

# Chapter 6 Decision Tree Learning



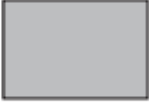
# Decision Tree Learning Model

Decision tree learning model, one of the most popular supervised predictive learning models, classifies data instances with high accuracy and consistency. The model performs an *inductive inference* that reaches a general conclusion from observed examples. This model is variably used for solving complex classification applications.

Decision tree is a concept tree which summarizes the information contained in the training dataset in the form of a tree structure. Once the concept model is built, test data can be easily classified.

1. Model will classify both categorical and continuous valued target variables.
2. For a given training set  $X$ , hypothesis function  $f(X)$  as a decision tree.
3. The Decision tree generated a complete hypothesis space as a tree and allow us to search through the possible set of hypothesis.

# Structure of a Decision Tree

|   |               |
|---|---------------|
|  | Root node     |
|  | Decision node |
|  | Leaf node     |

**Figure 6.1:** Nodes in a Decision Tree

**Note:** *A decision tree is not always a binary tree. It is a tree which can have more than two branches.*

- 1.It is a structure with root node ,internal node/decision nodes, branches and leaf nodes.
- 2.Topmost tree node is root node
- 3.Internal nodes are test nodes and also decision nodes
- 4.These nodes represent a choice or test an input attribute and the outputs for a test condition.
- 5.The branches are labelled as per the outcomes or output values of the test condition.
- 6.Each decision is a path to a leaf node .
- 7.The leaf nodes represent the label or outcomes of a decision path.
- 8.Every path from root to leaf node represents the logical rule corresponds to a conjunction of test attributes and the whole tree represents the disjunction of these conjunctions

# Structure of a Decision Tree

## ***BUILDING THE TREE***

**Goal** Construct a decision tree with the given training dataset. The tree is constructed in a top-down fashion. It starts from the root node. At every level of tree construction, we need to find the best split attribute or best decision node among all attributes. This process is recursive and continued until we end up in the last level of the tree or finding a leaf node which cannot be split further. The tree construction is complete when all the test conditions lead to a leaf node. The leaf node contains the target class or output of classification.

**Output** Decision tree representing the complete hypothesis space.

## ***KNOWLEDGE INFERENCE OR CLASSIFICATION***

**Goal** Given a test instance, infer to the target class it belongs to.

**Classification** Inferring the target class for the test instance or object is based on inductive inference on the constructed decision tree. In order to classify an object, we need to start traversing the tree from the root. We traverse as we evaluate the test condition on every decision node with the test object attribute value and walk to the branch corresponding to the test's outcome. This process is repeated until we end up in a leaf node which contains the target class of the test object.

**Output** Target label of the test instance.

# Advantages of Decision Trees

1. Easy to model and interpret
2. Simple to understand
3. The input and output attributes can be discrete or continuous predictor variables.
4. Can model a high degree of nonlinearity in the relationship between the target variables and the predictor variables
5. Quick to train

# Disadvantages of Decision Trees

Some of the issues that generally arise with a decision tree learning are that:

1. It is difficult to determine how deeply a decision tree can be grown or when to stop growing it.
2. If training data has errors or missing attribute values, then the decision tree constructed may become unstable or biased.
3. If the training data has continuous valued attributes, handling it is computationally complex and has to be discretized.
4. A complex decision tree may also be over-fitting with the training data.
5. Decision tree learning is not well suited for classifying multiple output classes.
6. Learning an optimal decision tree is also known to be NP-complete.

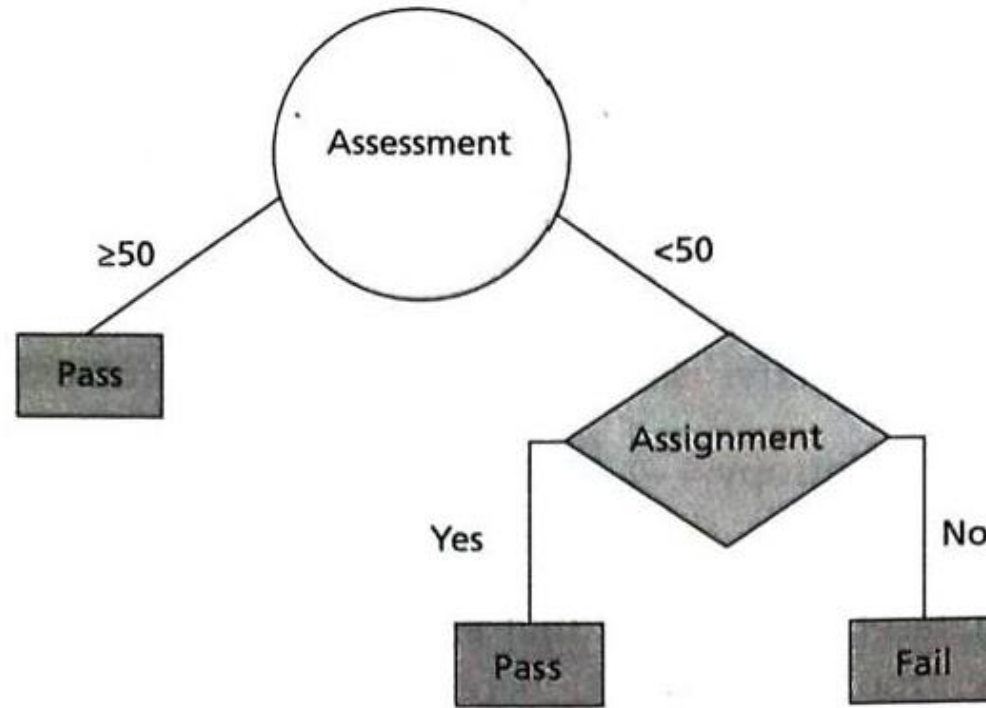


**Example 6.2:** Predict a student's academic performance of whether he will pass or fail based on the given information such as 'Assessment' and 'Assignment'. The following Table 6.2 shows the independent variables, Assessment and Assignment, and the target variable Exam Result with their values. Draw a binary decision tree.

**Table 6.2:** Attributes and Associated Values

| Attributes  | Values             |
|-------------|--------------------|
| Assessment  | $\geq 50$ , $< 50$ |
| Assignment  | Yes, No            |
| Exam Result | Pass, Fail         |

**Solution:** Consider the root node is 'Assessment'. If a student's marks are  $\geq 50$ , the root node is branched to leaf node 'Pass' and if the assessment marks are  $< 50$ , it is branched to another decision node. If the decision node in next level of the tree is 'Assignment' and if a student has submitted his assignment, the node branches to 'Pass' and if not submitted, the node branches to 'Fail'. Figure 6.2 depicts this rule.



**Figure 6.2:** Illustration of a Decision Tree

This tree can be interpreted as a sequence of logical rules as follows:

if (Assessment  $\geq 50$ ) then 'Pass'

else if (Assessment  $< 50$ ) then

    if (Assignment = Yes) then 'Pass'

    else if (Assignment = No) then 'Fail'

Now, if a test instance is given, such as a student has scored 42 marks in his assessment and has not submitted his assignment, then it is predicted with the decision tree that his exam result is 'Fail'.

---

# Fundamentals of Entropy

Entropy is the amount of uncertainty or randomness in the outcome of a random variable or an event. Moreover, entropy describes about the homogeneity of the data instances. The best feature is selected based on the entropy value. For example, when a coin is flipped, head or tail are the two outcomes, hence its entropy is lower when compared to rolling a dice which has got six outcomes.

Higher the entropy → Higher the uncertainty

Lower the entropy → Lower the uncertainty

Let  $P$  be the probability distribution of data instances from 1 to  $n$  as shown in Eq. (6.2).

$$\text{So, } P = P_1 \dots\dots\dots P_n \quad (6.2)$$

Entropy of  $P$  is the information measure of this probability distribution given in Eq. (6.3),

$$\begin{aligned} \text{Entropy\_Info}(P) &= \text{Entropy\_Info}(P_1 \dots\dots P_n) \\ &= -(P_1 \log_2(P_1) + P_2 \log_2(P_2) + \dots\dots\dots + P_n \log_2(P_n)) \end{aligned} \quad (6.3)$$

where,  $P_1$  is the probability of data instances classified as class 1 and  $P_2$  is the probability of data instances classified as class 2 and so on.

$$P_1 = |\text{No of data instances belonging to class 1}| / |\text{Total no of data instances in the training dataset}|$$

# General Algorithm for Decision Trees

## Algorithm 6.1: General Algorithm for Decision Trees

1. Find the best attribute from the training dataset using an attribute selection measure and place it at the root of the tree.
2. Split the training dataset into subsets based on the outcomes of the test attribute and each subset in a branch contains the data instances or tuples with the same value for the selected test attribute.
3. Repeat step 1 and step 2 on each subset until we end up in leaf nodes in all the branches of the tree.
4. This splitting process is recursive until the stopping criterion is reached.

### ***Stopping Criteria***

The following are some of the common stopping conditions:

1. The data instances are homogenous which means all belong to the same class  $C_i$  and hence its entropy is 0.
2. A node with some defined minimum number of data instances becomes a leaf (Number of data instances in a node is between 0.25 and 1.00% of the full training dataset).
3. The maximum tree depth is reached, so further splitting is not done and the node becomes a leaf node.



# Decision Tree Induction Algorithms

There are many decision tree algorithms, such as ID3, C4.5, CART, CHAID, QUEST, GUIDE, CRUISE, and CTREE, that are used for classification in real-time environment. The most commonly used decision tree algorithms are ID3 (Iterative Dichotomizer 3), developed by J.R Quinlan in 1986, and C4.5 is an advancement of ID3 presented by the same author in 1993. CART, that stands for Classification and Regression Trees, is another algorithm which was developed by Breiman et al. in 1984.

The accuracy of the tree constructed depends upon the selection of the best split attribute. Different algorithms are used for building decision trees which use different measures to decide on the splitting criterion. Algorithms such as ID3, C4.5 and CART are popular algorithms used in the construction of decision trees. The algorithm ID3 uses 'Information Gain' as the splitting criterion whereas the algorithm C4.5 uses 'Gain Ratio' as the splitting criterion. The CART algorithm is popularly used for classifying both categorical and continuous-valued target variables. CART uses GINI Index to construct a decision tree.

# ID3 Tree Construction

- 1.ID3 is an example of univariate decision tree with only one feature at each decision node →axis aligned splits.
- 2.It constructs the tree using greedy approach in a top down fashion by identifying the best attribute of each level in the tree.
- 3.Works fine for discrete, otherwise it has to partitioned attributes to discrete features

## Algorithm 6.2: Procedure to Construct a Decision Tree using ID3

1. Compute Entropy\_Info Eq. (6.8) for the whole training dataset based on the target attribute.
2. Compute Entropy\_Info Eq. (6.9) and Information\_Gain Eq. (6.10) for each of the attribute in the training dataset.
3. Choose the attribute for which entropy is minimum and therefore the gain 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.

1. The algorithm builds a tree using pruning measure “information Gain” with given training dataset and constructed tree to classify test data.
2. Works fine for larger dataset.
3. Small dataset overfitting
4. Not accurate if missing values.
5. C4.5 and CART handle discrete and continuous values
6. . C4.5 and CART handle missing data as well.
7. C4.5 is prone to outliers and CART handle it well.

## Definitions

Let  $T$  be the training dataset.

Let  $A$  be the set of attributes  $A = \{A_1, A_2, A_3, \dots, A_n\}$ .

Let  $m$  be the number of classes in the training dataset.

Let  $P_i$  be the probability that a data instance or a tuple ' $d$ ' belongs to class  $C_i$ .

It is calculated as,

$$P_i = \text{Total no of data instances that belongs to class } C_i \text{ in } T / \text{Total no of tuples in the training set } T \quad (6.6)$$

Mathematically, it is represented as shown in Eq. (6.7).

$$P_i = \frac{|d_{C_i}|}{|T|} \quad (6.7)$$

Expected information or Entropy needed to classify a data instance  $d'$  in  $T$  is denoted as Entropy\_Info( $T$ ) given in Eq. (6.8).

$$\text{Entropy\_Info}(T) = - \sum_{i=1}^m P_i \log_2 P_i \quad (6.8)$$

Entropy of every attribute denoted as Entropy\_Info( $T, A$ ) is shown in Eq. (6.9) as:

$$\text{Entropy\_Info}(T, A) = \sum_{i=1}^v \frac{|A_i|}{|T|} \times \text{Entropy\_Info}(A_i) \quad (6.9)$$

where, the attribute  $A$  has got ' $v$ ' distinct values  $\{a_1, a_2, \dots, a_v\}$ ,  $|A_i|$  is the number of instances for distinct value ' $i$ ' in attribute  $A$ , and Entropy\_Info( $A_i$ ) is the entropy for that set of instances.

Information\_Gain is a metric that measures how much information is gained by branching on an attribute  $A$ . In other words, it measures the reduction in impurity in an arbitrary subset of data.

It is calculated as given in Eq. (6.10):

$$\text{Information\_Gain}(A) = \text{Entropy\_Info}(T) - \text{Entropy\_Info}(T, A) \quad (6.10)$$

It can be noted that as entropy increases, information gain decreases. They are inversely proportional to each other.



**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.    | $\geq 9$ | Yes             | Very good           | Good                 | Yes       |
| 2.    | $\geq 8$ | No              | Good                | Moderate             | Yes       |
| 3.    | $\geq 9$ | No              | Average             | Poor                 | No        |
| 4.    | $< 8$    | No              | Average             | Good                 | No        |
| 5.    | $\geq 8$ | Yes             | Good                | Moderate             | Yes       |
| 6.    | $\geq 9$ | Yes             | Good                | Moderate             | Yes       |
| 7.    | $< 8$    | Yes             | Good                | Poor                 | No        |
| 8.    | $\geq 9$ | No              | Very good           | Good                 | Yes       |
| 9.    | $\geq 8$ | Yes             | Good                | Good                 | Yes       |
| 10.   | $\geq 8$ | Yes             | Average             | Good                 | Yes       |

**Solution:**

**Step 1:**

Calculate the Entropy for the target class 'Job Offer'.

$$\begin{aligned}\text{Entropy\_Info}(\text{Target Attribute} = \text{Job Offer}) &= \text{Entropy\_Info}(7, 3) = \\ &= -\left[\frac{7}{10} \log_2 \frac{7}{10} + \frac{3}{10} \log_2 \frac{3}{10}\right] = -(-0.3599 + -0.5208) = 0.8807\end{aligned}$$

**Iteration 1:**

**Step 2:**

Calculate the Entropy\_Info and Gain(Information\_Gain) for each of the attribute in the training dataset.

Table 6.4 shows the number of data instances classified with Job Offer as Yes or No for the attribute CGPA.

**Table 6.4: Entropy Information for CGPA**

| CGPA     | Job Offer = Yes | Job Offer = No | Total | Entropy |
|----------|-----------------|----------------|-------|---------|
| $\geq 9$ | 3               | 1              | 4     |         |
| $\geq 8$ | 4               | 0              | 4     | 0       |
| $< 8$    | 0               | 2              | 2     | 0       |

Entropy\_Info(T, CGPA)

$$\begin{aligned}&= \frac{4}{10} \left[ -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} \right] + \frac{4}{10} \left[ -\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4} \right] + \frac{2}{10} \left[ -\frac{0}{2} \log_2 \frac{0}{2} - \frac{2}{2} \log_2 \frac{2}{2} \right] \\ &= \frac{4}{10} (0.3111 + 0.4997) + 0 + 0 \\ &= 0.3243\end{aligned}$$

$$\begin{aligned}\text{Gain (CGPA)} &= 0.8807 - 0.3243 \\ &= 0.5564\end{aligned}$$

Table 6.5 shows the number of data instances classified with Job Offer as Yes or No for the attribute Interactiveness.

**Table 6.5: Entropy Information for Interactiveness**

| Interactiveness | Job Offer = Yes | Job Offer = No | Total | Entropy |
|-----------------|-----------------|----------------|-------|---------|
| YES             | 5               | 1              | 6     |         |
| NO              | 2               | 2              | 4     |         |

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

$$\begin{aligned}\text{Gain(Interactiveness)} &= 0.8807 - 0.7896 \\ &= 0.0911\end{aligned}$$

Table 6.6 shows the number of data instances classified with Job Offer as Yes or No for the attribute Practical Knowledge.

**Table 6.6: Entropy Information for Practical Knowledge**

| Practical Knowledge | Job Offer = Yes | Job Offer = No | Total | Entropy |
|---------------------|-----------------|----------------|-------|---------|
| Very Good           | 2               | 0              | 2     | 0       |
| Average             | 1               | 2              | 3     |         |
| Good                | 4               | 1              | 5     |         |

Entropy\_Info( $T$ , Practical Knowledge)

$$\begin{aligned}
 &= \frac{2}{10} \left[ -\frac{2}{2} \log_2 \frac{2}{2} - \frac{0}{2} \log_2 \frac{0}{2} \right] + \frac{3}{10} \left[ -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} \right] + \frac{5}{10} \left[ -\frac{4}{5} \log_2 \frac{4}{5} - \frac{1}{5} \log_2 \frac{1}{5} \right] \\
 &= \frac{2}{10}(0) + \frac{3}{10}(0.5280 + 0.3897) + \frac{5}{10}(0.2574 + 0.4641) \\
 &= 0 + 0.2753 + 0.3608 \\
 &= 0.6361
 \end{aligned}$$

$$\begin{aligned}
 \text{Gain(Practical Knowledge)} &= 0.8807 - 0.6361 \\
 &= 0.2446
 \end{aligned}$$

Table 6.7 shows the number of data instances classified with Job Offer as Yes or No for the attribute Communication Skills.

**Table 6.7: Entropy Information for Communication Skills**

| Communication Skills | Job Offer = Yes | Job Offer = No | Total |
|----------------------|-----------------|----------------|-------|
| Good                 | 4               | 1              | 5     |
| Moderate             | 3               | 0              | 3     |
| Poor                 | 0               | 2              | 2     |

Entropy\_Info( $T$ , Communication Skills)

$$\begin{aligned}
 &= \frac{5}{10} \left[ -\frac{4}{5} \log_2 \frac{4}{5} - \frac{1}{5} \log_2 \frac{1}{5} \right] + \frac{3}{10} \left[ -\frac{3}{3} \log_2 \frac{3}{3} - \frac{0}{3} \log_2 \frac{0}{3} \right] + \frac{2}{10} \left[ -\frac{0}{2} \log_2 \frac{0}{2} - \frac{2}{2} \log_2 \frac{2}{2} \right] \\
 &= \frac{5}{10}(0.5280 + 0.3897) + \frac{3}{10}(0) + \frac{2}{10}(0) \\
 &= 0.3609
 \end{aligned}$$

$$\begin{aligned}
 \text{Gain(Communication Skills)} &= 0.8813 - 0.36096 \\
 &= 0.5203
 \end{aligned}$$

The Gain calculated for all the attributes is shown in Table 6.8:

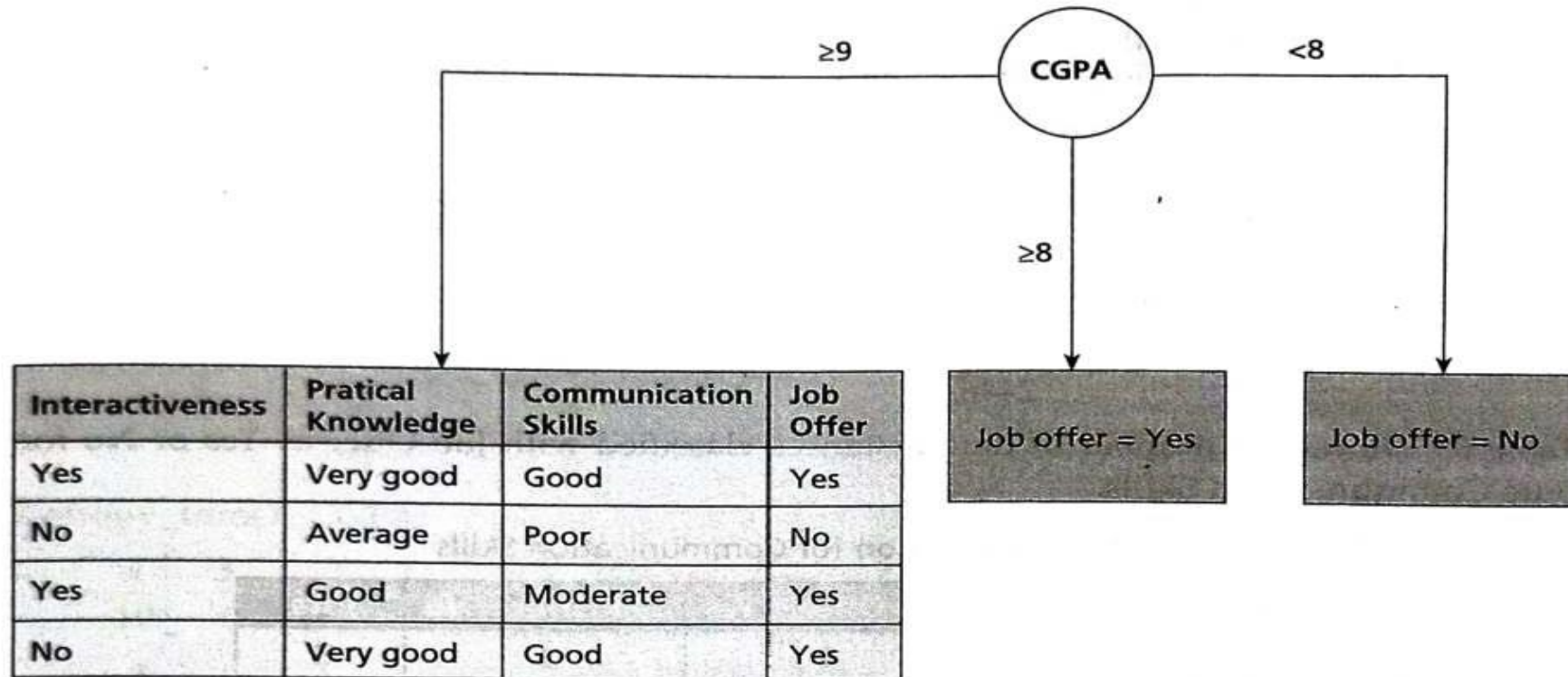
**Table 6.8: Gain**

| Attributes           | Gain   |
|----------------------|--------|
| CGPA                 | 0.5564 |
| Interactiveness      | 0.0911 |
| Practical Knowledge  | 0.2246 |
| Communication Skills | 0.5203 |



**Step 3:** From Table 6.8, choose the attribute for which entropy is minimum and therefore the gain is maximum as the best split attribute.

The best split attribute is CGPA since it has the maximum gain. So, we choose CGPA as the root node. There are three distinct values for CGPA with outcomes  $\geq 9$ ,  $\geq 8$  and  $< 8$ . The entropy value is 0 for  $\geq 8$  and  $< 8$  with all instances classified as Job Offer = Yes for  $\geq 8$  and Job Offer = No for  $< 8$ . Hence, both  $\geq 8$  and  $< 8$  end up in a leaf node. The tree grows with the subset of instances with CGPA  $\geq 9$  as shown in Figure 6.3.



**Figure 6.3:** Decision Tree After Iteration 1

Now, continue the same process for the subset of data instances branched with CGPA  $\geq 9$ .

**Iteration 2:**

In this iteration, the same process of computing the Entropy\_Info and Gain are repeated with the subset of training set. The subset consists of 4 data instances as shown in the above Figure 6.3.

$$\begin{aligned}\text{Entropy\_Info}(T) &= \text{Entropy\_Info}(3, 1) = \\ &= -\left[\frac{3}{4}\log_2 \frac{3}{4} + \frac{1}{4}\log_2 \frac{1}{4}\right] \\ &= -(-0.3111 + -0.4997) \\ &= 0.8108\end{aligned}$$

$$\begin{aligned}\text{Entropy\_Info}(T, \text{Interactiveness}) &= \frac{2}{4}\left[-\frac{2}{2}\log_2 \frac{2}{2} - \frac{0}{2}\log_2 \frac{0}{2}\right] + \frac{2}{4}\left[-\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{2}\log_2 \frac{1}{2}\right] \\ &= 0 + 0.4997\end{aligned}$$

$$\begin{aligned}\text{Gain}(\text{Interactiveness}) &= 0.8108 - 0.4997 \\ &= 0.3111\end{aligned}$$

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

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

$$\text{Entropy\_Info}(T, \text{Communication Skills})$$

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

$$= 0$$

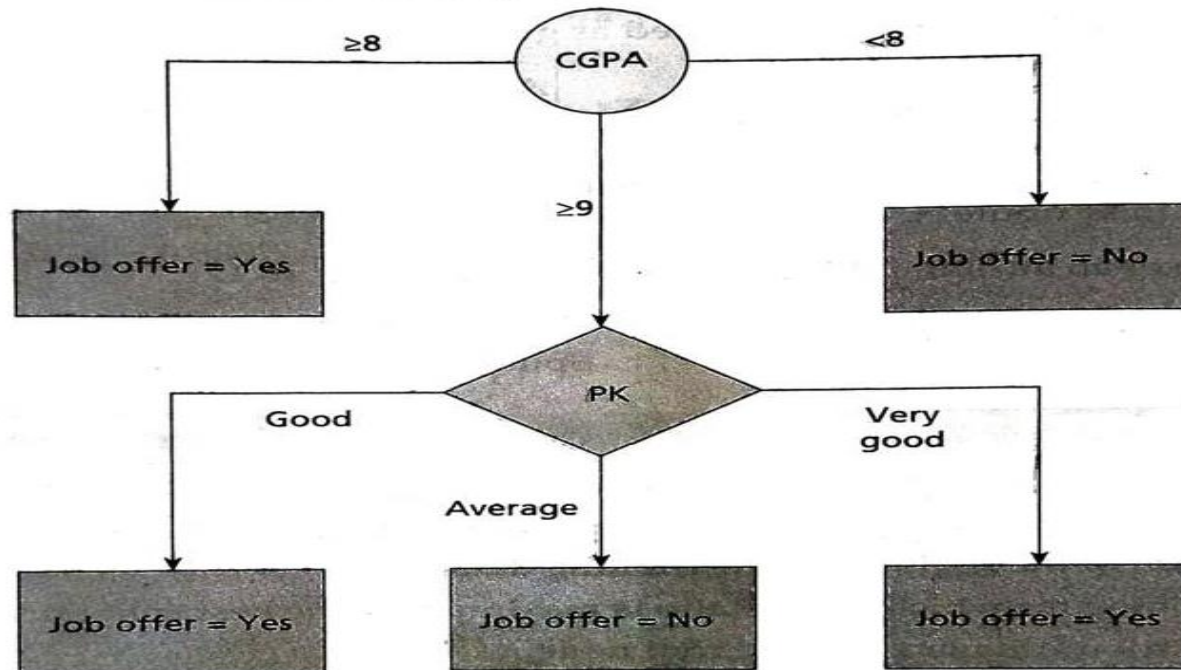
$$\text{Gain}(\text{Communication Skills}) = 0.8108$$

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**

# C4.5 Construction

1. Works with missing values by marking '?'
2. Builds C5.0
3. Occam's Razor problem, two correct solutions, the simple one is chosen.
4. Need larger dataset for accuracy.
5. Uses gain ratio as a measure during the construction of decision tree.

### 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.



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.

**Table 6.3:** Training Dataset *T*

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

Given a Training dataset  $T$ ,

The Split\_Info of an attribute  $A$  is computed as given in Eq. (6.11):

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

where, the attribute  $A$  has got ' $v$ ' distinct values  $\{a_1, a_2, \dots, a_v\}$ , 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):

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

---

**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'.

$$\begin{aligned}\text{Entropy\_Info}(\text{Target Attribute} = \text{Job Offer}) &= \text{Entropy\_Info}(7, 3) = \\ &= -\left[\frac{7}{10} \log_2 \frac{7}{10} + \frac{3}{10} \log_2 \frac{3}{10}\right] \\ &= (-0.3599 + -0.5208) \\ &= 0.8807\end{aligned}$$

**Step 2:** Calculate the Entropy\_Info, Gain(Info\_Gain), Split\_Info, Gain\_Ratio for each of the attri in the training dataset.

**CGPA:**

$$\begin{aligned}\text{Entropy Info}(T, \text{CGPA}) &= \frac{4}{10} \left[ -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} \right] + \frac{4}{10} \left[ -\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4} \right] \\ &\quad + \frac{2}{10} \left[ -\frac{0}{2} \log_2 \frac{0}{2} - \frac{2}{2} \log_2 \frac{2}{2} \right] \\ &= \frac{4}{10} (0.3111 + 0.4997) + 0 + 0 \\ &= 0.3243 \\ \text{Gain}(\text{CGPA}) &= 0.8807 - 0.3243 \\ &= 0.5564\end{aligned}$$

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

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

**Interactiveness:**

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

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

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

$$\begin{aligned}\text{Gain\_Ratio}(\text{Interactiveness}) &= \frac{\text{Gain}(\text{Interactiveness})}{\text{Split\_Info}(T, \text{Interactiveness})} \\ &= \frac{0.0911}{0.9704} \\ &= 0.0939\end{aligned}$$

$$= 0.0759$$

**Practical Knowledge:**

$$\begin{aligned} \text{Entropy\_Info}(T, \text{Practical Knowledge}) &= \frac{2}{10} \left[ -\frac{2}{2} \log_2 \frac{2}{2} - \frac{0}{2} \log_2 \frac{0}{2} \right] + \frac{3}{10} \left[ -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} \right] \\ &\quad + \frac{5}{10} \left[ -\frac{4}{5} \log_2 \frac{4}{5} - \frac{1}{5} \log_2 \frac{1}{5} \right] \\ &= \frac{2}{10}(0) + \frac{3}{10}(0.5280 + 0.3897) + \frac{5}{10}(0.2574 + 0.4641) \\ &= 0 + 0.2753 + 0.3608 = 0.6361 \end{aligned}$$

$$\begin{aligned} \text{Gain}(\text{Practical Knowledge}) &= 0.8807 - 0.6361 \\ &= 0.2448 \end{aligned}$$

$$\begin{aligned} \text{Split\_Info}(T, \text{Practical Knowledge}) &= -\frac{2}{10} \log_2 \frac{2}{10} - \frac{5}{10} \log_2 \frac{5}{10} - \frac{3}{10} \log_2 \frac{3}{10} \\ &= 1.4853 \end{aligned}$$

$$\begin{aligned} \text{Gain\_Ratio}(\text{Practical Knowledge}) &= \frac{\text{Gain}(\text{Practical Knowledge})}{\text{Split\_Info}(T, \text{Practical Knowledge})} \\ &= \frac{0.2448}{1.4853} \\ &= 0.1648 \end{aligned}$$



**Communication Skills:**

$$\begin{aligned}\text{Entropy\_Info}(T, \text{Communication Skills}) &= \frac{5}{10} \left[ -\frac{4}{5} \log_2 \frac{4}{5} - \frac{1}{5} \log_2 \frac{1}{5} \right] + \frac{3}{10} \left[ -\frac{3}{3} \log_2 \frac{3}{3} - \frac{0}{3} \log_2 \frac{0}{3} \right] \\ &\quad + \frac{2}{10} \left[ -\frac{0}{2} \log_2 \frac{0}{2} - \frac{2}{2} \log_2 \frac{2}{2} \right] \\ &= \frac{5}{10} (0.5280 + 0.3897) + \frac{3}{10} (0) + \frac{2}{10} (0) \\ &= 0.3609\end{aligned}$$

$$\begin{aligned}\text{Gain}(\text{Communication Skills}) &= 0.8813 - 0.36096 \\ &= 0.5202\end{aligned}$$

$$\begin{aligned}\text{Split\_Info}(T, \text{Communication Skills}) &= -\frac{5}{10} \log_2 \frac{5}{10} - \frac{3}{10} \log_2 \frac{3}{10} - \frac{2}{10} \log_2 \frac{2}{10} \\ &= 1.4853\end{aligned}$$

$$\begin{aligned}\text{Gain\_Ratio}(\text{Communication Skills}) &= \frac{\text{Gain}(\text{Communication Skills})}{\text{Split\_Info}(T, \text{Communication Skills})} \\ &= \frac{0.5202}{1.4853} = 0.3502\end{aligned}$$

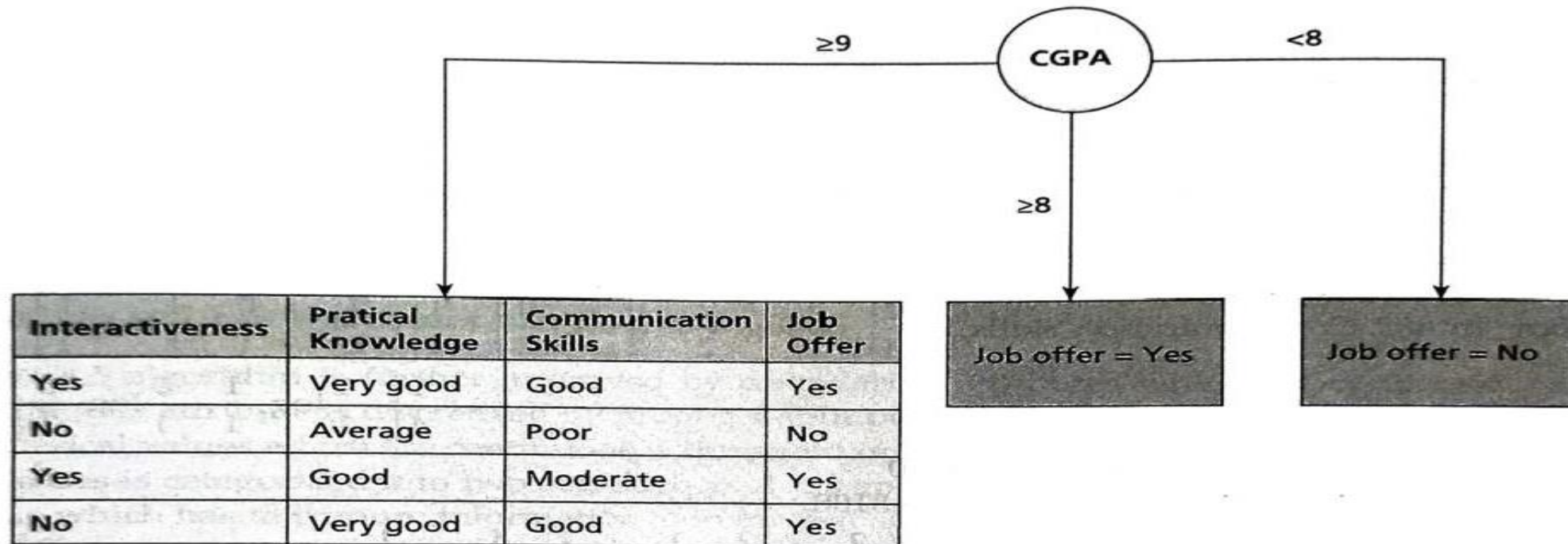
Table 6.10 shows the Gain\_Ratio computed for all the attributes.

**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.



**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.

$$\begin{aligned}
 \text{Entropy\_Info}(\text{Target Class} = \text{Job Offer}) &= -\frac{3}{4}\log_2 \frac{3}{4} - \frac{1}{4}\log_2 \frac{1}{4} \\
 &= 0.3112 + 0.5 \\
 &= 0.8112
 \end{aligned}$$

**Interactiveness:**

$$\begin{aligned}
 \text{Entropy\_Info}(T, \text{Interactiveness}) &= \frac{2}{4}\left[-\frac{2}{2}\log_2 \frac{2}{2} - \frac{0}{2}\log_2 \frac{0}{2}\right] + \frac{2}{4}\left[-\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{2}\log_2 \frac{1}{2}\right] \\
 &= 0 + 0.4997
 \end{aligned}$$

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

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

$$\begin{aligned}
 \text{Gain\_Ratio}(\text{Interactiveness}) &= \frac{\text{Gain}(\text{Interactiveness})}{\text{Split\_Info}(T, \text{Interactiveness})} \\
 &= \frac{0.3112}{1} = 0.3112
 \end{aligned}$$

**Practical Knowledge:**

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

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

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

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



### Communication Skills:

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

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

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

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

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.

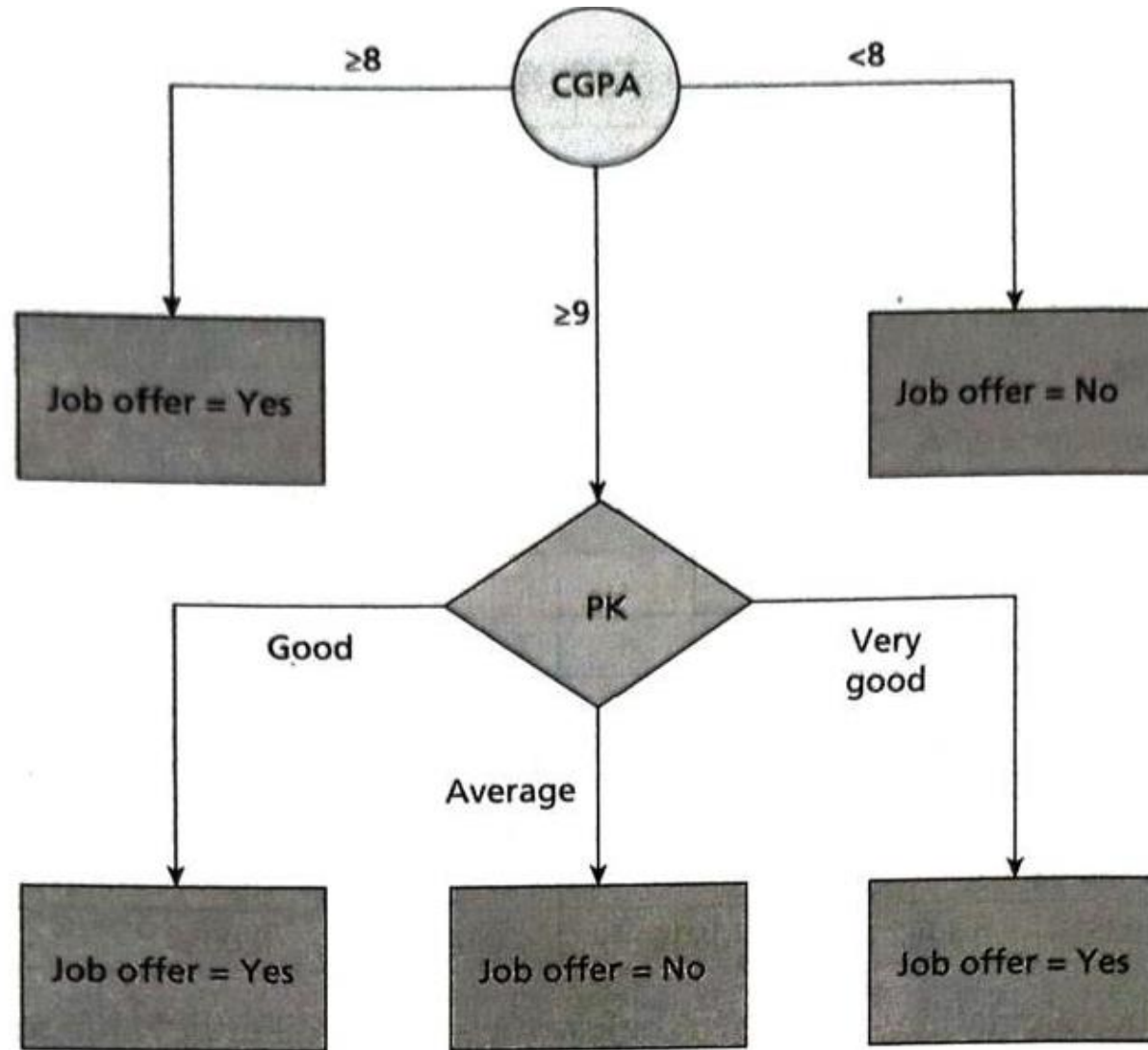


Figure 6.6: Final Decision Tree

## ***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 | 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.

$$\begin{aligned} \text{Entropy\_Info}(T, \text{Job\_Offer}) &= -\left[ \frac{7}{10} \log_2 \frac{7}{10} + \frac{3}{10} \log_2 \frac{3}{10} \right] \\ &= -(-0.3599 + -0.5209) \\ &= 0.8808 \end{aligned}$$

$$\begin{aligned} \text{Entropy}(7, 2) &= -\left[ \frac{7}{9} \log_2 \frac{7}{9} + \frac{2}{9} \log_2 \frac{2}{9} \right] \\ &= -(-0.2818 + -0.4819) \\ &= 0.7637 \end{aligned}$$

$$\begin{aligned} \text{Entropy\_Info}(T, \text{CGPA} \in 6.8) &= \frac{1}{10} \times \text{Entropy}(0, 1) + \frac{9}{10} \text{Entropy}(7, 2) \\ &= \frac{1}{10} \left[ -\frac{0}{1} \log_2 \frac{0}{1} - \frac{1}{1} \log_2 \frac{1}{1} \right] + \frac{9}{10} \left[ -\frac{7}{9} \log_2 \frac{7}{9} - \frac{2}{9} \log_2 \frac{2}{9} \right] \\ &= 0 + \frac{9}{10} (0.7637) \\ &= 0.6873 \end{aligned}$$

$$\begin{aligned} \text{Gain}(\text{CGPA} \in 6.8) &= 0.8808 - 0.6873 \\ &= 0.1935 \end{aligned}$$



**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             | $\leq 7.9$       | No        |
| 5.    | 8.5             | $>7.9$           | Yes       |
| 6.    | 9.5             | $>7.9$           | Yes       |
| 7.    | 7.9             | $\leq 7.9$       | No        |
| 8.    | 9.1             | $>7.9$           | Yes       |
| 9.    | 8.8             | $>7.9$           | Yes       |
| 10.   | 8.8             | $>7.9$           | Yes       |