





# DECISION TREE

PROF.BAVRABI GHOSH  
Department of CSE  
INSTITUTE OF ENGINEERING AND MANAGEMENT, KOLKATA



# Topics to be covered

- What is Decision Tree?
- Structure details of a Decision Tree
- How to build a Decision Tree?
- Decision Tree based on training dataset
- Entropy and Information Gain of Decision Tree
- Calculation of Entropy
- Calculation of Information Gain
- Conclusion

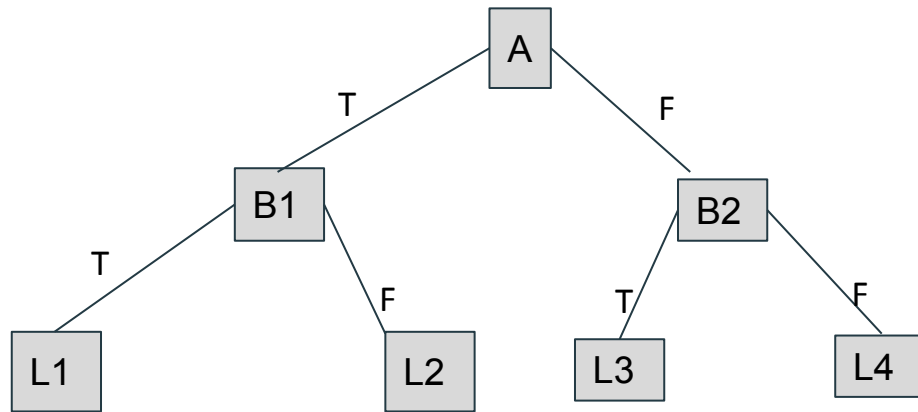
# DECISION TREE

- Decision tree is an algorithm for Classification type Machine Learning
- The model is built in form of a Tree like structure
- It is used for multi-dimensional analysis with multiple classes.
- Decision tree is one of user's favourite because of the following two reasons:
  - Fast execution
  - Ease in interpretation of rule

# DECISION TREE contd...

- **Goal** -> To create a model (on past data) or Past vector, where value of output variable is predicted based on input variable in feature vector
- **Each node** -> one of the feature vector
- **Each leaf node** -> one output value
- Which node is to be chosen for branching depends on which node gives highest information gain value.

# STRUCTURE DETAILS OF DECISION TREE



Feature Vector - A,B  
(attribute)

Leaf Node - L1, L2, L3 and L4  
(classification /assignment  
of a class)

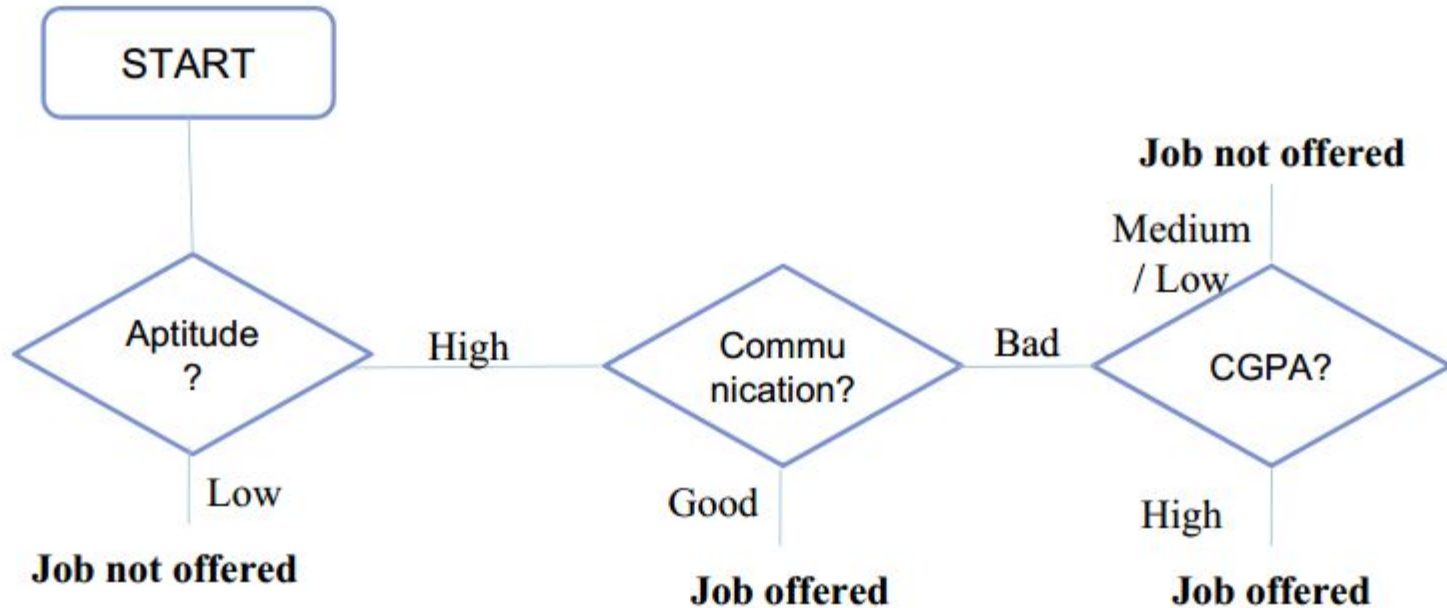
Root Node - A

Branch Node- B1, B2

# DATASET

<b>CGPA</b>	<b>Communication</b>	<b>Aptitude</b>	<b>Programming Skill</b>	<b>Job offered?</b>
High	Good	High	Good	Yes
Medium	Good	High	Good	Yes
Low	Bad	Low	Good	No
Low	Good	Low	Bad	No
High	Good	High	Bad	Yes
High	Good	High	Good	Yes
Medium	Bad	Low	Bad	No
Medium	Bad	Low	Good	No
High	Bad	High	Good	Yes
Medium	Good	High	Good	Yes
Low	Bad	High	Bad	No
Low	Bad	High	Bad	No
Medium	Good	High	Bad	Yes
Low	Good	Low	Good	No
High	Bad	Low	Bad	No
Medium	Bad	High	Good	No
High	Bad	Low	Bad	No
Medium	Good	High	Bad	Yes

# Decision Tree based on training dataset



# ENTROPY AND INFORMATION GAIN OF DECISION TREE

- Entropy is an expression of the disorder, or randomness of a system, or of the lack of information about it.
- Information gain is the parameter that decides which feature gives us maximum amount of information required to “decide” which will be the next feature to be chosen as “Node from where split happens” for the next stage.



# CALCULATION OF ENTROPY AND INFORMATION GAIN

$$\text{Entropy (S)} = -\sum_{i=1}^c p_i \log_2 p_i$$

$$\text{Information Gain (S, A)} = \text{Entropy (S}_{\text{bs}}) - \text{Entropy (S}_{\text{as}})$$

$$\text{Entropy (S}_{\text{as}}) = \sum_{i=1}^n w_i \text{Entropy}(p_i)$$

Entropy ( $S_{\text{bs}}$ ) -> Entropy Before Split

Entropy ( $S_{\text{as}}$ ) -> Entropy After Split

# CALCULATION OF ENTROPY AND INFORMATION GAIN contd..

(a) Original data set:

	Yes	No	Total
Count	8	10	18
pi	0.44	0.56	
-pi*log(pi)	0.52	0.47	0.99

Total Entropy = 0.99

CALCULATIONS

$$pi(\text{yes})8/18=0.44$$

$$-pi * \log_2(pi) = 0.52$$

$$pi(\text{no})10/18= 0.56$$

$$-pi * \log_2(pi) = 0.47$$

$$\text{Total entropy} = 0.52 + 0.47 = 0.99$$

# CALCULATION OF ENTROPY AND INFORMATION GAIN contd..

(b) Splitted data set (based on feature "CGPA"):

CGPA = High

	Yes	No	Total
Count	4	2	6
pi	0.67	0.33	
-pi*log(pi)	0.39	0.53	0.92

CGPA = Medium

	Yes	No	Total
Count	4	3	7
pi	0.57	0.43	
-pi*log(pi)	0.46	0.52	0.99

CGPA = Low

	Yes	No	Total
Count	0	5	5
pi	0.00	1.00	
-pi*log(pi)	0.00	0.00	0.00

Total Entropy = 0.69

Information Gain = 0.30

Total no of instances (row in dataset)= 18

Total Entropy(after split) =  $((6 \times 0.92) + (7 \times 0.99) + (5 \times 0.00)) / 18 = 0.69$

**(this is weighted entropy )**

Information Gain = Entropy (before split) - Entropy (after split) =  $0.99 - 0.69 = 0.30$

## CALCULATION OF ENTROPY AND INFORMATION GAIN contd..

### (c) Splitted data set (based on feature "Communication"):

Communication = Good

	Yes	No	Total
Count	7	2	9
pi	0.78	0.22	
$-pi \cdot \log(pi)$	0.28	0.48	0.76

Total Entropy = 0.63

Communication = Bad

	Yes	No	Total
Count	1	8	9
pi	0.11	0.89	
$-pi \cdot \log(pi)$	0.35	0.15	0.50

Information Gain = 0.36

Depending on information gain of "CGPA", "Communication", "Aptitude" and "Programming Skill", it is found that "Aptitude" has the highest information gain.

Therefore "Aptitude" is chosen as the first feature based on which split is to take place.

### (d) Splitted data set (based on feature "Aptitude"):

Aptitude = High

	Yes	No	Total
Count	8	3	11
pi	0.73	0.27	
$-pi \cdot \log(pi)$	0.33	0.51	0.85

Total Entropy = 0.52

Aptitude = Low

	Yes	No	Total
Count	0	7	7
pi	0.00	1.00	
$-pi \cdot \log(pi)$	0.00	0.00	0.00

Information Gain = 0.47

### (e) Splitted data set (based on feature "Programming Skill"):

Programming Skill = Good

	Yes	No	Total
Count	5	4	9
pi	0.56	0.44	
$-pi \cdot \log(pi)$	0.47	0.52	0.99

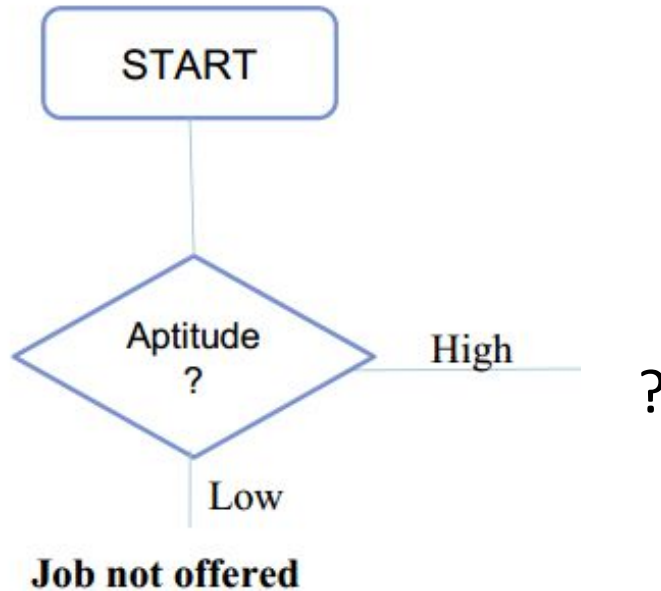
Total Entropy = 0.95

Programming Skill = Bad

	Yes	No	Total
Count	3	6	9
pi	0.33	0.67	
$-pi \cdot \log(pi)$	0.53	0.39	0.92

Information Gain = 0.04

# CALCULATION OF ENTROPY AND INFORMATION GAIN contd..



To choose feature for the next node we will carry on with "Entropy" and "Information Gain" calculation. But this time the size of dataset decreases as we eliminate the instances for which "Aptitude" is **low**. Because Low aptitude does not fetch you a job. So that branch of the tree freezes.

## CALCULATION OF ENTROPY AND INFORMATION GAIN contd..

(a) Level-2 starting set:

	Yes	No	Total
Count	8	3	11
pi	0.73	0.27	
$-pi \cdot \log(pi)$	0.33	0.51	0.85

Total Entropy = 0.85

(b) Splitted data set (based on feature "CGPA"):

CGPA = High

	Yes	No	Total
Count	4	0	4
pi	1.00	0.00	
$-pi \cdot \log(pi)$	0.00	0.00	0.00

Total Entropy = 0.33

CGPA = Medium

	Yes	No	Total
Count	4	1	5
pi	0.80	0.20	
$-pi \cdot \log(pi)$	0.26	0.46	0.72

Information Gain = 0.52

CGPA = Low

	Yes	No	Total
Count	0	2	2
pi	0.00	1.00	
$-pi \cdot \log(pi)$	0.00	0.00	0.00

Aptitude = High

CGPA	Communication	Programming Skill	Job offered?
High	Good	Good	Yes
Medium	Good	Good	Yes
High	Good	Bad	Yes
High	Good	Good	Yes
High	Bad	Good	Yes
Medium	Good	Good	Yes
Low	Bad	Bad	No
Low	Bad	Bad	No
Medium	Good	Bad	Yes
Medium	Bad	Good	No
Medium	Good	Bad	Yes

## CALCULATION OF ENTROPY AND INFORMATION GAIN contd..

### (c) Splitted data set (based on feature "Communication"):

#### Communication = Good

	Yes	No	Total
Count	7	0	7
pi	1.00	0.00	
$-pi \cdot \log(pi)$	0.00	0.00	0.00

Total Entropy = 0.30

### (d) Splitted data set (based on feature "Programming Skill"):

#### Programming Skill = Good

	Yes	No	Total
Count	5	1	6
pi	0.83	0.17	
$-pi \cdot \log(pi)$	0.22	0.43	0.65

Total Entropy = 0.80

#### Communication = Bad

	Yes	No	Total
Count	1	3	4
pi	0.25	0.75	
$-pi \cdot \log(pi)$	0.50	0.31	0.81

Information Gain = 0.55

#### Programming Skill = Bad

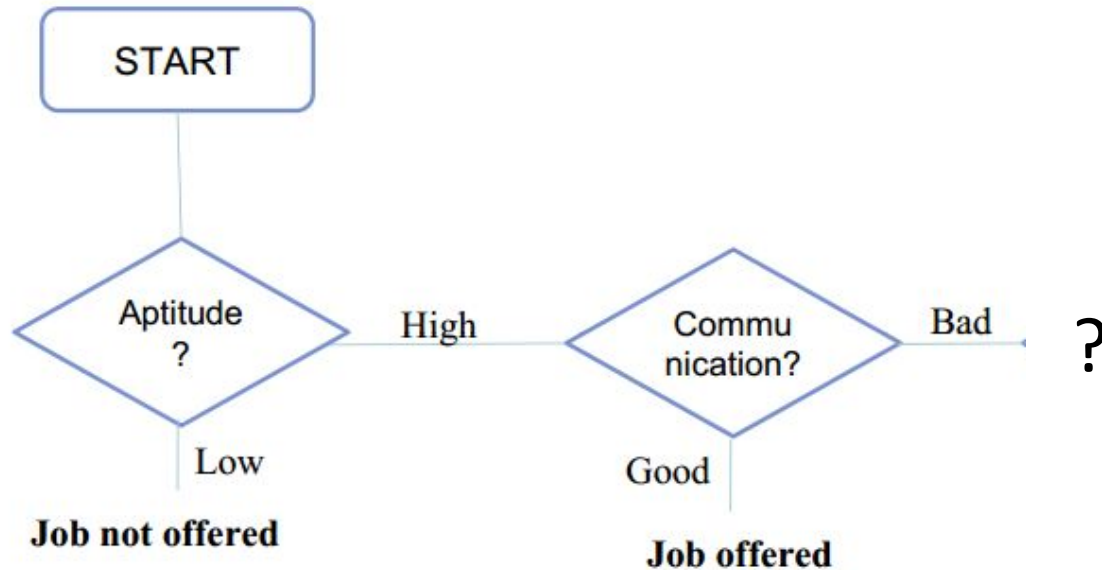
	Yes	No	Total
Count	3	2	5
pi	0.60	0.40	
$-pi \cdot \log(pi)$	0.44	0.53	0.97

Information Gain = 0.05

Aptitude = High

CGPA	Communication	Programming Skill	Job offered?
High	Good	Good	Yes
Medium	Good	Good	Yes
High	Good	Bad	Yes
High	Good	Good	Yes
High	Bad	Good	Yes
Medium	Good	Good	Yes
Low	Bad	Bad	No
Low	Bad	Bad	No
Medium	Good	Bad	Yes
Medium	Bad	Good	No
Medium	Good	Bad	Yes

# CALCULATION OF ENTROPY AND INFORMATION GAIN contd..





CALCULATION OF ENTROPY AND INFORMATION GAIN contd..

**Aptitude = High & Communication = Bad**

<b>CGPA</b>	<b>Programming Skill</b>	<b>Job offered ?</b>
High	Good	Yes
Low	Bad	No
Low	Bad	No
Medium	Good	No

**(a) Level-2 starting set:**

	<b>Yes</b>	<b>No</b>	<b>Total</b>
Count	1	3	4
pi	0.25	0.75	
-pi*log(pi)	0.50	0.31	0.81

**Total Entropy = 0.81**

## CALCULATION OF ENTROPY AND INFORMATION GAIN contd..

### (b) Splitted data set (based on feature "CGPA"):

CGPA = High

	Yes	No	Total
Count	1	0	1
pi	1.00	0.00	
$-pi \cdot \log(pi)$	0.00	0.00	0.00

Total Entropy = 0.00

CGPA = Medium

	Yes	No	Total
Count	0	1	1
pi	0.00	1.00	
$-pi \cdot \log(pi)$	0.00	0.00	0.00

Information Gain = 0.81

CGPA = Low

	Yes	No	Total
Count	0	2	2
pi	0.00	1.00	
$-pi \cdot \log(pi)$	0.00	0.00	0.00

### (c) Splitted data set (based on feature "Programming Skill"):

Programming Skill = Good

	Yes	No	Total
Count	1	1	2
pi	0.50	0.50	
$-pi \cdot \log(pi)$	0.50	0.50	1.00

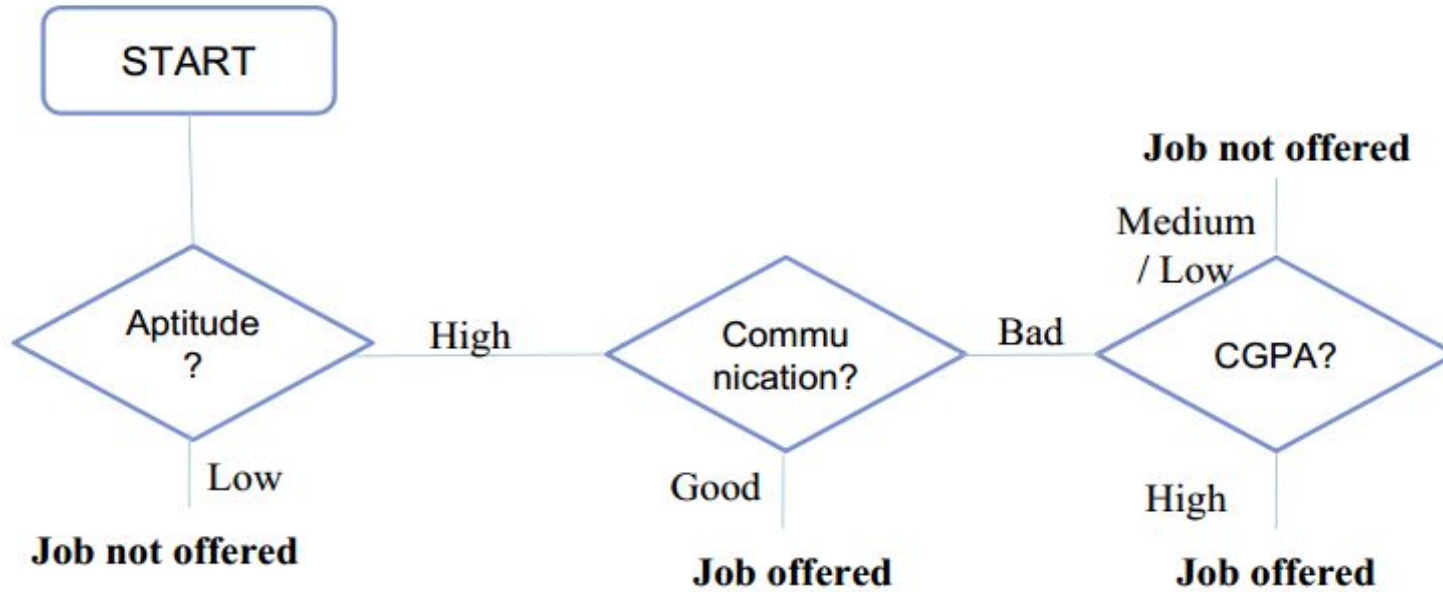
Total Entropy = 0.50

Programming Skill = Bad

	Yes	No	Total
Count	0	2	2
pi	0.00	1.00	
$-pi \cdot \log(pi)$	0.00	0.00	0.00

Information Gain = 0.31

# CALCULATION OF ENTROPY AND INFORMATION GAIN contd..



This is the final decision tree

# CODE SNIPPET

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.tree import DecisionTreeClassifier
data=pd.read_csv("../input/apndcts/apndcts.csv")

predictors= data.iloc[:,0:7]   #To separate features that will act as predictors

target = data.iloc[:,7]       #To separate features that will act as target(class variable)

predictors_train, predictors_test, target_train, target_test = train_test_split(predictors, target, test_size=0.3 , random_state = 123)

dtree_entropy = DecisionTreeClassifier (criterion="entropy" , random_state=100 , max_depth = 3, min_samples_leaf=5)

# this is to fit the model(training)
model= dtree_entropy.fit(predictors_train, target_train)

#this is to test the model (to see whether it gives right prediction)
prediction= dtree_entropy.predict(predictors_test)

#this is to check accuracy
accuracy_score(target_test, prediction, normalize=True)
```

Out[10] 0.84375

THANK YOU