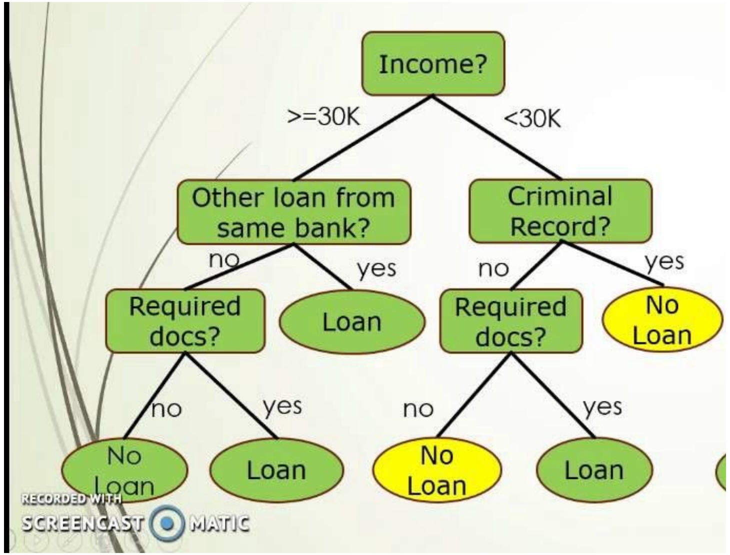


Selected leaf node

Most popular intermediate node (need to include selected leaf node)

Prunning

After remove most popular intermediate node, calculate classification error



If maintained accuracy, remove node

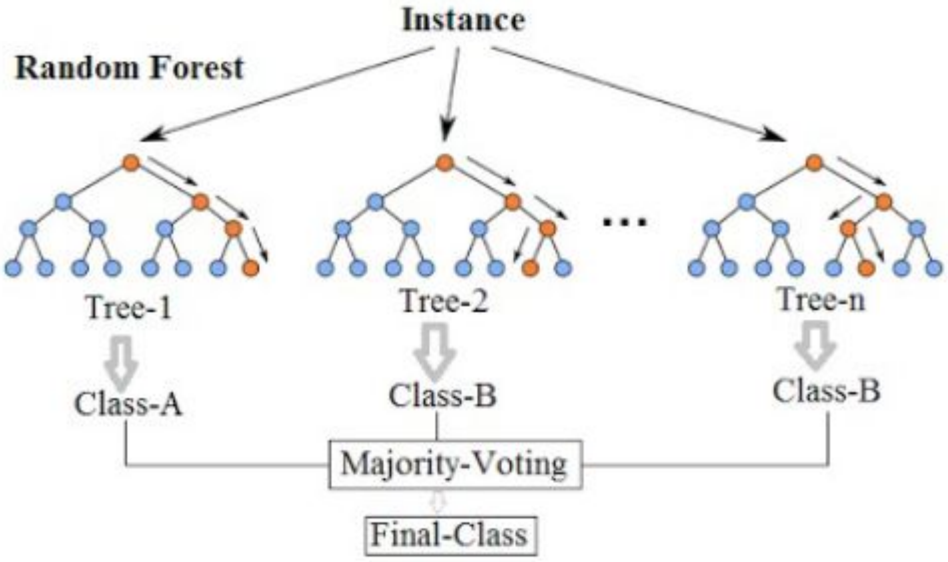
Advantage and disadvantage

- Advantage
  - Simple to understand, interpret, visualize.
  - Can *handle both numerical and categorical data*. Can also *handle multi-output problems*.
  - Decision trees require relatively little effort from users for data preparation.
- Disadvantage
  - **Decision-tree learners can create over-complex trees that do not generalize the data well.**
  - **Decision trees can be unstable because small variations in the data might result in a completely different tree being generated.**
  - Greedy algorithms cannot guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees, where the features and samples are randomly sampled with replacement.
  - Decision tree learners create *biased trees if some classes dominate*. It is therefore recommended to balance the data set prior to fitting with the decision tree.

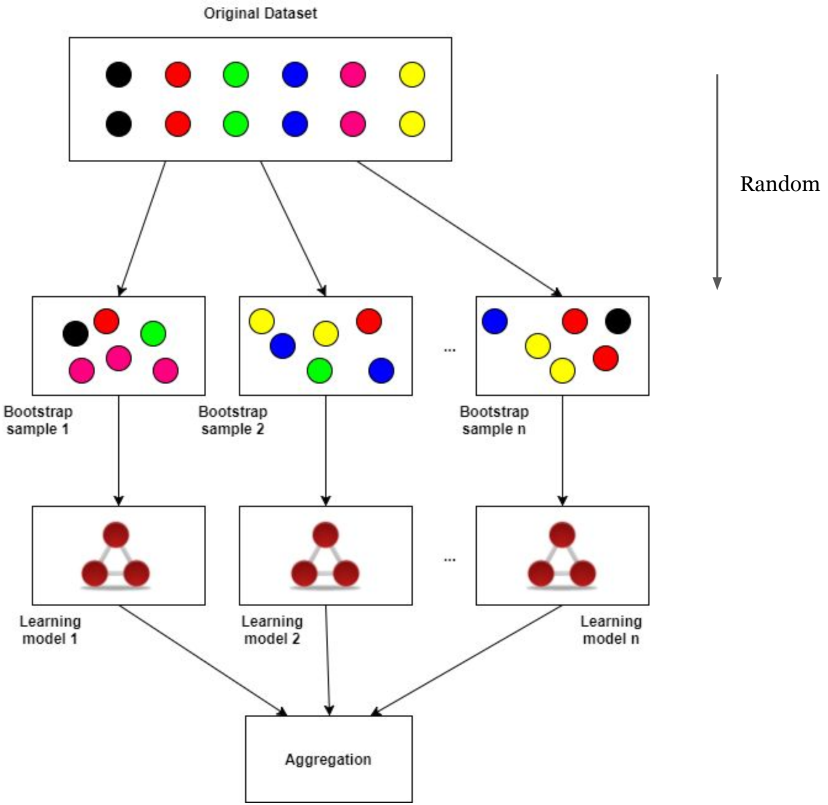
Random forest

What is the random forest?

- large number of **individual** decision trees that operate as an ensemble.
    - Construct random trees using bootstrap samples
    - For this reason, overfitting can be prevented.



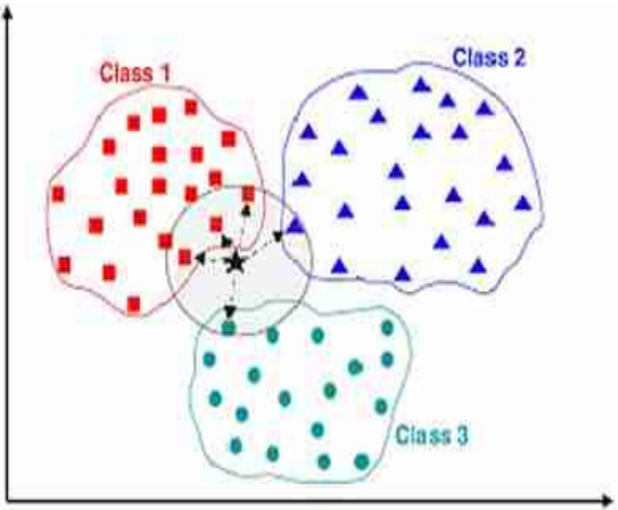
How to construct individual decision trees?



KNN

What is the KNN?

- An object is classified by a vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors
  - k is a positive integer, typically small
- Non-parametric model and lazy learning algorithm
  - lazy learning: user need not construct model
  - non-parametric model: Algorithm computational complexity (# of parar is related # of training samples



- Distance measure
  - **Minkowski Distance**: Generalization of Euclidean and Manhattan distance.

$$D(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^{d} |x_{il} - x_{jl}|^{1/p} \right)^p .$$