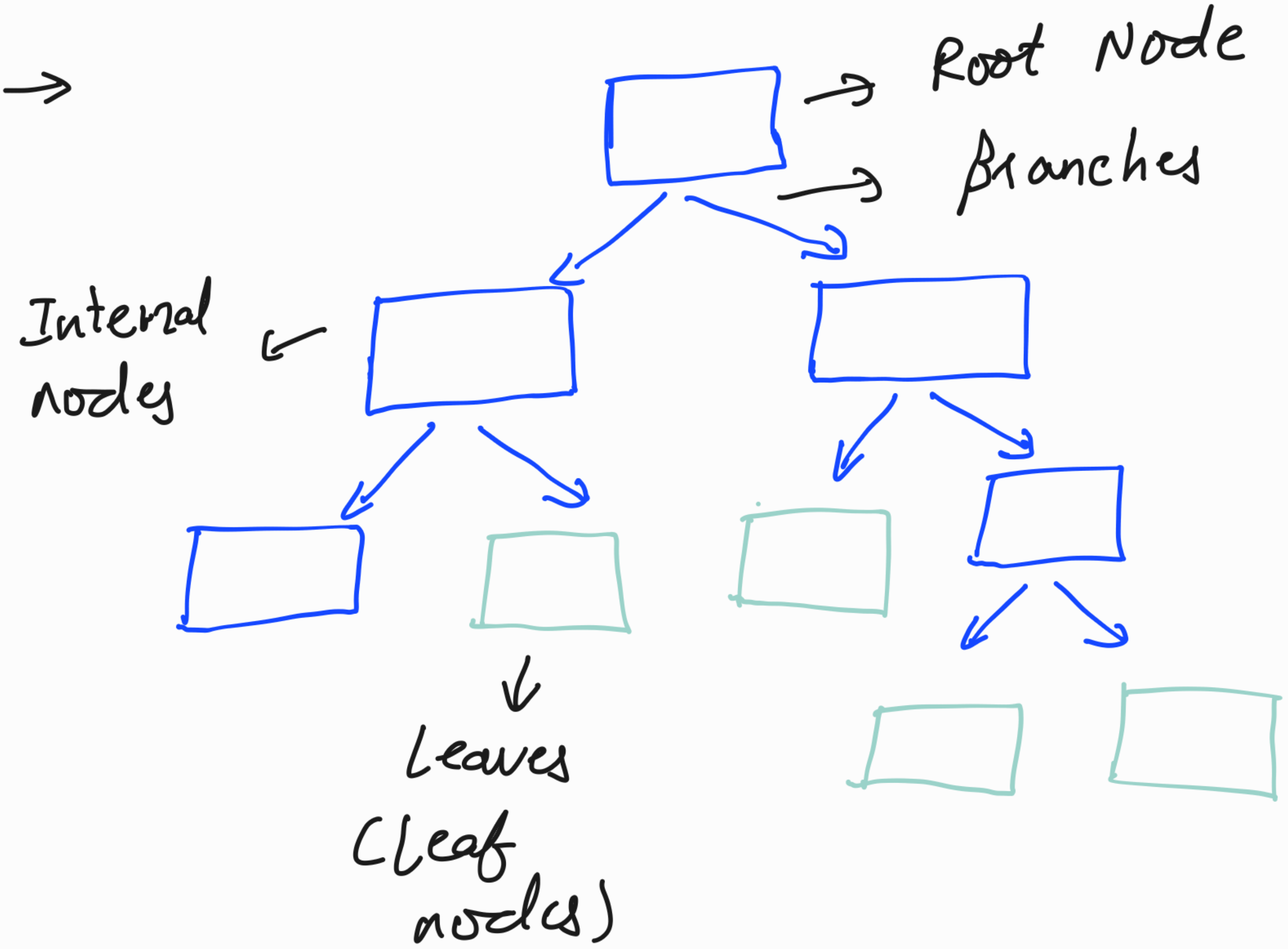
→ If the tree is trying to predict a categorical output then it is called classification, if it is predicting a numerical value then it is called regression tree.

→



→ If leaves contain mixture of decision in the leaves then it is called as impure.

→ ways to quantify impurity

i) Gini impurity

ii) Entropy

iii) Information gain

Gini impurity of a leaf

= 1 − (the probability of yes)^2 − (Probability of No)^2

Total Gini impurity = weighted average of gini impurities for the leaves

→ To prevent overfitting we can prune the trees.

→ In regression trees, like how we use gini impurity to decide the root node and subsequent division for internal nodes, in this we use SSR (Sum of Squared Residuals).

→ As a general rule of thumb min 20 data points should be there to contribute to the split.

→ Tree pruning:

Tree score = $SSR + \alpha T$

SSR ⇒ Sum of squared Residuals

$$\left.\begin{array}{l}\alpha \Rightarrow \text{Found using CV}\\ T \Rightarrow \text{Total nr of leaves}\end{array}\right\}\text{Tree complexity penalty}$$

→ This process above is known as cost complexity pruning.

Random Forests

- → create a bootstrapped dataset
- → create a decision tree using the bootstrapped dataset. (considering a random subset of variables at each step)
- → Bootstrapping the data plus using the aggregate to make a decision is called Bagging.