Criteria to be considered while constructing Decision Tree is

- Entropy
- Information gain

CONSTRUCTION OF DECISION TREE V

Constructing Decision Tree

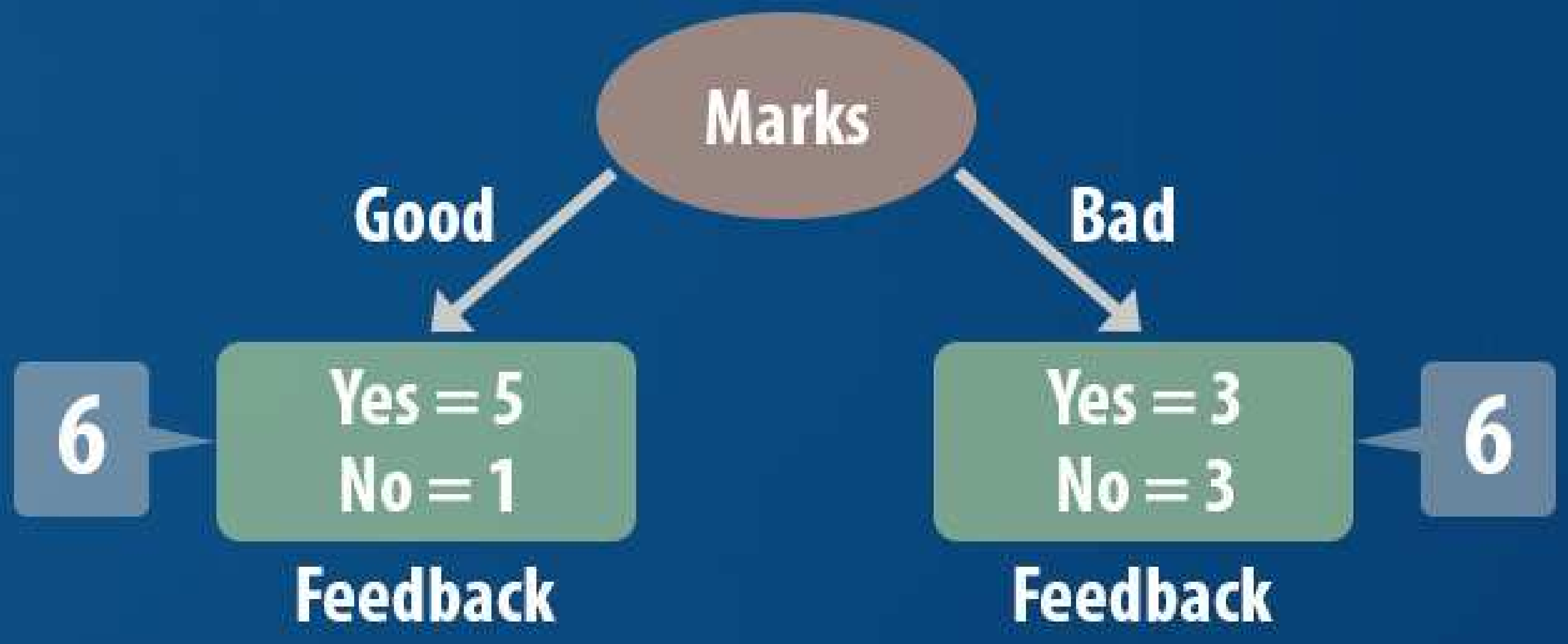
- Constructing a decision tree is all about finding attribute that returns the homogeneous branches. The criteria on which we measure the split or branches in tree is "information gain".
- Information gain is based on entropy. It is calculated for every attribute in data set. The values are sorted, and attributes are placed in the tree by following the order i.e, the attribute with a high value (in case of information gain) is placed at the root.
- Information gain is calculated as follows:

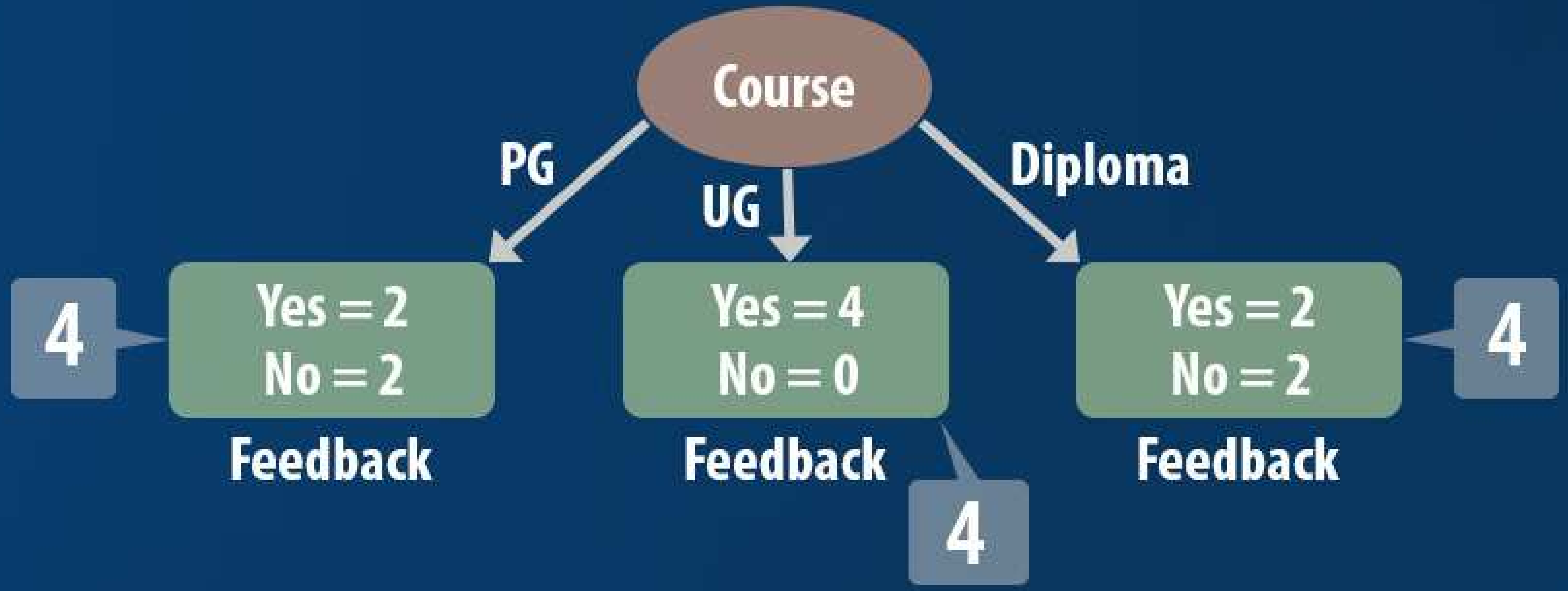
$$\triangle_{\text{Informationgain}} = \text{Entropy (parent) - Average [Entropy (Child)]}$$

- Information gain determine the goodness of an attribute test condition, for this we need to compare the degree of impurity of the parent node with the weighted degree of impurity of the child nodes. The larger their difference, the better the test condition.
- It implies that we need to decrease the Average [Entropy (child)].

CONSTRUCTION OF DECISION TREE VI

Let us first understand Average entropy at child by using simple





$$\text{Entropy (Marks = good)} = -\frac{5}{6}\log_2\frac{5}{6} - \frac{1}{6}\log_2\frac{1}{6}$$

$$= +\frac{5}{6}(+0.263) + \frac{1}{6}(+2.585) = 0.65$$

$$\text{Entropy (Marks = bad)} = -\frac{3}{3}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = -\frac{9}{2}\log_2\frac{1}{2} = 1$$

$$\text{Average [Entropy (children)]} = -\frac{6}{12}\times 0.65 + \frac{6}{12}\times 1$$

$$= -0.325 + 0.5 = 0.175$$

Thus, Marks has a lower weighted entropy than Course.

$$\text{Entropy (Course = PG)} = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

$$\text{Entropy (Course = UG)} = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$$

$$\text{Entropy (Course=Diploma)} = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

$$\text{Average [Entropy (children)]} = -\frac{4}{12}\times 1 + \frac{4}{12}\times 1 + \frac{4}{12}\times 0 = 0.66$$

CONSTRUCTION OF DECISION TREE VII

Constructing Decision Tree - Example illustration

Consider we are given data set below for which decision tree has to be constructed.

Course	Marks	Attendance	Feedback
UG	Bad	Good	Yes
UG	Good	Poor	Yes
PG	Bad	Good	No
PG	Good	Good	Yes
Diploma	Bad	Good	Yes
Diploma	Bad	Poor	No
PG	Bad	Poor	No
PG	Good	Poor	Yes
Diploma	Good	Good	Yes
Diploma	Good	Poor	No
UG	Good	Good	Yes
UG	Bad	Poor	Yes

Step 1

To construct Decision tree, we need to check entropy at the decision variable. Here, in this example it is Feedback. We have two classes for variables Feedback namely, Yes and No. Eight examples correspond to Feedback = Yes and 4 examples to Feedback = No.

$$\text{Entropy (Feedback)} = -\frac{4}{12}\log_2\frac{4}{12} - \frac{8}{12}\log_2\frac{8}{12} = 1.8089$$

CONSTRUCTION OF DECISION TREE VIII

Step 2

We compute $\triangle_{\text{Informationgain}}$ of variables Marks, Course and Attendance with Parent node computed in Step 1. The variable with highest information gain will a root node of Decision tree.

1. $\triangle_{\text{Informationgain}}(\text{Course}) = 1.8089 - \frac{4}{12} \times 1.0 - \frac{4}{12} \times 1.0 = 1.142$

2. $\triangle_{\text{Informationgain}}(\text{Marks}) = 1.8089 - \frac{6}{12} \times 0.65 - \frac{6}{12} \times 1.0 = 0.9839$

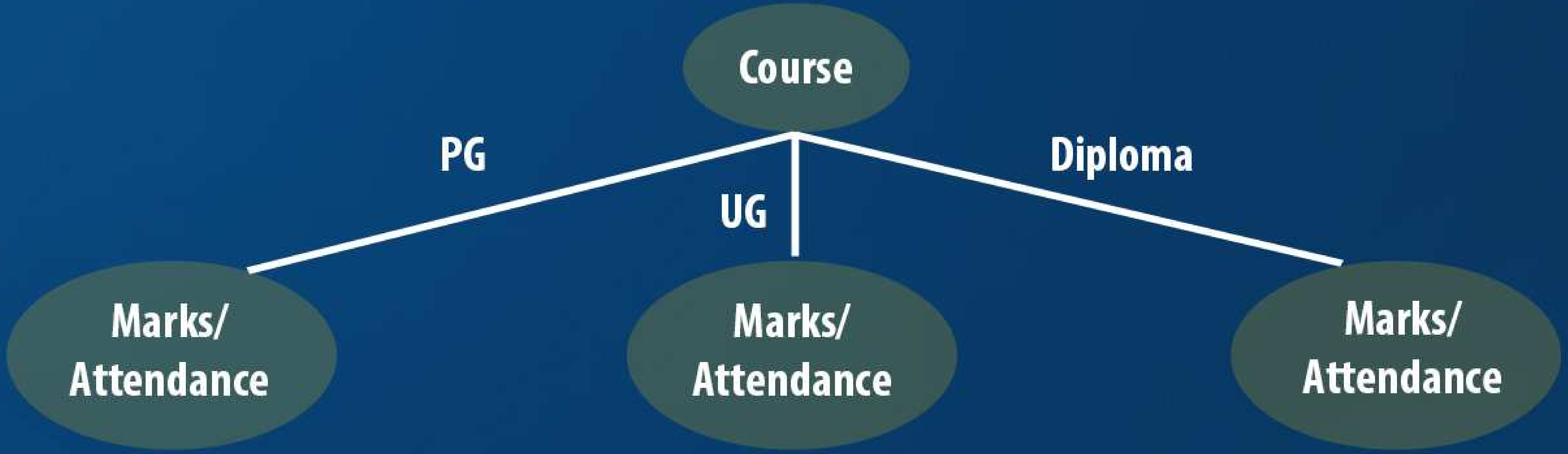
3. $\triangle_{\text{Informationgain}}(\text{Attendence}) = 1.8089 - \frac{6}{12} \times 0.65 - \frac{6}{12} \times 1.0 = 0.9839$

The attribute with the largest gain is Course and therefore, Course is chosen as the first attribute in the decision tree.

CONSTRUCTION OF DECISION TREE IX

Step 3

Till now, we have obtained the root node, now out of remaining two features namely, Marks and Attendance, we need to decide where these features fit best (It is on the left side, the right side or in the middle).



Let us consider all cases:

Case 1: (Course = UG)

Course = UG has 4 examples belonging to class Feedback = Yes and 0 example for Feedback = No, giving an entropy of 0. This branch has reached the leaf node (Clear decision).

CONSTRUCTION OF DECISION TREE X

Case 2: (Course = PG)

It has 2 examples belonging to Feedback = Yes and 2 examples for Feedback = No.
The entropy for Course = PG is therefore

$$E\ (\text{Course=PG}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1.0$$

For this branch we need to decide between features Marks and Attendance. We compute Information gain at Course = PG for both the features. The features with maximum information gain with Course = PG will fit in left branch.

Case 2(a): (Course = PG and Marks)

Marks = Bad has 2 examples for Feedback = No and 0 examples belonging to Feedback = Yes, giving entropy of 0.

Marks = Good has 2 examples for Feedback = Yes and 0 examples belonging to Feedback = No, giving entropy of 0.

$$\triangle_{\text{Informationgain}}\ (\text{Marks}) = 1 - 0 = 1$$

CONSTRUCTION OF DECISION TREE XI

Case 2(b): (Course = PG and Attendance)

Attendance = Good has 1 example for Feedback = Yes and 1 example belonging to Feedback = No

$$E\ (\text{Attendance=Good}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{1}\log_2\frac{1}{2} = 1.0$$

Attendance = Poor has 1 example for Feedback = Yes and 1 examples belonging to Feedback = No, giving entropy of 0.

$$E\ (\text{Attendance=Good}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1.0$$

$$\triangle_{\text{Informationgain}}\ (\text{Attendence}) = 1 - \frac{2}{4}\times 1.0 - \frac{2}{4}\times 1.0 = 0$$

Since Marks has a higher gain, it is chosen as the next attribute after Course = PG.

Case : 3 (Course = Diploma)

The same process as in case 2 is followed for this case.

CONSTRUCTION OF DECISION TREE XII

The final decision tree is as shown in Figure below. It can be seen that if Course = UG, we do not require information about Marks and Attendance to get the right classification. If Course = PG, we do not require information on Attendance. The leaf nodes represents the class labels.

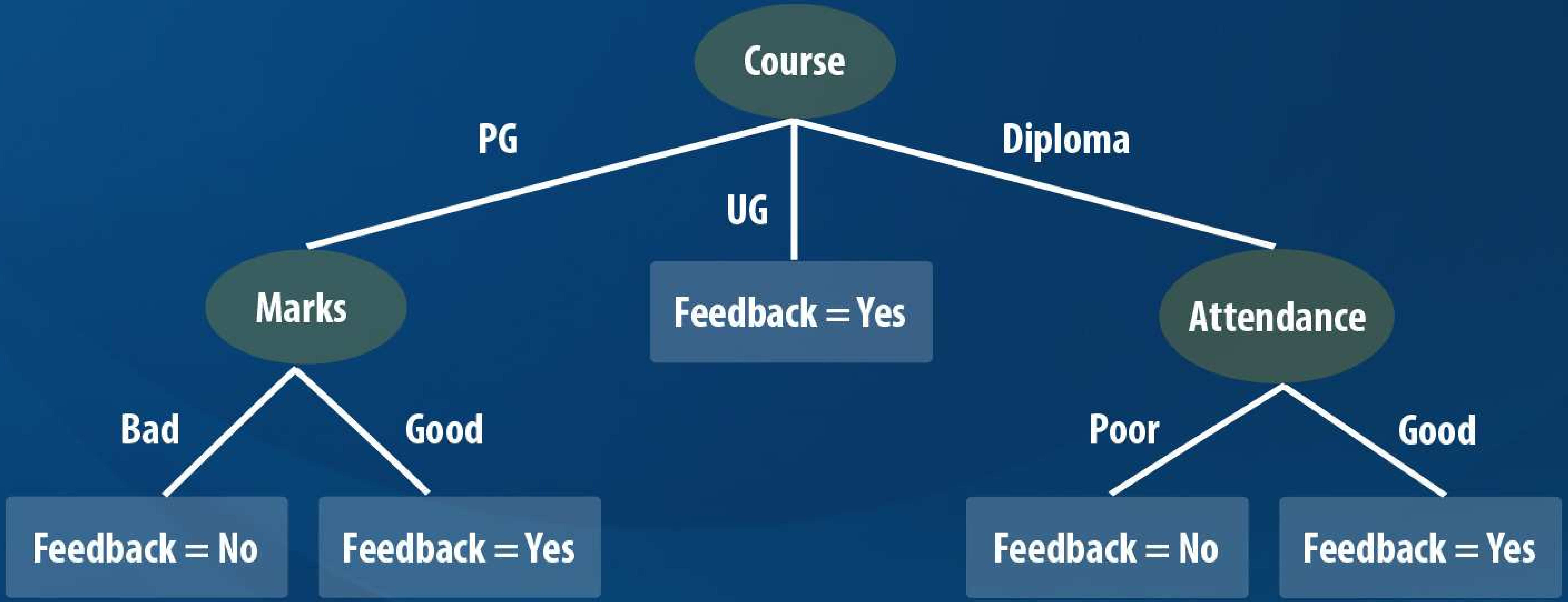
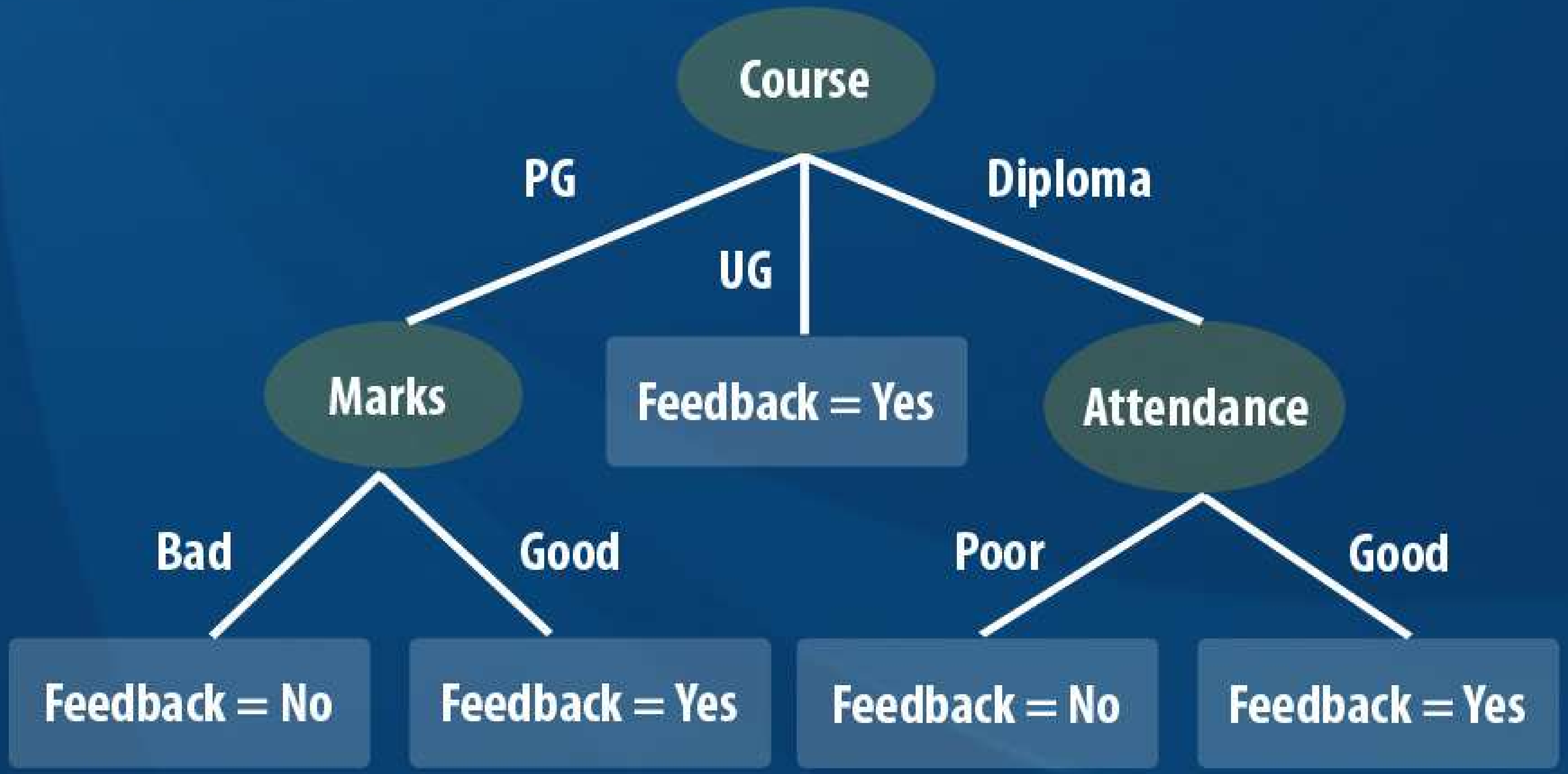
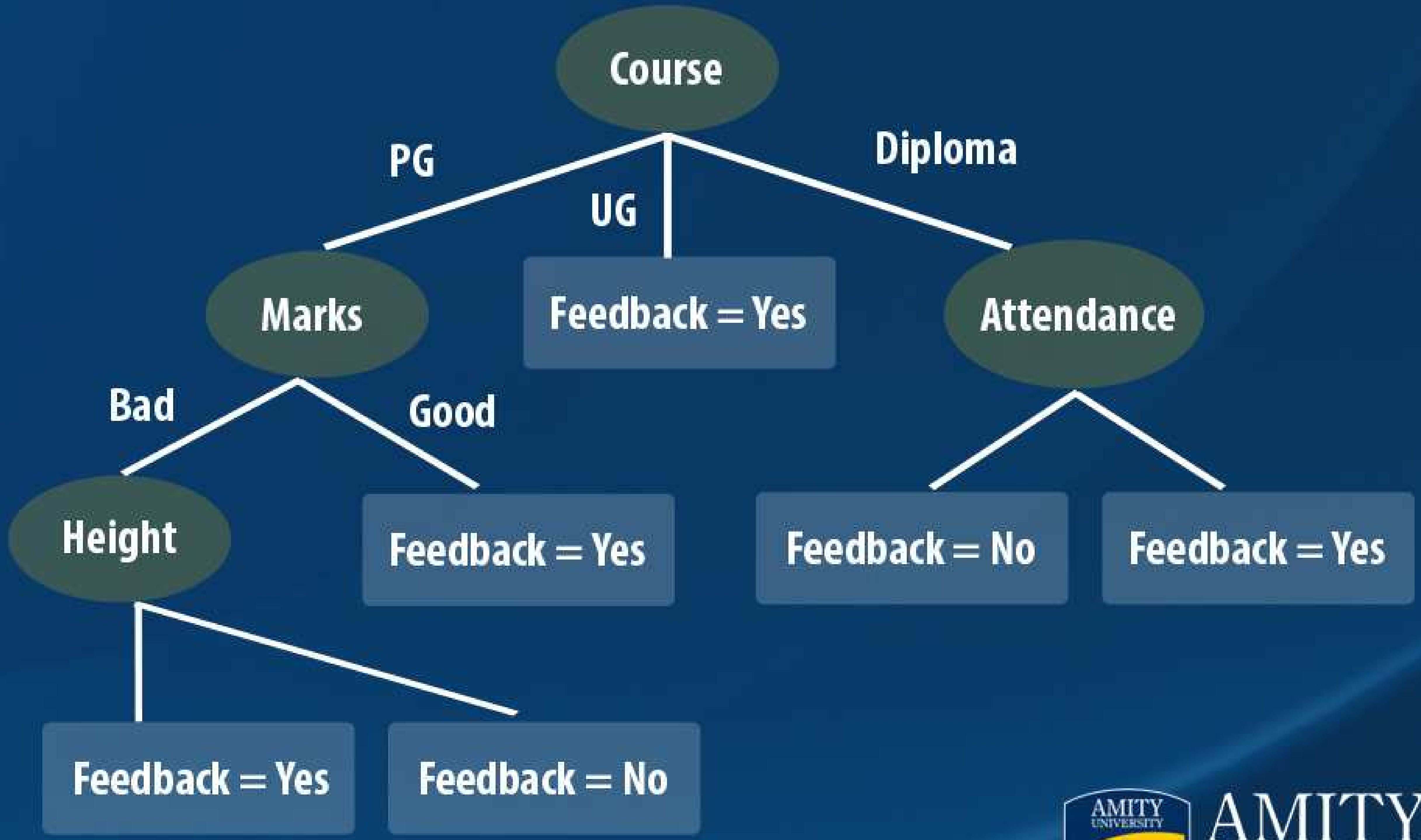


Figure 74:Decision tree induced from training data

PRUNING IN DECISION TREE I

- Pruning is the process of reducing the size of the tree by turning some branch nodes into leaf nodes, and removing the leaf nodes under the original branch.
- Pruning is useful because classification trees may fit the training data well, but may do a poor job of classifying new values.





PRUNING IN DECISION TREE II

Approaches to Tree pruning

- There are several approaches to avoiding overfitting in building decision trees.

  Here are the two approaches:

  Pre-pruning

  Post-pruning

1. Pre-pruning:
   - Ceases the growth of the tree, before it perfectly classifies the training set.
2. Post-pruning:
   - Allows the tree to perfectly classify the training set, and then post prune the tree.

- The idea of pruning in Decision tree is to find minimum depth so that we reach to the solution and remove any irrelevant features from the data set.