B. Working

1. Selection of attribute or feature for root node.
2. Spliting of node to subnodes.

→ Creation of subnodes increases the homogeneity of resultant subnodes or purity of node increases with respect to target.

→ DT splits the node in all possible classes and select the one with more homogenity or less impurity for further split to reach leaf node.

→ Various possible use cases:

- Features (continous) – Target (categorical)
  ➡ Classification problem
- Features (categorical) – Target (categorical)
  ➡ classification problem
- Features (continous) – Target (continous)
  ➡ Regression problem

→ Algorithm of selection:

- selection of attribute to do split is done by algorithms.
- The algorithm to use for this purpose depends on target and feature variable types.
- Algorithms use attribute selection measure to select attribute.
- Several attribute selection measures are:

Entropy, Information Gain, Gini Index, Gain Ratio, Chi Square, Reduction in Variance.

- These values are calculated for each attribute and then sorted and put into tree by that order.
- The values are sorted as increasing order of homogenity and decreasing order of impurity.

→ Entropy $\propto$ Impurity $\left(\frac{1}{\text{Homogenity}}\right)$

→ Information Gain $\propto$ $1/\text{Impurity}$ $\left(\frac{1}{\text{Homogenity}}\right)$

→ Gini Index $\propto$ $1/\text{Impurity}$.

→ Chi square $\propto$ Homogenity $(1/\text{Impurity})$

→ Various selection algorithm:

- **Iterative Dichotomiser 3 (ID3)**
  - extension of D3
  - Uses Gini Index and Information Gain
  - Used for classification problem (continous)
- **C4.5**
  - Sucessor of ID3 (use Gain ratio)
  - Uses Information Gain (Normalized)
  - Used for classification problem (with continous or categorical features)
  - Does this by making threshold for split.
- **Classification and Regression Tree (CART)**
  - Uses Gini index
  - Greedy algorithm
  - Used for classification and regression both.
- **CHAID**
  - Chi square automatic interaction detection
  - performs multi level split while computin
  - Recursive algorithm
  - Works for classification and regression both.

C. Pruning

- A technique which removes part of decision tree which prevent it from growing to full length or depth.
- This restricts decision tree to go over complex and do overfitting.
- We do this by tunning the hyperparameters of our Decision Tree.

- **Pre Pruning**
  - prunes the DT prior to training pipeline.
  - It uses heuristic approch called 'early stopping' which stops it from growing to full depth.
  - It stops tree building process to avoid producing leaves with small samples.
  - During each stage cross validation error is monitored and if value of error dont decrease we stop the growth.
  - Hyperparameters to tune:
    'max_depth', 'min_samples_leaf', 'min_samples_split'.
- **Post Pruning**
  - Once DT grows to full depth, branches are removed to avoid overfitting.
  - One issue is that, introduction of new data point lead to wrong prediction.
  - Hyperparameter to tune is Cost complexity pruning:
    'ccp_alpha'