Decision Tree

(ID3, C4.5, CART)

By
Assistant Prof. M. B. Narnaware
(WCE-IT)

Agenda: To understand the mathematics behind Decision Tree

Key variations of the Decision Tree

1. ID3: Iterative Dichotomiser 3
2. C4.5
    a. For Discrete Variables
    b. For Continuous Variables.
3. CART: Classification and Regression Tree

ID3: Iterative Dichotomiser 3

Key term to understand for ID3:

1. Entropy: Definition
2. Entropy before split
3. Entropy after split
4. Information Gain

Key Term: Entropy

Definitions from Web:

➔ A way of measuring the amount of order/uncertainty present or absent in a system.

➔ Entropy is a scientific concept that is most commonly associated with a state of disorder, randomness, or uncertainty.

★ Higher the Entropy → Higher the Uncertainty.

★ Lower the Entropy → Lower the Uncertainty.

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

Where: p_i **= Probability of Event**

Key Term: Information Gain

Information gain for a particular feature A is calculated by the difference in entropy before a split (or S_{bs}) with the entropy after the split (S_{as}).

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

For calculating the entropy after split, entropy for all partitions needs to be considered. Then, the weighted summation of the entropy for each partition can be taken as the total entropy after split. For performing weighted summation, the proportion of examples falling into each partition is used as weight.

$$\textbf{Entropy}(S_{as}) = \sum_{i=1}^{n} w_i \,\text{Entropy}\,(p_i)$$

Example: Entropy Calculation

CGPA	Communication	Aptitude	Programming Skill	Job offered?
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

Entropy Calculation:

(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

Details

- Original dataset contains total 18 rows/entries.
- Entries with **YES** labels are 8.
- Entries with **NO** labels are 10.
- Probability of **YES** is 0.44
- Probability of **NO** is 0.56
- By formula:
    - Total Entropy = 0.99; rounded to two decimals

Entropy Calculation:

(b) Splitted data set (based on the 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

Entropy Calculation:

(c) Splitted data set (based on the feature 'Communication'):

Communication = Good

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

Communication = Bad

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

Total Entropy = 0.63

Information Gain = 0.36

Entropy Calculation:

(d) Splitted data set (based on the feature 'Aptitude'):

Aptitude = High

	Yes	No	Total
Count	8	3	11
pi	0.73	0.27	
-pi*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*log(pi)	0.00	0.00	0.00

Information Gain = 0.47

Entropy Calculation:

(e) Splitted data set (based on the feature 'Programming Skill'):

Programming Skill = Good

	Yes	No	Total
Count	5	4	9
pi	0.56	0.44	
-pi*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*log(pi)	0.53	0.39	092

Information Gain = 0.04

Best IG as Splitting Criterion:

- Thus Aptitude give best IG among all the features.
- Hence it should be noted that Aptitude will be the criterion of first split.
- After using Aptitude as split criterion, the original dataset will be divided into Aptitude== Low and Aptitude == High.
- For Aptitude == Low ⇒ Job Offer == NO. Hence conclusion is reached.
- For Aptitude == High ⇒ Job Offer == NO Or Yes.
- Hence, Aptitude == High, Part of the decision needs to be further explored.
- Next slide contains table when Aptitude == High.
- Now the same process needs to be repeated till conclusion is reached or stopping criterion is satisfied (To be discussed separately).

Reduced Table for Aptitude == High

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

(a) Level 2 starting set:

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

Total Entropy = 0.85

(b) Splitted data set (based on the feature 'CGPA'):

CGPA = High

	Yes	No	Total
Count	4	0	4
pi	1.00	0.00	
-pi*log(pi)	0.00	0.00	0.00

CGPA = Medium

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

CGPA = Low

	Yes	No	Total
Count	0	2	2
pi	0.00	1.00	
-pi*log(pi)	0.00	0.00	0.00

Total Entropy = 0.33

Information Gain = 0.52

(c) Splitted data set (based on the feature 'Communication'):

Communication = Good

	Yes	No	Total
Count	7	0	7
pi	1.00	0.00	
-pi*log(pi)	0.00	0.00	0.00

Communication = Bad

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

Total Entropy = 0.30

Information Gain = 0.55

(d) Spitted data set (based on the feature 'Programming Skill'):

Programming Skill = Good

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

Programming Skill = Bad

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

Total Entropy = 0.80

Information Gain = 0.05

Aptitude = High & Communication = Bad

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

(a) Level 1 starting set:

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

Total Entropy = 0.81

(b) Splitted data set (based on the feature 'CGPA'):

CGPA = High

	Yes	No	Total
Count	1	0	1
pi	1.00	0.00	
-pi*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*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*log(pi)	0.00	0.00	0.00

(c) Splitted data set (based on the feature 'Programming Skill'):

Programming Skill = Good

	Yes	No	Total
Count	1	1	2
pi	0.50	0.50	
-pi*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*log(pi)	0.00	0.00	0.00

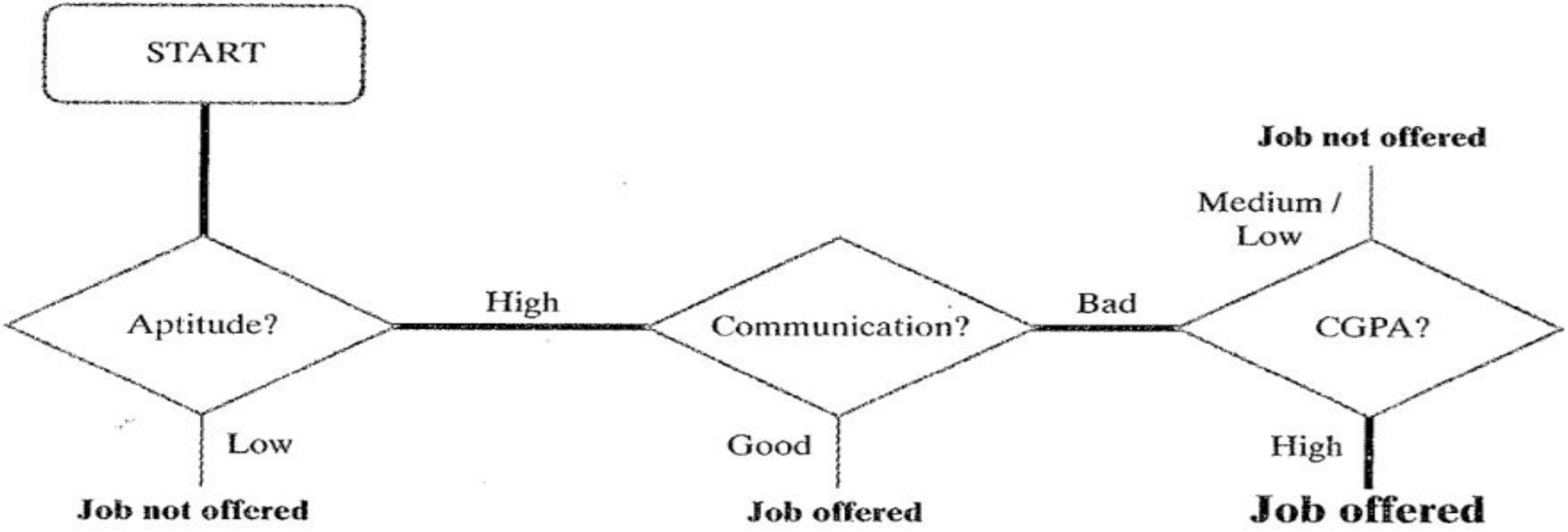
Information Gain = 0.31

FIG.
Entropy and information gain calculation (Level 3)

Stopping Criterion

- After level three Stopping Criterion is reached as Uncertainty reduces to Zero.
- The Answer is clear YES or NO.

Final Decision Tree:



Algorithm C4.5

1. C4.5 used for
    a. Successor of ID3
    b. For Discrete Variables/Feature
    c. For Continuous Variables/Feature
2. Concepts required:
    a. Entropy (As Before)
    b. Entropy after split (As Before)
    c. Information Gain (As Before)
    d. Split Info (New/Additional concept)
    e. Gain Ratio (New/ Additional concept)

Split Info:

Given a Training dataset T,

The Split_Info of an attribute A is computed as given in Eq. (6.11):

$$\text{Split\_Info}(T, A) = -\sum_{i=1}^{v} \frac{|A_i|}{|T|} \times \log_2 \frac{|A_i|}{|T|} \quad (6.1$$

where, the attribute A has got 'v' distinct values $\{a_1, a_2, \ldots a_v\}$, and $|A_i|$ is the number of instan

for distinct value 'i' in attribute A.

Gain Ratio:

The Gain_Ratio of an attribute A is computed as

$$\text{Gain\_Ratio}(A) = \frac{\text{Info\_Gain}(A)}{\text{Split\_Info}(T, A)}$$

Consider the Problem and Solve by C4.5

Example 6.3: Assess a student's performance during his course of study and predict whether a student will get a job offer or not in his final year of the course. The training dataset *T* consists of 10 data instances with attributes such as 'CGPA', 'Interactiveness', 'Practical Knowledge' and 'Communication Skills' as shown in Table 6.3. The target class attribute is the 'Job Offer'.

Table 6.3: Training Dataset *T*

S.No.	CGPA	Interactiveness	Practical Knowledge	Communication Skills	Job Offer
1.	≥9	Yes	Very good	Good	Yes
2.	≥8	No	Good	Moderate	Yes
3.	≥9	No	Average	Poor	No
4.	<8	No	Average	Good	No
5.	≥8	Yes	Good	Moderate	Yes
6.	≥9	Yes	Good	Moderate	Yes
7.	<8	Yes	Good	Poor	No
8.	≥9	No	Very good	Good	Yes
9.	≥8	Yes	Good	Good	Yes
10.	≥8	Yes	Average	Good	Yes

Iteration 1:

Step 1: Calculate the Class_Entropy for the target class 'Job Offer'.

Entropy_Info(Target Attribute = Job Offer) = Entropy_Info(7, 3) =

$$= -\left[\frac{7}{10}\log_2\frac{7}{10} + \frac{3}{10}\log_2\frac{3}{10}\right]$$

$$= (-0.3599 + -0.5208)$$

$$= 0.8807$$

Step 2: Calculate the Entropy_Info, Gain(Info_Gain), Split_Info, Gain_Ratio for each of the attribute in the training dataset.

CGPA:

$$\text{Entropy Info}(T, \text{CGPA}) = \frac{4}{10}\left[-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right] + \frac{4}{10}\left[-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right] + \frac{2}{10}\left[-\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2}\right]$$

$$= \frac{4}{10}(0.3111 + 0.4997) + 0 + 0$$

$$= 0.3243$$

$$\text{Gain(CGPA)} = 0.8807 - 0.3243$$

$$= 0.5564$$

$$\text{Split\_Info}(T, \text{CGPA}) = -\frac{4}{10}\log_2\frac{4}{10} - \frac{4}{10}\log_2\frac{4}{10} - \frac{2}{10}\log_2\frac{2}{10}$$

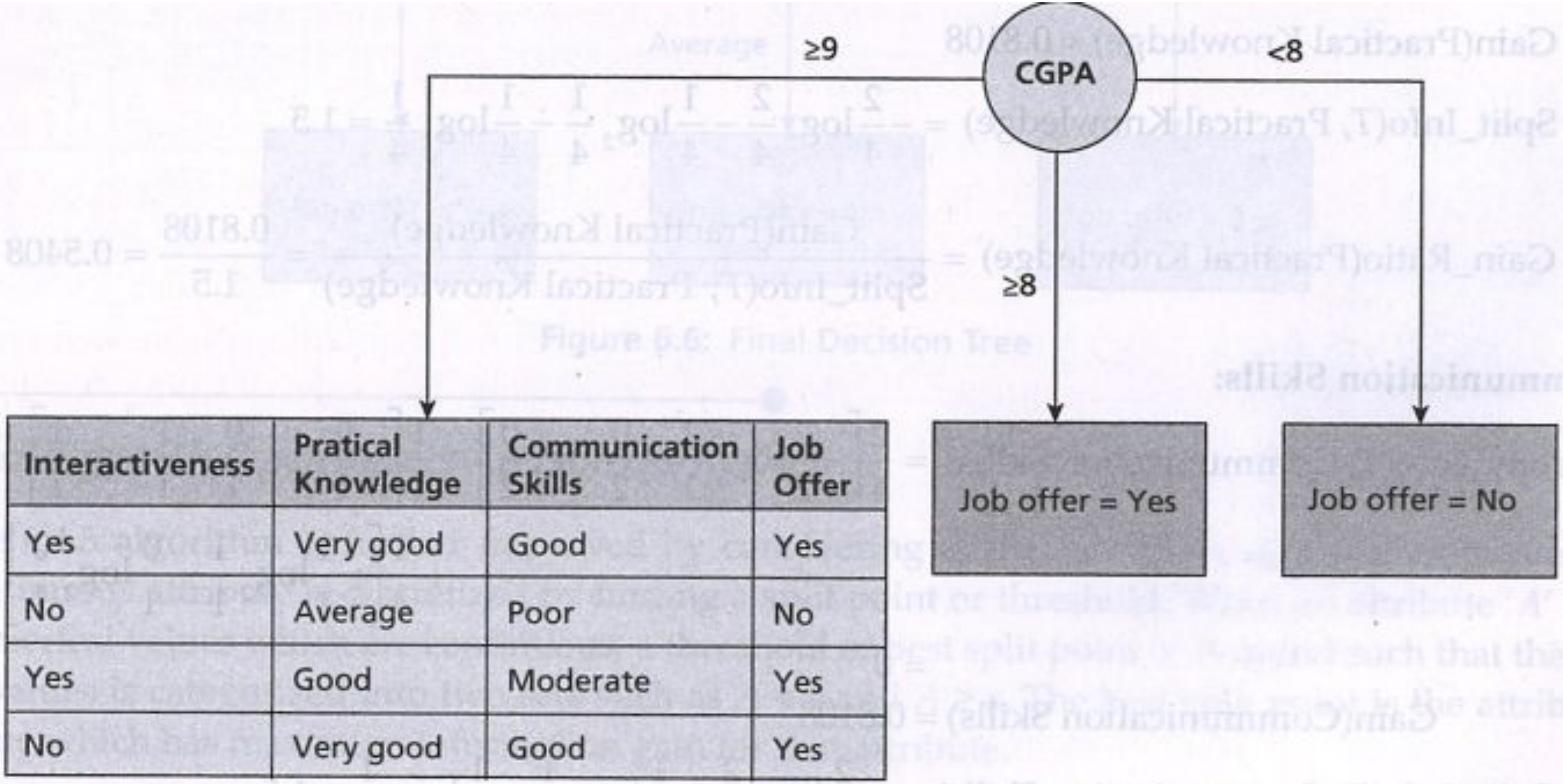
$$= 0.5285 + 0.5285 + 0.4641$$

$$= 1.5211$$

$$\text{Gain Ratio(CGPA)} = (\text{Gain(CGPA)})/(\text{Split\_Info}(T, \text{CGPA}))$$

$$= \frac{0.5564}{1.5211} = 0.3658$$

Attribute	Gain_Ratio
CGPA	0.3658
INTERACTIVENESS	0.0939
PRACTICAL KNOWLEDGE	0.1648
COMMUNICATION SKILLS	0.3502



Interactiveness	Pratical Knowledge	Communication Skills	Job Offer
Yes	Very good	Good	Yes
No	Average	Poor	No
Yes	Good	Moderate	Yes
No	Very good	Good	Yes

Figure 6.5: Decision Tree after Iteration 1

Attributes	Gain_Ratio
Interactiveness	0.3112
Practical Knowledge	0.5408
Communication Skills	0.5408

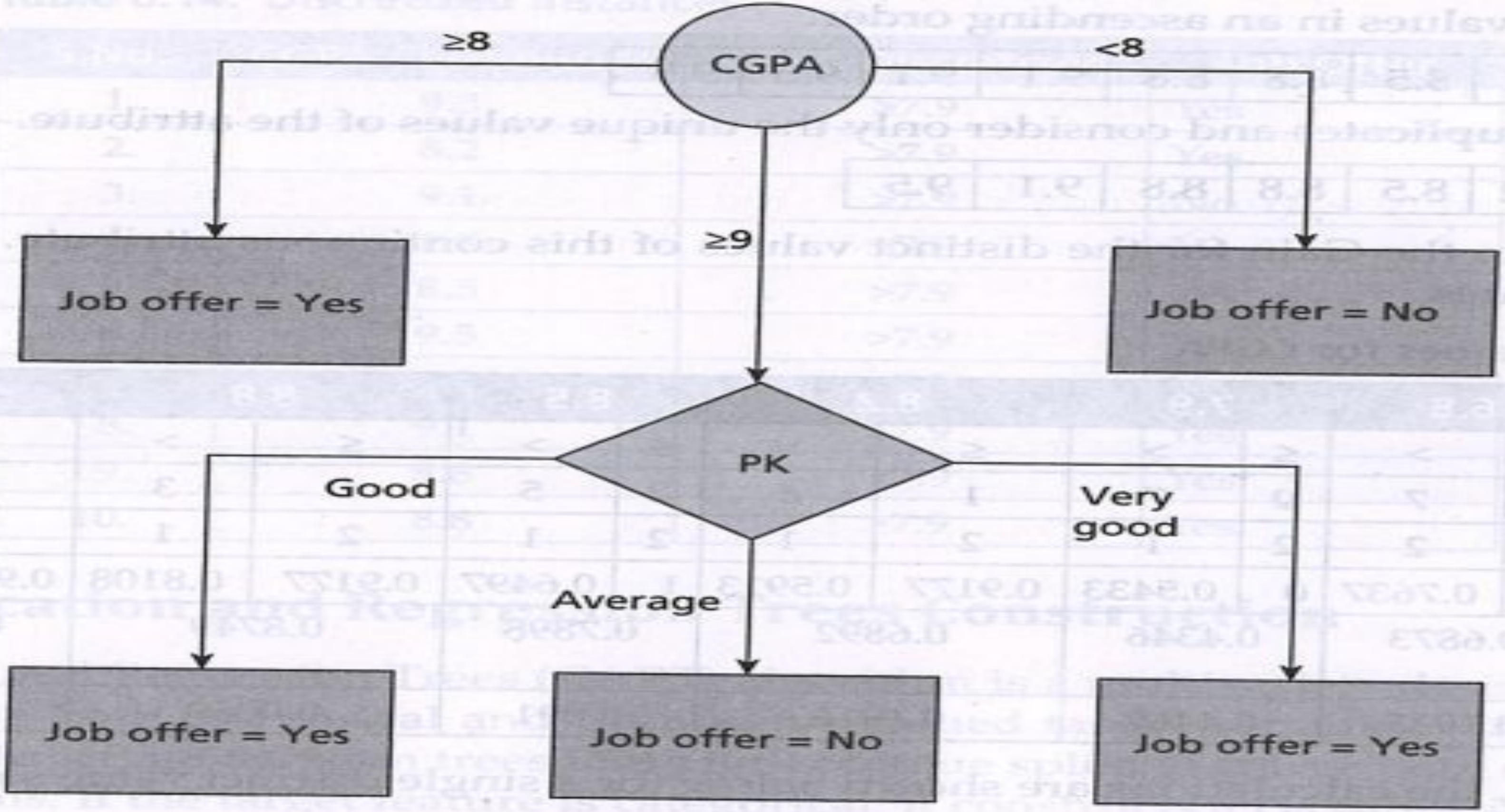


Figure 6.6: Final Decision Tree

C4.5 for Continuous Variable:

Now, let us consider the set of continuous values for the attribute CGPA in the sample dataset

Sample Dataset:

S.No.	CGPA	Job Offer
1.	9.5	Yes
2.	8.2	Yes
3.	9.1	No
4.	6.8	No
5.	8.5	Yes
6.	9.5	Yes
7.	7.9	No
8.	9.1	Yes
9.	8.8	Yes
10.	8.8	Yes

Next:

First, sort the values in an ascending order.

6.8	7.9	8.2	8.5	8.8	8.8	9.1	9.1	9.5	9.5

Remove the duplicates and consider only the unique values of the attribute.

6.8	7.9	8.2	8.5	8.8	8.8	9.1	9.5

Next:

	6.8		7.9		8.2		8.5		8.8		9.1	
Range	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>
Yes	0	7	0	7	1	6	2	5	4	3	5	2
No	1	2	2	1	2	1	2	1	2	1	3	0
Entropy	0	0.7637	0	0.5433	0.9177	0.5913	1	0.6497	0.9177	0.8108	0.9538	0
Entropy_Info (*S*, *T*)	0.6873		0.4346		0.6892		0.7898		0.8749		0.7630	
Gain	0.1935		**0.4462**		0.1916		0.091		0.0059		0.1178	

For a sample, the calculation

Finally

Table 6.14: Discretized Instances

S.No.	CGPA Continuous	CGPA Discretized	Job Offer
1.	9.5	>7.9	Yes
2.	8.2	>7.9	Yes
3.	9.1	>7.9	No
4.	6.8	≤7.9	No
5.	8.5	>7.9	Yes
6.	9.5	>7.9	Yes
7.	7.9	≤7.9	No
8.	9.1	>7.9	Yes
9.	8.8	>7.9	Yes
10.	8.8	>7.9	Yes

CART: Classification And Regression Tree

Key term to understand for CART:

1. Gini Index: Significance
2. Gini before split
3. Gini after split
4. Difference in Gini Index after split

Gini Index: Significance

Higher the GINI value, higher is the homogeneity of the data instances.

Gini_Index(T) is computed as given .

$$\text{Gini\_Index}(T) = 1 - \sum_{i=1}^{m} P_i^2$$

where,

P_i be the probability that a data instance or a tuple 'd' belongs to class C_i. It is computed as:

P_i = |No. of data instances belonging to class i|/|Total no of data instances in the training dataset T|

Gini before split

Gini_Index(T) is computed as given .

$$\text{Gini\_Index}(T) = 1 - \sum_{i=1}^{m} P_i^2$$

where,

P_i be the probability that a data instance or a tuple 'd' belongs to class C_i. It is computed as:

P_i = |No. of data instances belonging to class i|/|Total no of data instances in the training dataset T|

Gini after split

Gini_Index(T, A) is computed as given in Eq. (6.14).

$$\text{Gini\_Index}(T, A) = \frac{|S_1|}{|T|}\text{Gini}(S_1) + \frac{|S_2|}{|T|}\text{Gini}(S_2)$$

- Where S1 & S2 are the subset after split.
- The split with minimum Gini Index on Subset S1 & S2 is taken forward.

Difference in Gini Index after split

ΔGini is computed as given

$$\Delta\text{Gini}(A) = \text{Gini}(T) - \text{Gini}(T, A)$$

Let us solve one PRoblem

Example 6.3: Assess a student's performance during his course of study and predict whether a student will get a job offer or not in his final year of the course. The training dataset T consists of 10 data instances with attributes such as 'CGPA', 'Interactiveness', 'Practical Knowledge' and 'Communication Skills' as shown in Table 6.3. The target class attribute is the 'Job Offer'.

Table 6.3: Training Dataset T

S.No.	CGPA	Interactiveness	Practical Knowledge	Communication Skills	Job Offer
1.	≥9	Yes	Very good	Good	Yes
2.	≥8	No	Good	Moderate	Yes
3.	≥9	No	Average	Poor	No
4.	<8	No	Average	Good	No
5.	≥8	Yes	Good	Moderate	Yes
6.	≥9	Yes	Good	Moderate	Yes
7.	<8	Yes	Good	Poor	No
8.	≥9	No	Very good	Good	Yes
9.	≥8	Yes	Good	Good	Yes
10.	≥8	Yes	Average	Good	Yes

Total Gini before applying any split:

$$\text{Gini\_Index}(T) = 1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2$$

$$= 1 - 0.49 - 0.09$$

$$= 1 - 0.58$$

$$\text{Gini\_Index}(T) = 0.42$$

Let us first consider CGPA as split criterion:

Following the out come of CGPA:

CGPA	Job Offer = Yes	Job Offer = No
≥9	3	1
≥8	4	0
<8	0	2

With Three CGPA; there will be total 3 split criterions.

We need to calculate the Gini for all seven criterion and need to consider minimum one.

Gini Index of CGPA:

Table 6.16: Gini_Index of CGPA

Subsets		Gini_Index
(≥9, ≥8)	<8	0.1755
(≥9, <8)	≥8	0.3
(≥8, <8)	≥9	0.417

Similarly Gini Index on other

Table 6.22: Gini_Index and ΔGini for all Attributes

Attribute	Gini_Index	ΔGini
CGPA	0.1755	0.2445
Interactiveness	0.368	0.052
Practical knowledge	0.3054	0.1146
Communication Skills	0.1755	0.2445

Now,

CGPA or Communication can be taken as first criterion.

And the process is repeated till stopping criterion.

Thanks and Regards

Questions if any