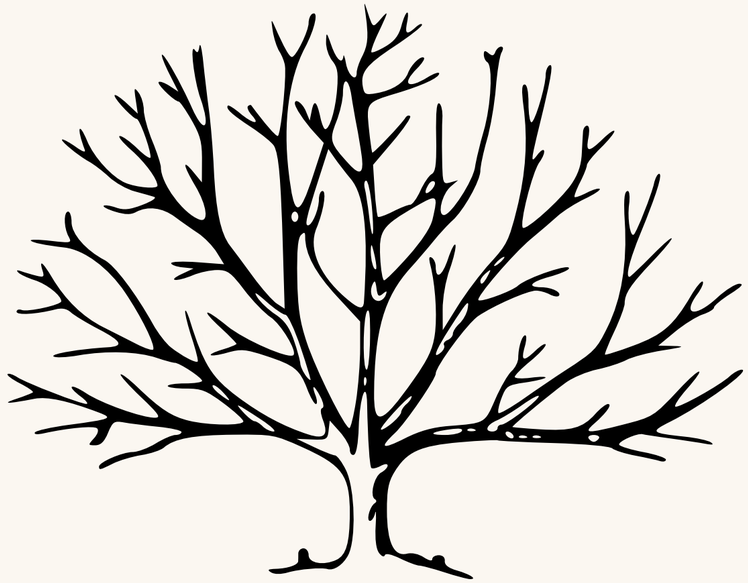
DECISION TREE



Agenda

1. Introduction
2. Advantage And limitations
3. Overfitting
4. Early Stoping
5. Pruning

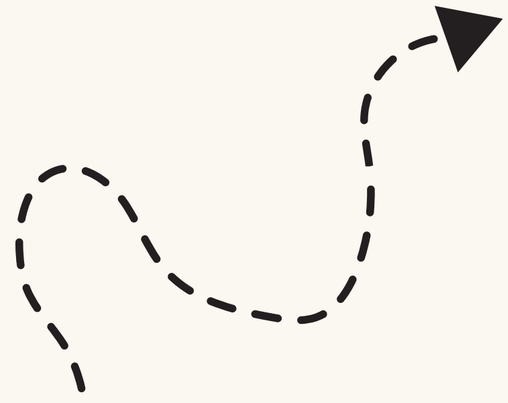
Introduction

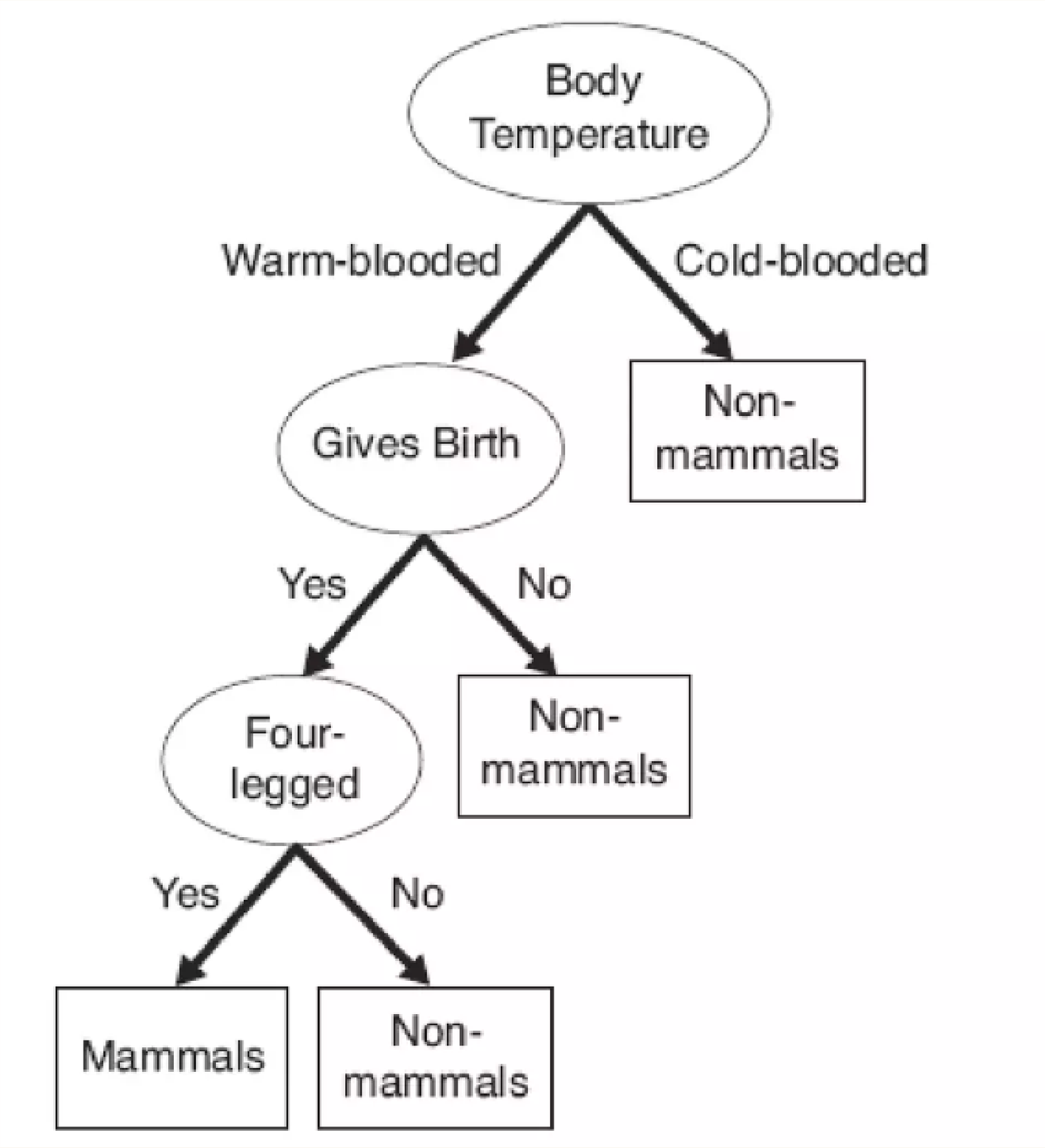
Decision Tree is supervised algorithm for classification and regression.

The decision tree algorithm builds a tree-like model of decisions and their possible consequences.

Each branch node represents choice between number of alternatives.

Each leaf node represents decision.




Body Temperature
Warm-blooded
Cold-blooded
Gives Birth
Non-mammals
Yes
No
Four-legged
Non-mammals
Yes
No
Mammals
Non-mammals

Advantages

1. Simple to understand and to interpret
2. handle both numerical and categorical features
3. Requires little data preparation
4. can capture nonlinear relationships between features and the target variable

Limitations

1. Overfitting
2. Lack of robustness(sensetive to small changes)
3. Bias towards features with more levels

OVERFITTING

If decision tree is fully grown, it may loses some generalization capabilities.
This phenomena is called Overfitting

There are different techniques to avoid overfitting in decision tree. They are:

1. Early Stopping
2. Pruning

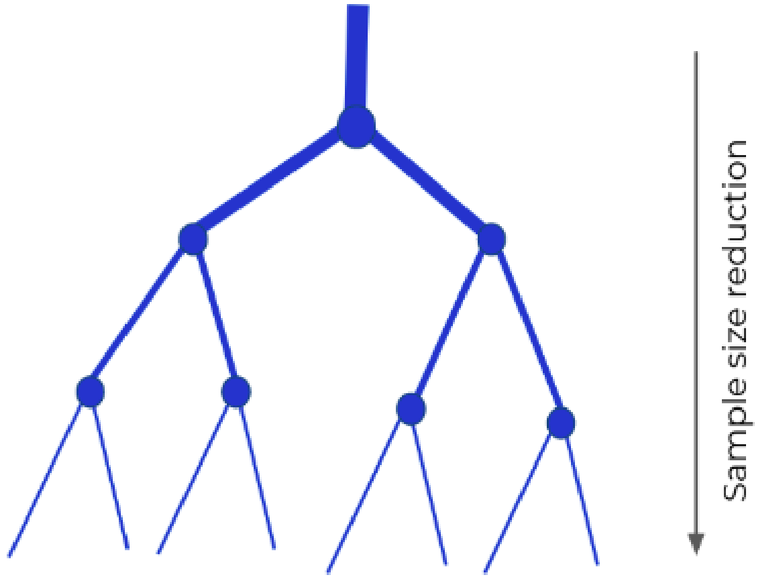


Figure 1: Decision tree with thickness of edge representing the fraction of samples available while making split.

While going deeper, the size of samples at each node decreases.

The decision(split) based on a small subset is specific to the local pattern present in those subsets.

So the decisions, at the deeper part of the tree, based on the local patterns in a subset of data, cause overfitting in the decision tree.

Early Stoping

- Top-down approach (from the root node to the leaf node)
- Prevents the generation of branch/nodes by imposing certain early stopping criteria.
- These criteria are set before training a tree and checked while generating split of a node in a decision tree
- For example, check the max_depth limitation before generating a new split.

Pruning

- Down up approach (from the leaf node to the root node)
- A decision tree is grown to its entirety, and the nodes/branches are removed or pruned.
- If the removal of node/branch results in a reduction in a validation error, then we keep the change, else revert it.