DECISION TREE

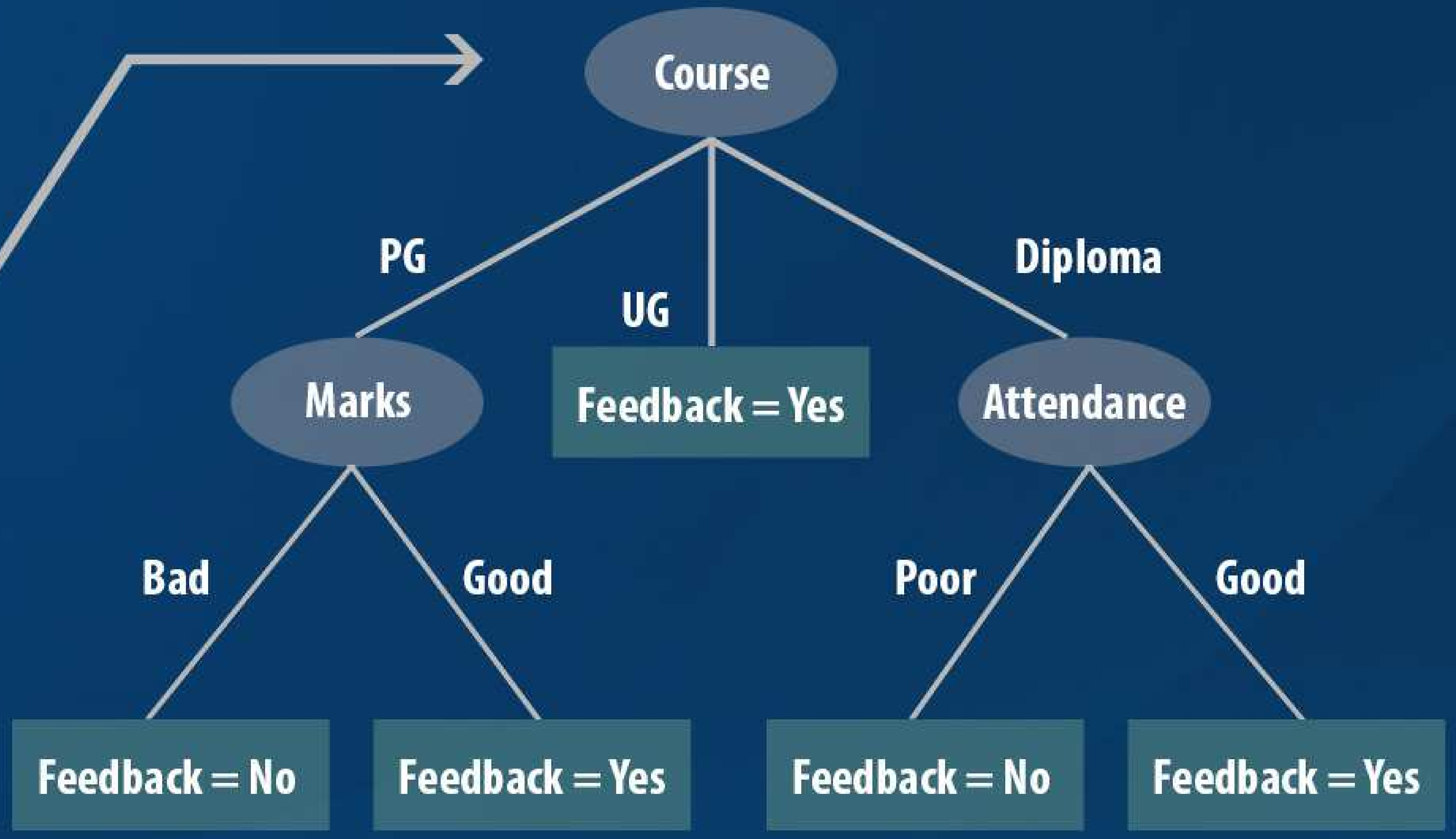
- It is a predictive model.
- Decision Tree works on supervised data where, class variable can be discrete or continuous in nature.
- It is like tree like structure with several branches and an end node at each branch represents the decision/prediction.
- It is non-parametric model. It means that it does not make any assumptions on the underlying data distribution.

DECISION TREE - INTRODUCTION

Decision tree is a supervised algorithm used for predictions. Given a training data set, the model results in a tree like structure composed of features and relationships used to forecast an unknown event.

Indicator variable			Class variable
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

Training data set

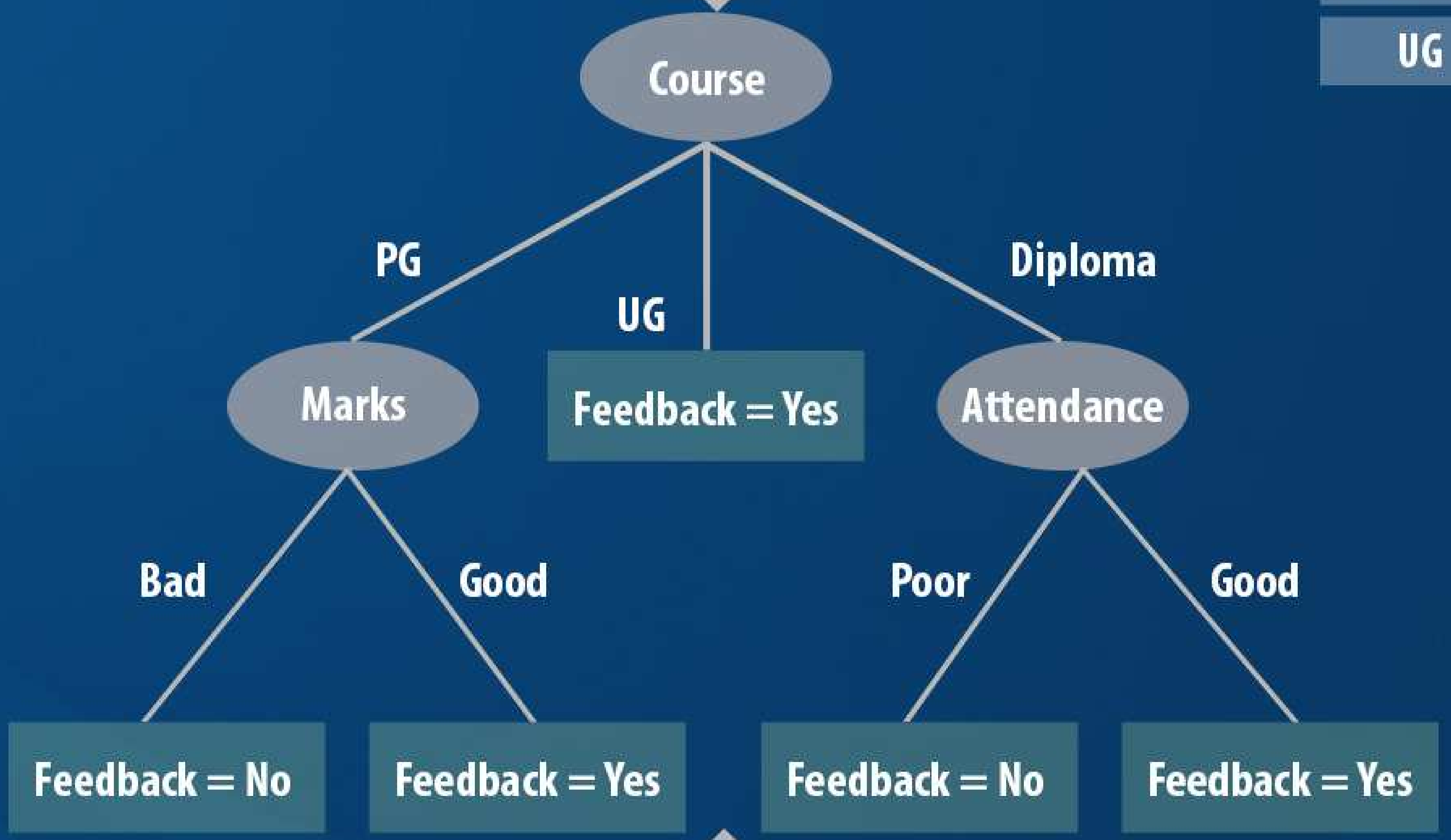


Decision Tree Model

PREDICTION USING DECISION TREE- INTRODUCTION

Course	Marks	Attendance	Feedback
Diploma	Good	Poor	?
UG	Good	Good	?

Test set

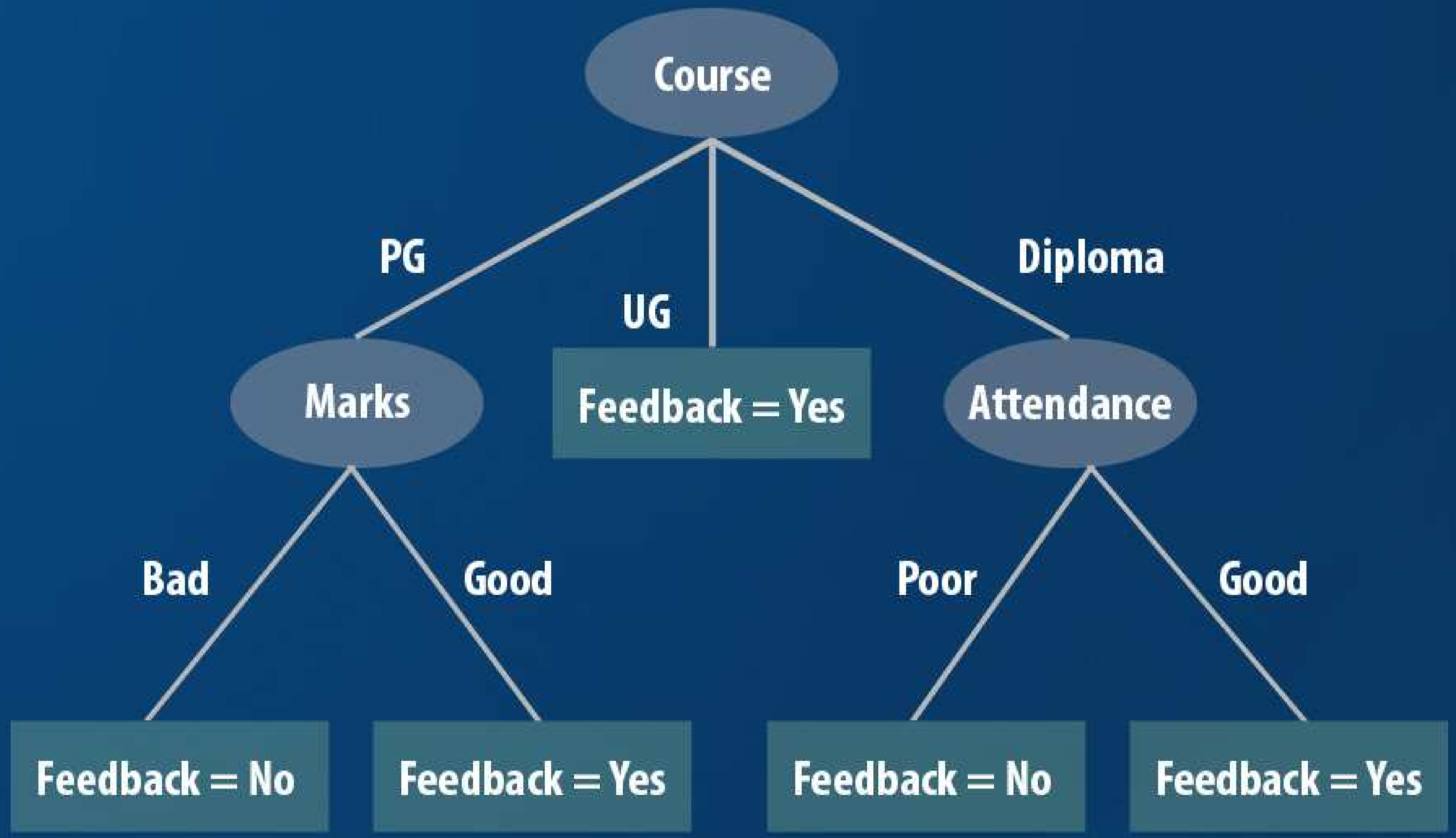


Course	Marks	Attendance	Feedback
Diploma	Good	Poor	No
UG	Good	Good	Yes

Prediction on Test set

DECISION TREE BASICS I

In Decision tree, indicator variable is addressed as internal node and predictor variable is addressed as leaf node.



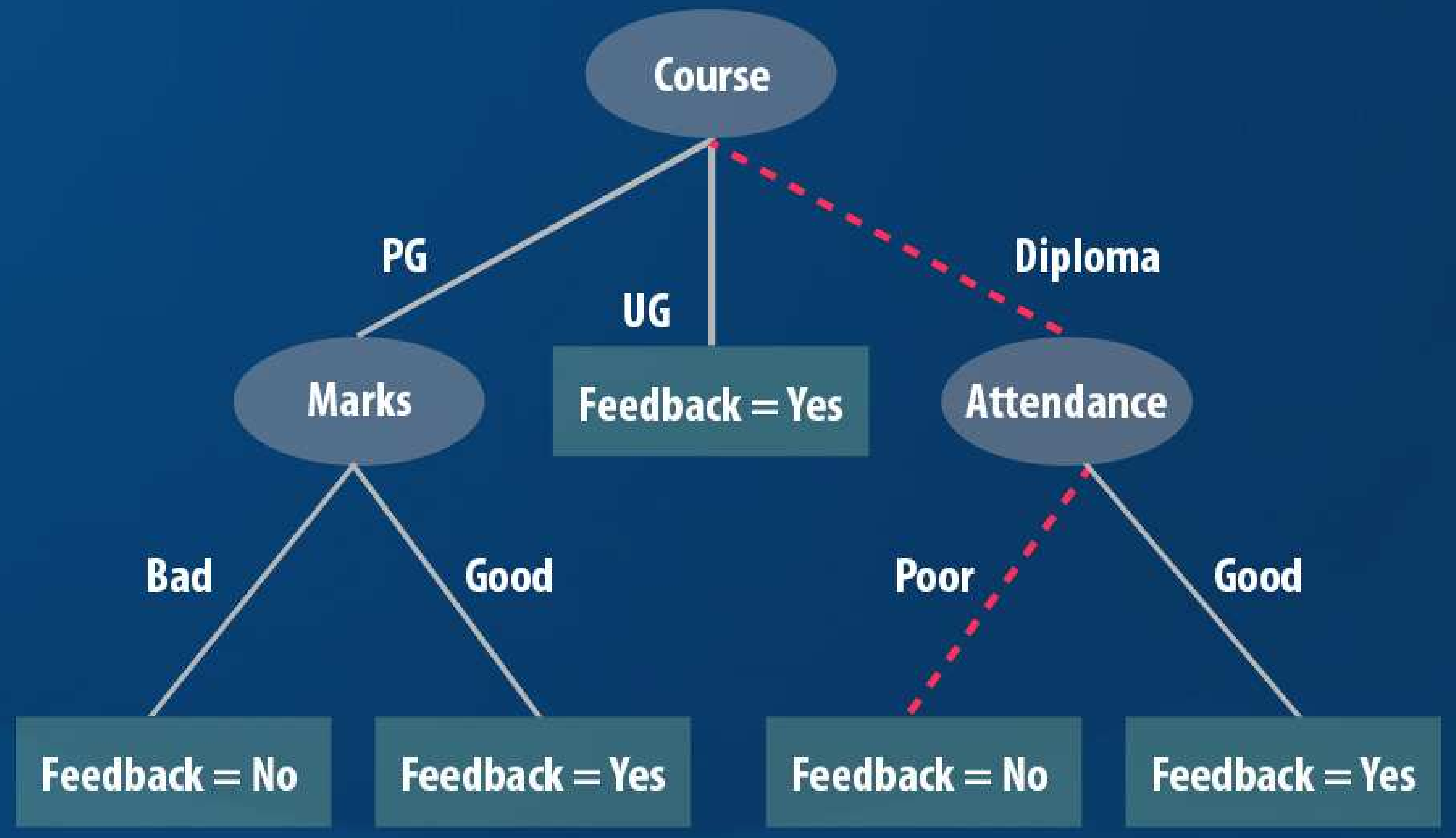
Here,

1. Internal nodes: Course, Marks and Attendance
2. Leaf node: Feedback

DECISION TREE BASICS III

Decision Tree Basics

- Classification results are represented in every branch of the decision tree.
- If we start from root node and reach to the leaf node, we get classification result.

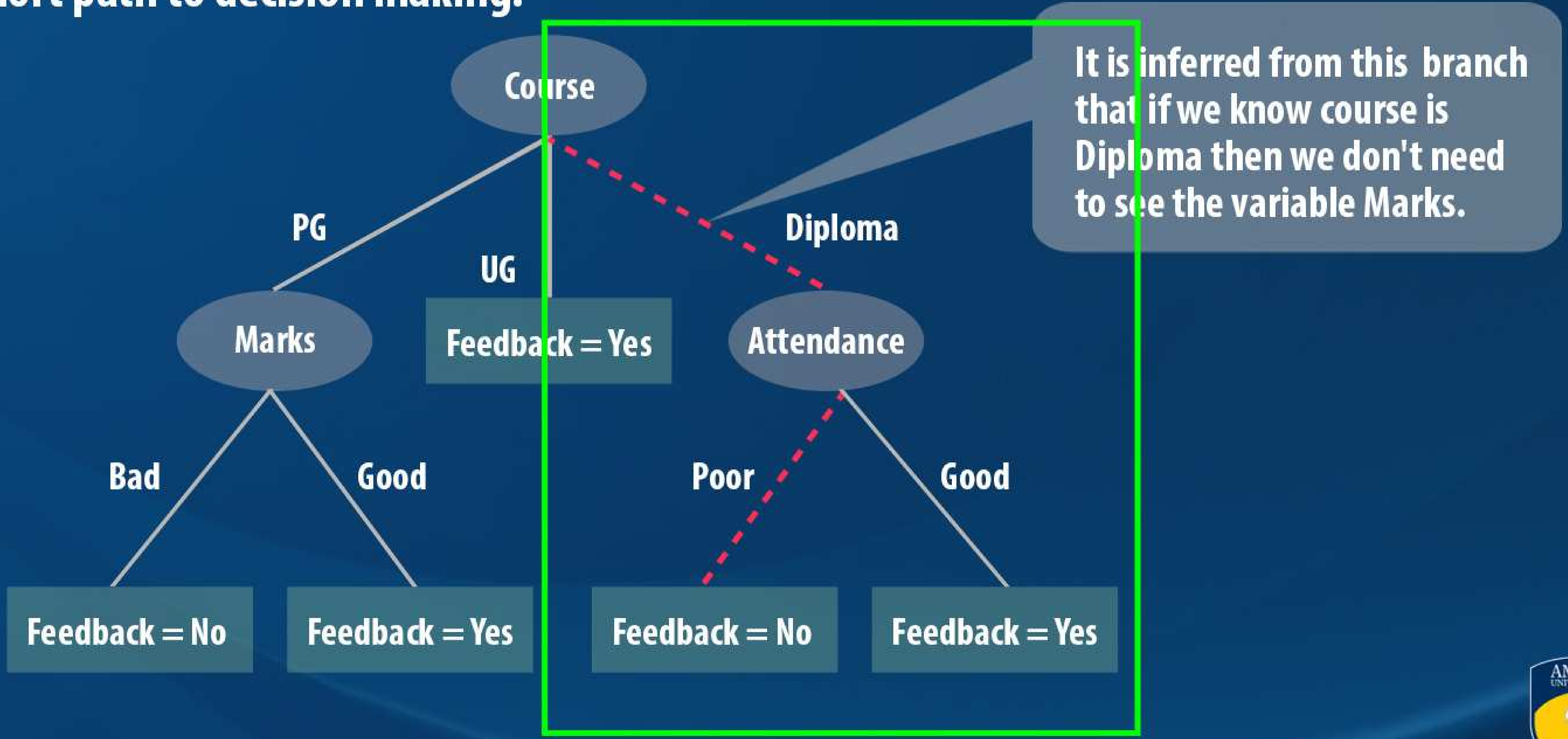


- Consider the extreme left branch of the tree, starting from Course = PG, Marks = Bad, we reach to the classification result of Feedback = No.

DECISION TREE BASICS IV

Decision Tree Basics

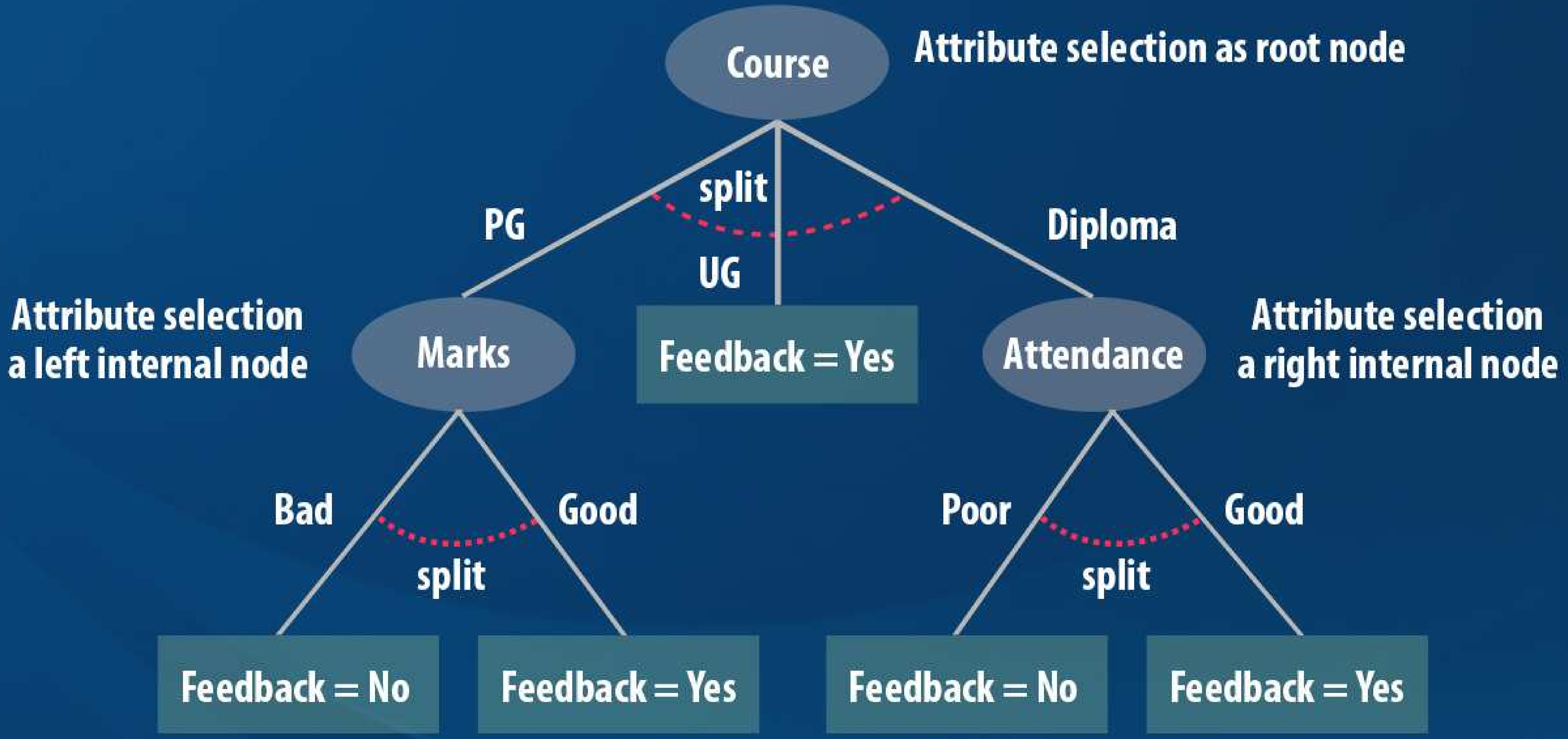
Decision tree are smart in eliminating the irrelevant features from the model and uses simple, direct, short path to decision making.



CONSTRUCTION OF DECISION TREE I

There are two major concerns in development of Decision tree model.

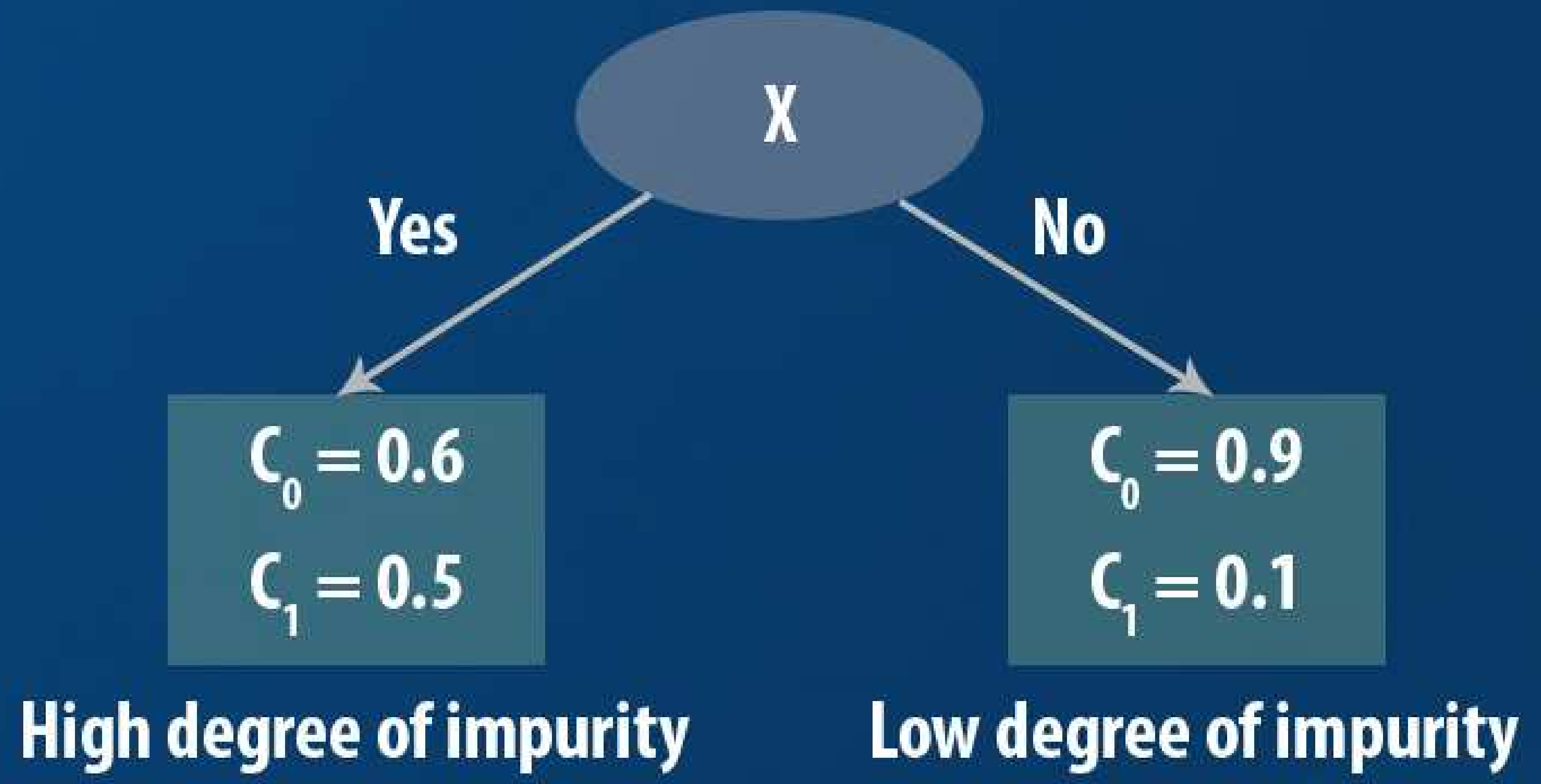
1. Attribute selection
2. Splitting at the attribute (node)



Entropy is measure of impurity.

CONSTRUCTION OF DECISION TREE II

- To construct a Decision tree, one or more attribute have to be chosen at each node to make a decision.
- The most important attribute is the one that leads to a clear decision. A clear decision in context with classification is the one where there is purity of classes.
- The primary challenge in the decision tree implementation is to identify which attributes do we need to consider as the root node and each level. Handling this is known as the attributes selection.



- In the figure above, choosing the right branch is more pure than the left branch. It is for the reason that right branch is giving pure distribution of classes than in left branch.
- A pure distribution of classes make the decision clear and hence attributes that gives such paths in the decision tree are preferred.

CONSTRUCTION OF DECISION TREE III

In order to compute the purity of a node, we may choose any of the following metric:

1. Entropy

$$E(x) = -\sum_{i=0}^{c-1} P_i(x) Log_2 P_i(x) \quad (40)$$

2. Gini Index

$$1 - \sum_{i=0}^{c-1} P_i(x)^2 \quad (41)$$

Where,

c represents number of classes.

$P_i(x)$ indicates the probability of instances that belong to class i at node x.

Example

X_1	Frequency
C_0	0
C_1	6

Table 16

X_2	Frequency
C_0	1
C_1	5

Table 17

Gini(x)= $1 - (0/6)^2 - (6/6)^2 = 0$

E(x)= $-(0/6)\log_2(0/6) - (6/6)\log_2(6/6) = 0$

Gini(x)= $1 - (1/6)^2 - (5/6)^2 = 0.278$

E(x)= $-(1/6)\log_2(1/6) - (5/6)\log_2(5/6) = 0.65$

Here, variable x_1 is preferred over x_2.