Gain(Practical Knowledge) = 0.8108

Entropy_Info(T, Communication Skills)

$$= \frac{2}{4}\left[-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}\right] + \frac{1}{4}\left[-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right] + \frac{1}{4}\left[-\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1}\right]$$

$= 0$

Gain(Communication Skills) = 0.8108

The gain calculated for all the attributes is shown in Table 6.9.

Table 6.9: Total Gain

Attributes	Gain
Interactiveness	0.3111
Practical Knowledge	**0.8108**
Communication Skills	**0.8108**

Here, both the attributes 'Practical Knowledge' and 'Communication Skills' have the same Gain. So, we can either construct the decision tree using 'Practical Knowledge' or 'Communication Skills'. The final decision tree is shown in Figure 6.4.

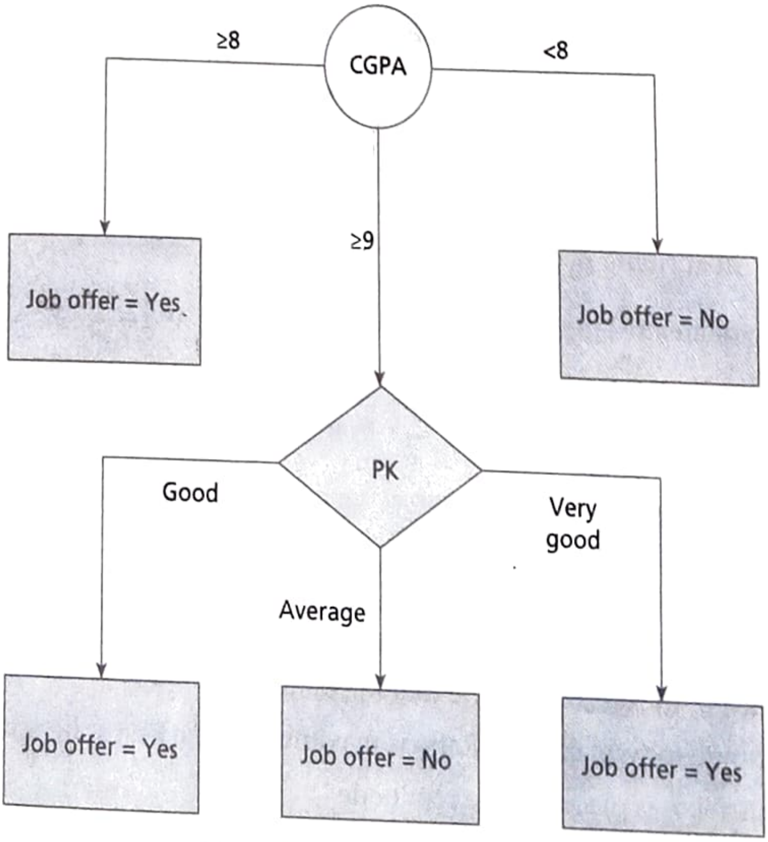


Figure 6.4: Final Decision Tree

6.2.2 C4.5 Construction

C4.5 is an improvement over ID3. C4.5 works with continuous and discrete attributes and missing values, and it also supports post-pruning. C5.0 is the successor of C4.5 and is more efficient and used for building smaller decision trees. C4.5 works with missing values by marking as '?', but these missing attribute values are not considered in the calculations.

The algorithm C4.5 is based on Occam's Razor which says that given two correct solutions, the simpler solution has to be chosen. Moreover, the algorithm requires a larger training set for better accuracy. It uses Gain Ratio as a measure during the construction of decision trees. ID3 is more biased towards attributes with larger values. For example, if there is an attribute called 'Register No' for students it would be unique for every student and will have distinct value for every data instance resulting in more values for the attribute. Hence, every instance belongs to a category and would have higher Information Gain than other attributes. To overcome this bias issue, C4.5 uses a purity measure Gain ratio to identify the best split attribute. In C4.5 algorithm, the Information Gain measure used in ID3 algorithm is normalized by computing another factor called Split_Info. This normalized information gain of an attribute called as Gain_Ratio is computed by the ratio of the calculated Split_Info and Information Gain of each attribute. Then, the attribute with the highest normalized information gain, that is, highest gain ratio is used as the splitting criteria.

As an example, we will choose the same training dataset shown in Table 6.3 to construct a decision tree using the C4.5 algorithm.

Given a Training dataset *T*,

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

$$\text{Split\_Info}(T, A) = -\sum_{i=1}^{v} \frac{|A_i|}{|T|} \times \log_2 \frac{|A_i|}{|T|} \tag{6.11}$$

where, the attribute *A* has got '*v*' distinct values $\{a_1, a_2, \ldots a_v\}$, and $|A_i|$ is the number of instances for distinct value '*i*' in attribute *A*.

The Gain_Ratio of an attribute *A* is computed as given in Eq. (6.12):

$$\text{Gain\_Ratio}(A) = \frac{\text{Info\_Gain}(A)}{\text{Split\_Info}(T, A)} \tag{6.12}$$

Algorithm 6.3: Procedure to Construct a Decision Tree using C4.5

1. Compute Entropy_Info Eq. (6.8) for the whole training dataset based on the target attribute.
2. Compute Entropy_Info Eq. (6.9), Info_Gain Eq. (6.10), Split_Info Eq. (6.11) and Gain_Ratio Eq. (6.12) for each of the attribute in the training dataset.
3. Choose the attribute for which Gain_Ratio is maximum as the best split attribute.
4. The best split attribute is placed as the root node.
5. The root node is branched into subtrees with each subtree as an outcome of the test condition of the root node attribute. Accordingly, the training dataset is also split into subsets.
6. Recursively apply the same operation for the subset of the training set with the remaining attributes until a leaf node is derived or no more training instances are available in the subset.

Example 6.4: Make use of Information Gain of the attributes which are calculated in ID3 algorithm in Example 6.3 to construct a decision tree using C4.5.

Solution:

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:

$$\text{Entropy Info}(T, \text{CGPA}) = \frac{4}{10}\left[-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right] + \frac{4}{10}\left[-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right] + \frac{2}{10}\left[-\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2}\right]$$
$$= \frac{4}{10}(0.3111 + 0.4997) + 0 + 0$$
$$= 0.3243$$

$$\text{Gain(CGPA)} = 0.8807 - 0.3243$$
$$= 0.5564$$

$$\text{Split\_Info}(T, \text{CGPA}) = -\frac{4}{10}\log_2\frac{4}{10} - \frac{4}{10}\log_2\frac{4}{10} - \frac{2}{10}\log_2\frac{2}{10}$$
$$= 0.5285 + 0.5285 + 0.4641$$
$$= 1.5211$$

$$\text{Gain Ratio(CGPA)} = (\text{Gain(CGPA)})/(\text{Split\_Info}(T, \text{CGPA}))$$
$$= \frac{0.5564}{1.5211} = 0.3658$$

Interactiveness:

$$\text{Entropy Info}(T, \text{Interactiveness}) = \frac{6}{10}\left[-\frac{5}{6}\log_2\frac{5}{6} - \frac{1}{6}\log_2\frac{1}{6}\right] + \frac{4}{10}\left[-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right]$$
$$= \frac{6}{10}(0.2191 + 0.4306) + \frac{4}{10}(0.4997 + 0.4997)$$
$$= 0.3898 + 0.3998 = 0.7896$$

$$\text{Gain(Interactiveness)} = 0.8807 - 0.7896 = 0.0911$$

$$\text{Gain(Interactiveness)} = -\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10} = 0.9704$$

$$\text{Gain\_Ratio(Interactiveness)} = \frac{\text{Gain(Interactiveness)}}{\text{Split\_Info}(T, \text{ Interactiveness})}$$

$$= \frac{0.0911}{0.9704}$$

$$= 0.0939$$

Practical Knowledge:

$$\text{Entropy\_Info}(T, \text{Practical Knowledge}) = \frac{2}{10}\left[-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}\right] + \frac{3}{10}\left[-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right] + \frac{5}{10}\left[-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}\right]$$

$$= \frac{2}{10}(0) + \frac{3}{10}(0.5280 + 0.3897) + \frac{5}{10}(0.2574 + 0.4641)$$

$$= 0 + 0.2753 + 0.3608 = 0.6361$$

$$\text{Gain(Practical Knowledge)} = 0.8807 - 0.6361$$

$$= 0.2448$$

$$\text{Split\_Info}(T, \text{Practical Knowledge}) = -\frac{2}{10}\log_2\frac{2}{10} - \frac{5}{10}\log_2\frac{5}{10} - \frac{3}{10}\log_2\frac{3}{10}$$

$$= 1.4853$$

$$\text{Gain\_Ratio(Practical Knowledge)} = \frac{\text{Gain(Practical Knowledge)}}{\text{Split\_Info}(T, \text{ Practical Knowledge})}$$

$$= \frac{0.2448}{1.4853}$$

$$= 0.1648$$

Communication Skills:

$$\text{Entropy\_Info}(T, \text{Communication Skills}) = \frac{5}{10}\left[-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}\right] + \frac{3}{10}\left[-\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}\right] + \frac{2}{10}\left[-\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2}\right]$$

$$= \frac{5}{10}(0.5280 + 0.3897) + \frac{3}{10}(0) + \frac{2}{10}(0)$$

$$= 0.3609$$

$$\text{Gain(Communication Skills)} = 0.8813 - 0.36096$$

$$= 0.5202$$

$$\text{Split\_Info}(T, \text{ Communication Skills}) = -\frac{5}{10}\log_2\frac{5}{10} - \frac{3}{10}\log_2\frac{3}{10} - \frac{2}{10}\log_2\frac{2}{10}$$

$$= 1.4853$$

$$\text{Gain\_Ratio(Communication Skills)} = \frac{\text{Gain(Communication Skills)}}{\text{Split\_Info}(T, \text{ Communication Skills})}$$

$$= \frac{0.5202}{1.4853} = 0.3502$$

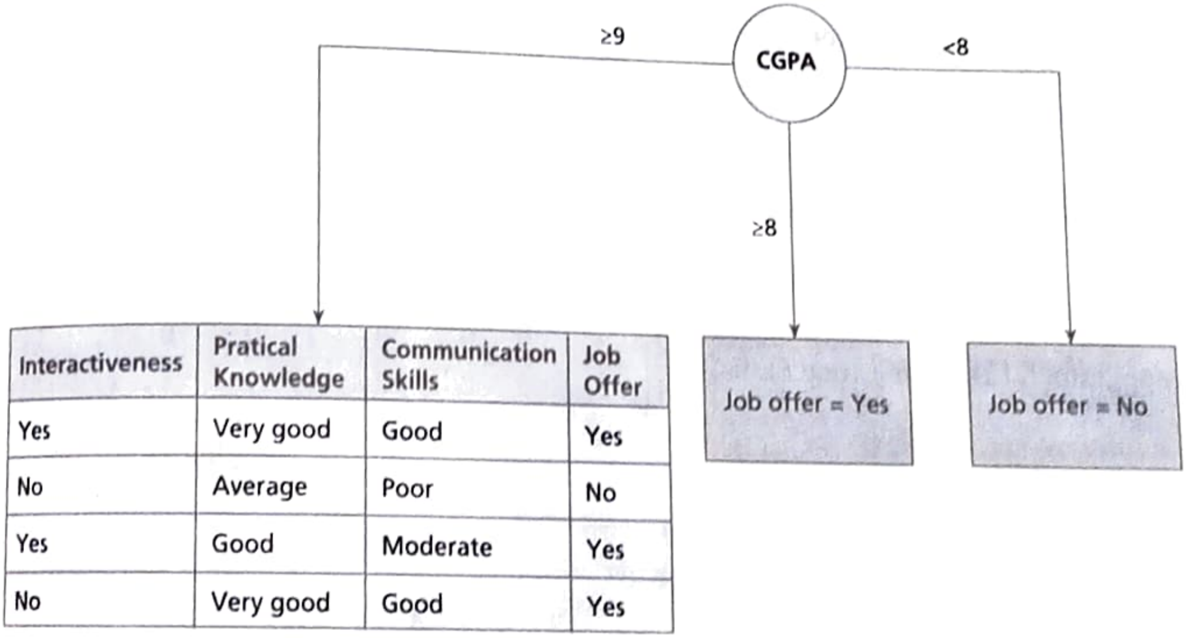
Table 6.10 shows the Gain_Ratio computed for all the attributes.

Table 6.10: Gain_Ratio

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

Step 3: Choose the attribute for which Gain_Ratio is maximum as the best split attribute.

From Table 6.10, we can see that CGPA has highest gain ratio and it is selected as the best split attribute. We can construct the decision tree placing CGPA as the root node shown in Figure 6.5. The training dataset is split into subsets with 4 data instances.



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

Iteration 2:

Total Samples: 4

Repeat the same process for this resultant dataset with 4 data instances.

Job Offer has 3 instances as Yes and 1 instance as No.

$$\text{Entropy\_Info(Target Class = Job Offer)} = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$$

$$= 0.3112 + 0.5$$

$$= 0.8112$$

Interactiveness:

$$\text{Entropy\_Info}(T, \text{Interactiveness}) = \frac{2}{4}\left[-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}\right] + \frac{2}{4}\left[-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right]$$

$$= 0 + 0.4997$$

$$\text{Gain(Interactiveness)} = 0.8108 - 0.4997 = 0.3111$$

$$\text{Split\_Info}(T, \text{Interactiveness}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 0.5 + 0.5 = 1$$

$$\text{Gain\_Ratio(Interactiveness)} = \frac{\text{Gain(Interactiveness)}}{\text{Split\_Info}(T, \text{Interactiveness})}$$

$$= \frac{0.3112}{1} = 0.3112$$

Practical Knowledge:

$$\text{Entropy\_Info}(T, \text{Practical Knowledge}) = \frac{2}{4}\left[-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}\right] + \frac{1}{4}\left[-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right]$$
$$+ \frac{1}{4}\left[-\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1}\right]$$

$$= 0$$

$$\text{Gain(Practical Knowledge)} = 0.8108$$

$$\text{Split\_Info}(T, \text{Practical Knowledge}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} = 1.5$$

$$\text{Gain\_Ratio(Practical Knowledge)} = \frac{\text{Gain(Practical Knowledge)}}{\text{Split\_Info}(T, \text{Practical Knowledge})} = \frac{0.8108}{1.5} = 0.5408$$

Communication Skills:

$$\text{Entropy\_Info}(T, \text{Communication Skills}) = \frac{2}{4}\left[-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}\right] + \frac{1}{4}\left[-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right]$$
$$+ \frac{1}{4}\left[-\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1}\right]$$

$$= 0$$

$$\text{Gain(Communication Skills)} = 0.8108$$

$$\text{Split\_Info}(T, \text{Communication Skills}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} = 1.5$$

$$\text{Gain\_Ratio(Communication Skills)} = \frac{\text{Gain(Practical Knowledge)}}{\text{Split\_Info}(T, \text{Practical Knowledge})} = \frac{0.8108}{1.5} = 0.5408$$

Table 6.11 shows the Gain_Ratio computed for all the attributes.

Table 6.11: Gain-Ratio

Attributes	Gain_Ratio
Interactiveness	0.3112
Practical Knowledge	0.5408
Communication Skills	0.5408

Both 'Practical Knowledge' and 'Communication Skills' have the highest gain ratio. So, the best splitting attribute can either be 'Practical Knowledge' or 'Communication Skills', and therefore, the split can be based on any one of these.

Here, we split based on 'Practical Knowledge'. The final decision tree is shown in Figure 6.6.

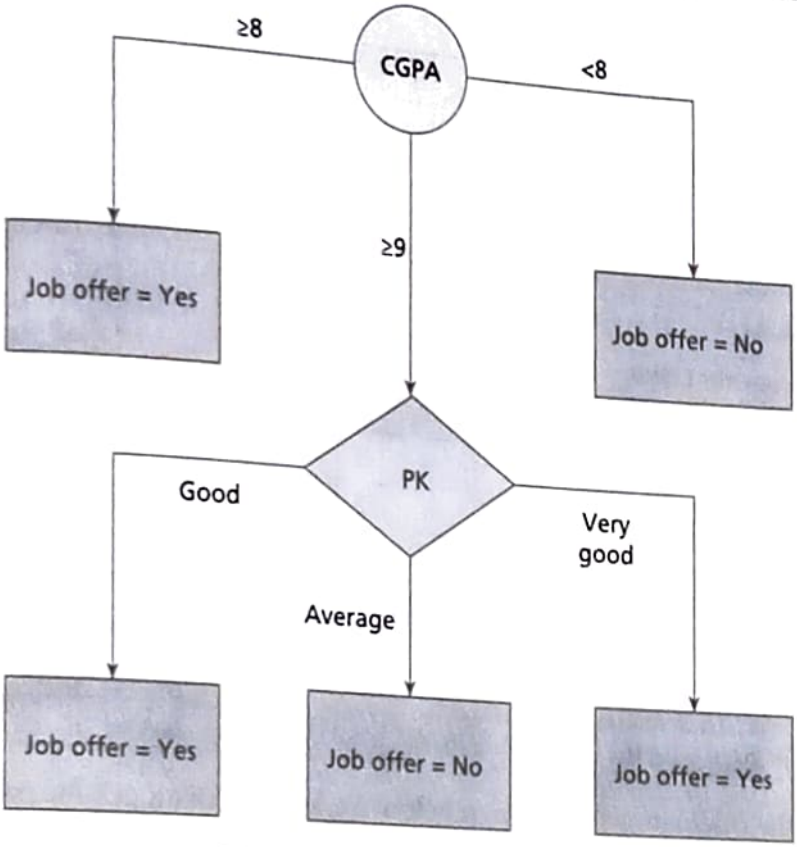


Figure 6.6: Final Decision Tree

Dealing with Continuous Attributes in C4.5

The C4.5 algorithm is further improved by considering attributes which are continuous, and a continuous attribute is discretized by finding a split point or threshold. When an attribute '*A*' has numerical values which are continuous, a threshold or best split point '*s*' is found such that the set of values is categorized into two sets such as $A < s$ and $A \geq s$. The best split point is the attribute value which has maximum information gain for that attribute.

Now, let us consider the set of continuous values for the attribute CGPA in the sample dataset as shown in Table 6.12.

Table 6.12: 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

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

Now, compute the Gain for the distinct values of this continuous attribute. Table 6.13 shows the computed values.

Table 6.13: Gain Values for CGPA

	6.8		7.9		8.2		8.5		8.8		9.1		9.5	
Range	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>
Yes	0	7	0	7	1	6	2	5	4	3	5	2	7	0
No	1	2	2	1	2	1	2	1	2	1	3	0	3	0
Entropy	0	0.7637	0	0.5433	0.9177	0.5913	1	0.6497	0:9177	0.8108	0.9538	0	0.8808	0
Entropy_Info (*S*, *T*)	0.6873		0.4346		0.6892		0.7898		0.8749		0.7630		0.8808	
Gain	0.1935		**0.4462**		0.1916		0.091		0.0059		0.1178		0	

For a sample, the calculations are shown below for a single distinct value say, CGPA ∈ 6.8.

$$\text{Entropy\_Info}(T, \text{Job\_Offer}) = -\left[\frac{7}{10}\log_2\frac{7}{10} + \frac{3}{10}\log_2\frac{3}{10}\right]$$

$$= -(-0.3599 + -0.5209)$$

$$= 0.8808$$

$$\text{Entropy}(7, 2) = -\left[\frac{7}{9}\log_2\frac{7}{9} + \frac{2}{9}\log_2\frac{2}{9}\right]$$

$$= -(-0.2818 + -0.4819)$$

$$= 0.7637$$

$$\text{Entropy\_Info}(T, \text{CGPA} \in 6.8) = \frac{1}{10} \times \text{Entropy}(0, 1) + \frac{9}{10}\,\text{Entropy}(7, 2)$$

$$= \frac{1}{10}\left[-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right] + \frac{9}{10}\left[-\frac{7}{9}\log_2\frac{7}{9} - \frac{2}{9}\log_2\frac{2}{9}\right]$$

$$= 0 + \frac{9}{10}(0.7637)$$

$$= 0.6873$$

$$\text{Gain}(\text{CGPA} \in 6.8) = 0.8808 - 0.6873$$

$$= 0.1935$$

Similarly, the calculations are done for each of the distinct value for the attribute CGPA and a table is created. Now, the value of CGPA with maximum gain is chosen as the threshold value or the best split point. From Table 6.13, we can observe that CGPA with 7.9 has the maximum gain as 0.4462. Hence, CGPA ∈ 7.9 is chosen as the split point. Now, we can dicretize the continuous values of CGPA as two categories with CGPA ≤ 7.9 and CGPA > 7.9. The resulting discretized instances are shown in Table 6.14.

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

6.2.3 Classification and Regression Trees Construction

The Classification and Regression Trees (CART) algorithm is a multivariate decision tree learning used for classifying both categorical and continuous-valued target variables. CART algorithm is an example of multivariate decision trees that gives oblique splits. It solves both classification and regression problems. If the target feature is categorical, it constructs a classification tree and if the target feature is continuous, it constructs a regression tree. CART uses GINI Index to construct a decision tree. GINI Index is defined as the number of data instances for a class or it is the proportion of instances. It constructs the tree as a binary tree by recursively splitting a node into two nodes. Therefore, even if an attribute has more than two possible values, GINI Index is calculated for all subsets of the attributes and the subset which has maximum value is selected as the best split subset. For example, if an attribute A has three distinct values say $\{a_1, a_2, a_3\}$, the possible subsets are $\{\ \}$, $\{a_1\}$, $\{a_2\}$, $\{a_3\}$, $\{a_1, a_2\}$, $\{a_1, a_3\}$, $\{a_2, a_3\}$, and $\{a_1, a_2, a_3\}$. So, if an attribute has 3 distinct values, the number of possible subsets is 2^3, which means 8. Excluding the empty set $\{\ \}$ and the full set $\{a_1, a_2, a_3\}$, we have 6 subsets. With 6 subsets, we can form three possible combinations such as:

$\{a_1\}$ with $\{a_2, a_3\}$

$\{a_2\}$ with $\{a_1, a_3\}$

$\{a_3\}$ with $\{a_1, a_2\}$

Hence, in this CART algorithm, we need to compute the best splitting attribute and the best split subset i in the chosen attribute.

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

Gini_Index(T) is computed as given in Eq. (6.13).

$$\text{Gini\_Index}(T) = 1 - \sum_{i=1}^{m} P_i^2 \tag{6.13}$$

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 Index assumes a binary split on each attribute, therefore, every attribute is considered as a binary attribute which splits the data instances into two subsets S_1 and S_2.