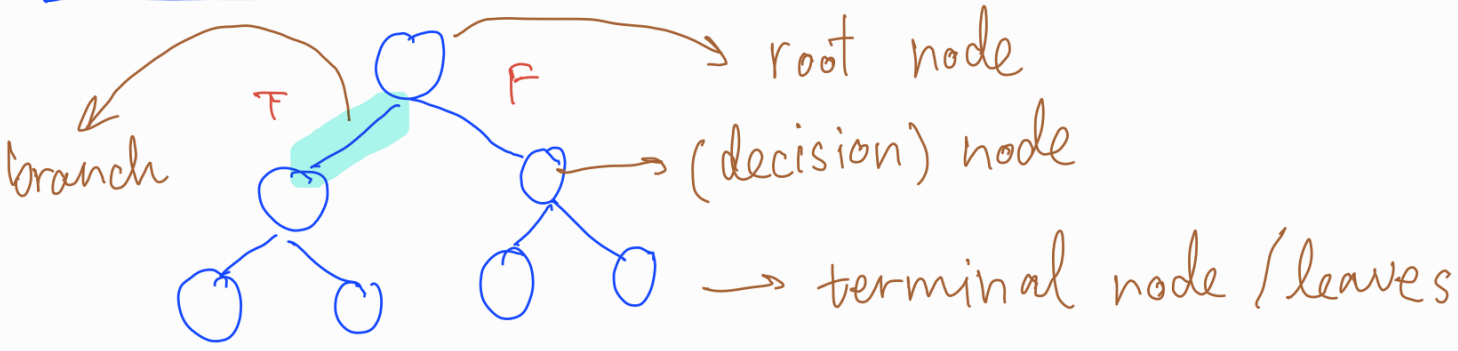
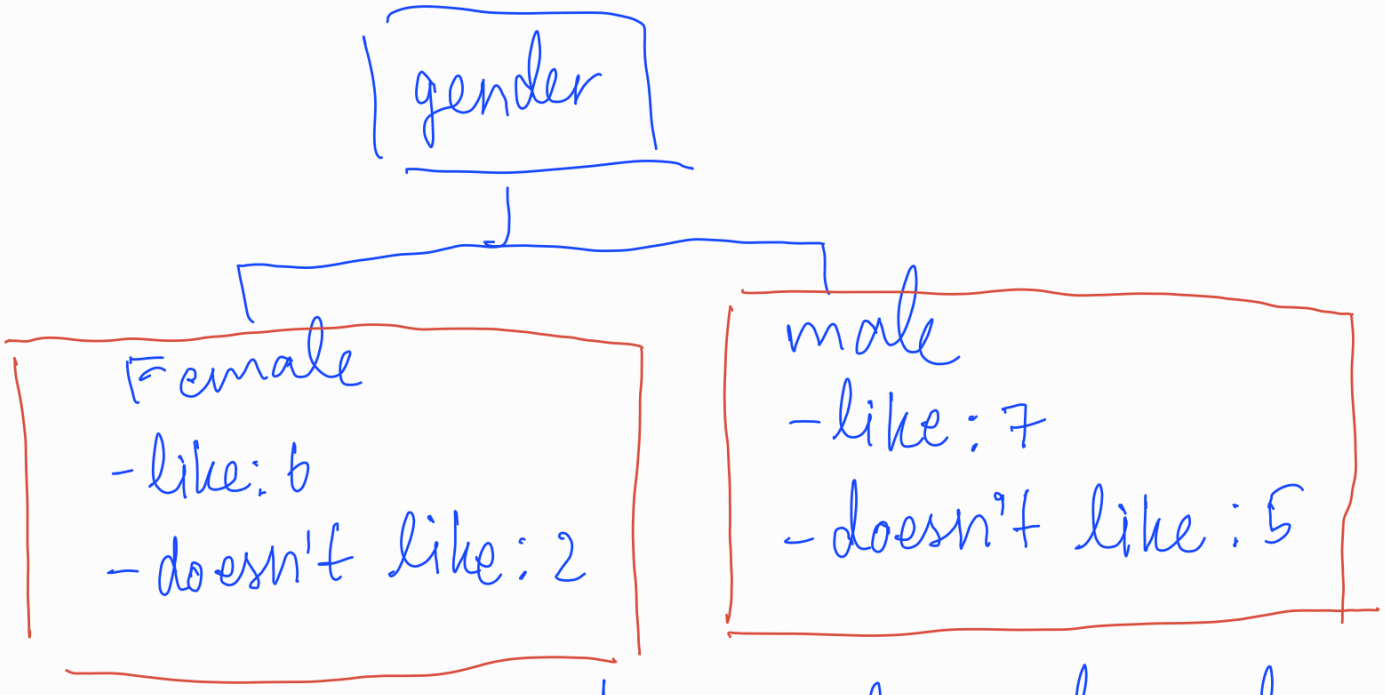
Decision tree (binary)



create decision tree

- Predict who like Taylor Swift
  - attribute → gender, age, income



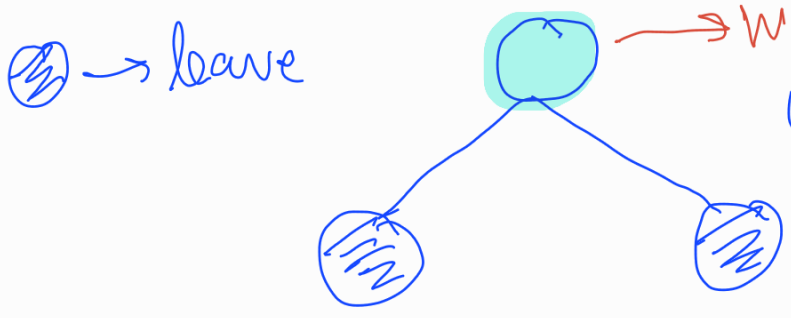
* We can see that Female and male node is n't impure
  - node can't split data purely (In female, it still have "like" & "doesn't like" group

To calculate impurity of node, we used "Gini Coefficient"

↳ low value, more pure (0 = pure)

* weighted gini coefficient *

↳ compute gini coefficient for parent node



* Decision tree can lead to "overfitting"
* Decision tree can be used with discrete & continuous value

To solve overfitting problem, they introduce "CART algorithm"

↳ "pruning" help prevent overfit (start from leave back to parent node)

reduce the tree size

Disadvantage of decision tree

- for categorical data, it can be bias.
  If we have more "like" data than "dislike", it leads to bias since DT is based on majority vote.

Random Forest

- ensemble of many decision trees
  Laggregate the decision/result from decision trees and use majority vote/averaging to come up w/ one answer
- RF provide insight on importanc of eac ature/ attribute
- RF related to bootstrapping & bagging (bootstrap aggregation) concept