

# Decision Tree Algorithm:-

→ Supervised Algorithm (Classification & Regression)

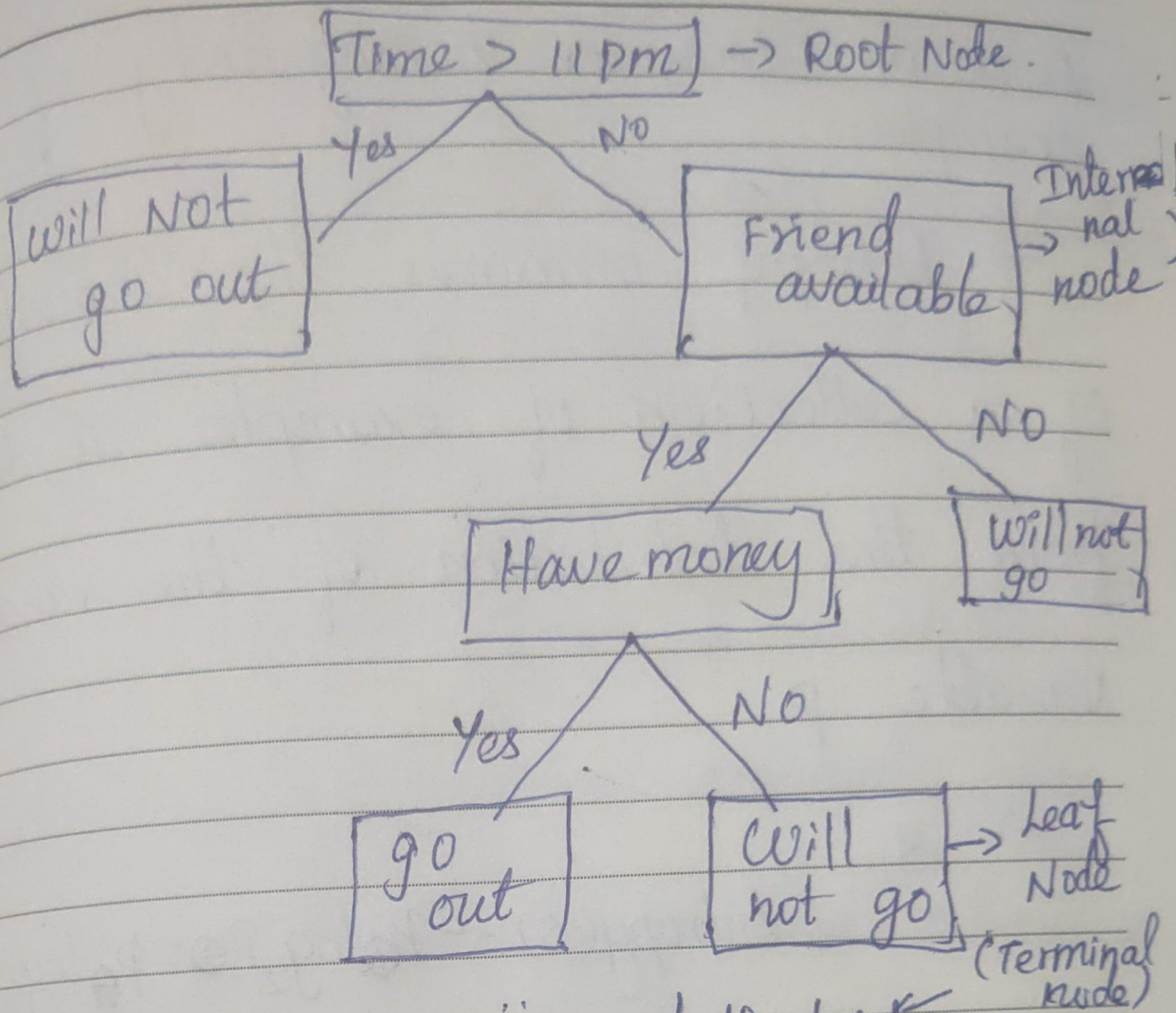
→ It is simple nested if-else condition.

## Decision Tree?

Task - decide whether I should go out or not?

Data:	Time	Friend Available	have money	go out
	3Pm	Yes	No	No
	2Pm	No	Yes	No
	11 pm	Yes	Yes	No
	6 pm	Yes	Yes	Yes
	2 am	no	Yes	No
	4 pm	Yes	Yes	Yes.





represents the label.

DT:-

It uses tree representation to solve the problem.

→ To select the correct feature we can do split or get pure split we can use here

→ Entropy

→ Gini Impurity



ENTROPY Measures homogeneity of examples:-

→ Entropy measures the impurity of a collection of examples. It depends from the distribution of the random variable  $p$ .

Ex 8

$$\text{Entropy}(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

$S$  is a collection of training of examples.

$P_+$  the proportion of positive example in  $S$

$P_-$  the proportion of negative example in  $S$ .

Ex:

$$\text{Entropy}([14+, 0-]) = -14/14 \log_2 (14/14) - 0 \log_2 (0) = 0$$



If Entropy = 1, then its the impure split  
Entropy = 0,  $\rightarrow$  Pure split. (leafnode=1)  
 $\rightarrow$  Entropy range always be 0 to 1.

Information Gain:-

$\rightarrow$  Information gain is used to determine which feature/attribute gives us the maximum information about a class.

$\rightarrow$  IG is based on the concept of entropy, which is the degree of uncertainty, impurity or disorder.

$\rightarrow$  IG aims to reduce the level of entropy starting from the root



node to the leaf nodes.

→ The greater the reduction in the uncertainty, the more information is gained about  $Y$  from  $X$ .

$$\text{Gain}(T, X) = \text{Entropy}(T) - \text{Entropy}(T, X)$$

$T$  → parent node before split

$X$  → split node from  $T$ .

### GINI INDEX:-

→ Gini Index or Gini Impurity measures the degree or probability of a particular variable.

$$\text{Gini Index} = 1 - \sum_{i=1}^n (P_i)^2$$

$$\text{Gini Range} \rightarrow 0 \text{ to } 0.5$$

## Grid Search CV:-

→ It helps to find the best params and fit the estimator (model) on training data.

→ It is a hyperparameter tuning.

## Drawback of DT:-

→ Overfitting.