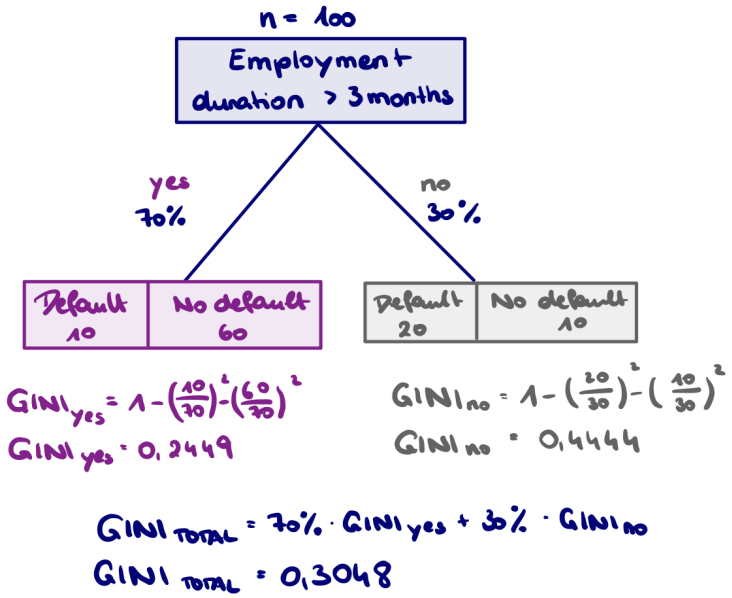
Decision Tree Building Process

- Sample imported with target and different features, which can be numerical or categorical
- Categorical feature:
  - Calculate frequencies of target class per category
  - Compute GINI per category
  - Determine weight of each category and calculate total GINI
- Numerical feature:
  - Each distinct value can be used as split → evaluate impurity for each possible split
  - Create fictive split (bucket 1, bucket 2) at one value and calculate frequencies of target class per bucket
  - Compute GINI per bucket, determine weight of each bucket and calculate total GINI
  - These steps need to be done for each distinct value of numerical feature

$$GINI = 1 - \sum_{i=1}^{k} p_i^2$$



Decision Tree Building Process

- After GINIs of all features (numerical and categorical) was calculated, the split with the lowest GINI is selected as the next split if size of each data set is not too small
- Process is repeated for all features until all splits end as leaves:
  - if GINI of next possible split is lower and size of each data set is not too small, then the split will be performed. No further split will be done, if resulting node is pure.
  - if GINI of next possible split is higher or size of one data set is too small, then the split will not be performed.
- The most occurring target class or average value is the predicted value of the terminal node