RANDOM FOREST TECHNIQUE

The Random Forest technique is illustrated in the Figure 4 below:

Figure 4: Random Forest Technique

RANDOM FOREST CONSTRUCTION

- The construction of Random forest consist of several steps as detailed below:

1. Create Bootstrapped data set.
2. Build Decision tree using the bootstrapped data set, but only use random subset of variables at each step.
3. Repeat Step 1 and 2, i.e., make a new bootstrapped data set and build a new tree considering subset of variables at each step.
4. The steps 1 and 2 are repeated n number of times to build random forest of n decision trees.

RANDOM FOREST CONSTRUCTION - EXAMPLE ILLUSTRATION

Original data set

Course	Marks	Height	Attendance	Feedback
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	No
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	No
--------	--------	--------	--------	No
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	No
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes

Figure 5: Sample Data set

RANDOM FOREST CONSTRUCTION - EXAMPLE ILLUSTRATION

Step 1: Create Bootstrapped data set from original data set

Original data set

Course	Marks	Height	Attendance	Feedback
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	No
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	No
--------	--------	--------	--------	No
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	No
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes

Bootstrapped data set

Course	Marks	Height	Attendance	Feedback	
--------	--------	--------	--------	Yes	
--------	--------	--------	--------	Yes	2
--------	--------	--------	--------	Yes	
--------	--------	--------	--------	Yes	4
--------	--------	--------	--------	Yes	
--------	--------	--------	--------	No	
--------	--------	--------	--------	No	
--------	--------	--------	--------	Yes	8
--------	--------	--------	--------	Yes	
--------	--------	--------	--------	No	
--------	--------	--------	--------	No	11
--------	--------	--------	--------	Yes	

Figure 6: Creating Bootstrapped data set

RANDOM FOREST CONSTRUCTION - EXAMPLE ILLUSTRATION

Step 2 (a): Build Decision tree using the bootstrapped data set, but only use random subset of variables at each step

Randomly select 2 variables out of 4 to decide for root node. Let's assume, Mark and Height.

For the sake of example, assume Height as identified as better choice than Marks, based on the information gain criteria of decision tree.

Bootstrapped data set

Course	Marks	Height	Attendance	Feedback
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	No
--------	--------	--------	--------	No
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	No
--------	--------	--------	--------	No
--------	--------	--------	--------	Yes

Figure 7: Deciding root node of Decision tree from random variables selection

RANDOM FOREST CONSTRUCTION - EXAMPLE ILLUSTRATION

Step 2 (b): Build Decision tree using the bootstrapped data set, but only use random subset of variables at each step

Randomly select 2 variables excluding root.

Bootstrapped data set

Course	Marks	Height	Attendance	Feedback
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	No
--------	--------	--------	--------	No
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	No
--------	--------	--------	--------	No
--------	--------	--------	--------	Yes

For the sake of example, below decision tree is formed based on information gain criteria.

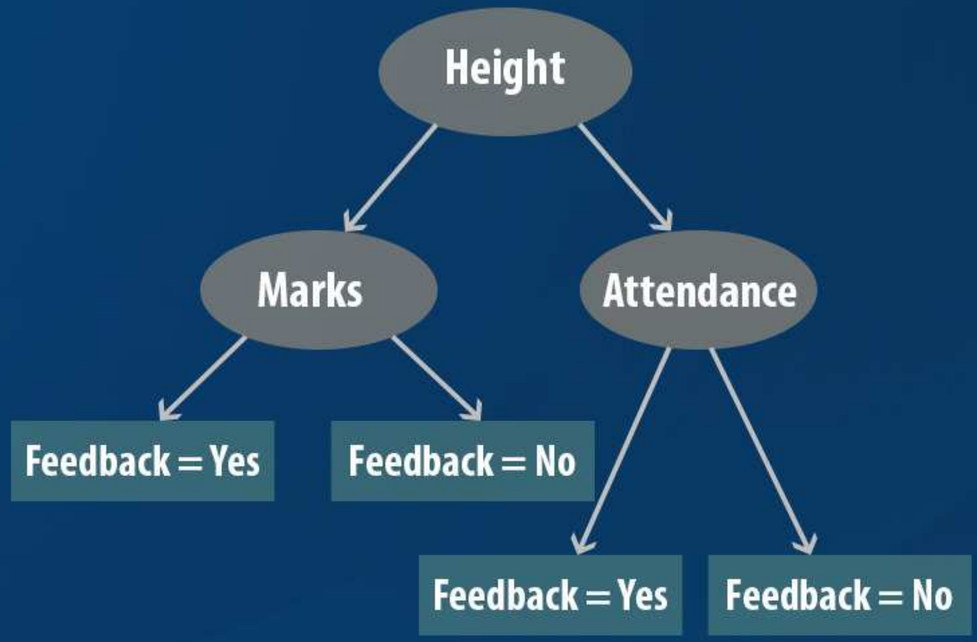

Figure 8: Decision tree from random variable selection

RANDOM FOREST CONSTRUCTION - EXAMPLE ILLUSTRATION

Step 2 (a): Build a new Decision tree using the bootstrapped data set, but only use random subset of variables at each step

Randomly select 2 variables out of 4 to decide for root node. Let's assume, Height and Attendance.

For the sake of example, assume Attendance as identified as better choice than Height based on the information gain criteria of decision tree.

Bootstrapped data set

Course	Marks	Height	Attendance	Feedback
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	No
--------	--------	--------	--------	No
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	No
--------	--------	--------	--------	No
--------	--------	--------	--------	Yes

Figure 9: Deciding root node of Decision tree from random variables selection

RANDOM FOREST CONSTRUCTION - EXAMPLE ILLUSTRATION

Step 2 (b): Build Decision tree using the bootstrapped data set, but only use random subset of variables at each step

Randomly select 2 variables excluding root.

Bootstrapped data set

Course	Marks	Height	Attendance	Feedback
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	No
--------	--------	--------	--------	No
--------	--------	--------	--------	Yes
--------	--------	--------	--------	Yes
--------	--------	--------	--------	No
--------	--------	--------	--------	No
--------	--------	--------	--------	Yes

For the sake of example, below decision tree is formed based in information gain criteria.

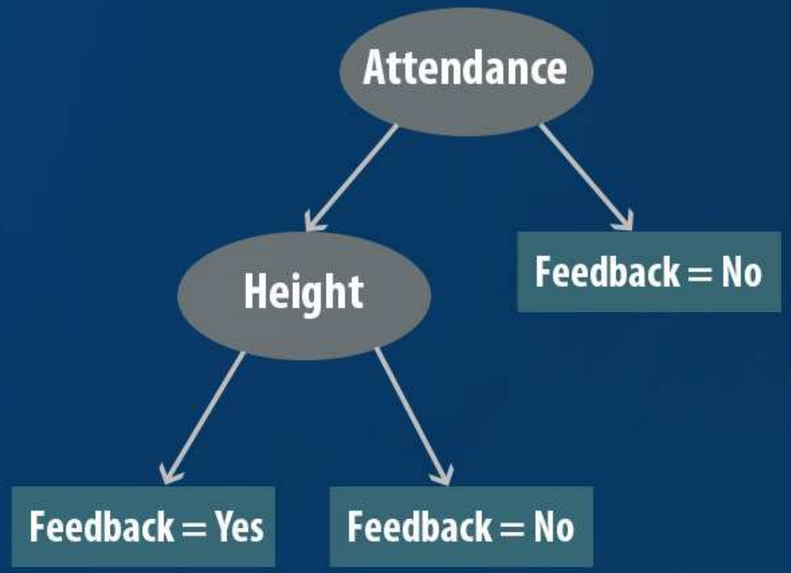

Figure 10: Decision tree from random variable selection

The process of constructing decision trees is continued till new decision trees are discovered.

RANDOM FOREST CONSTRUCTION - EXAMPLE ILLUSTRATION

Step 3: Collect random forest of all Decision Trees. For the sake of example, assume 4 decision trees are discovered

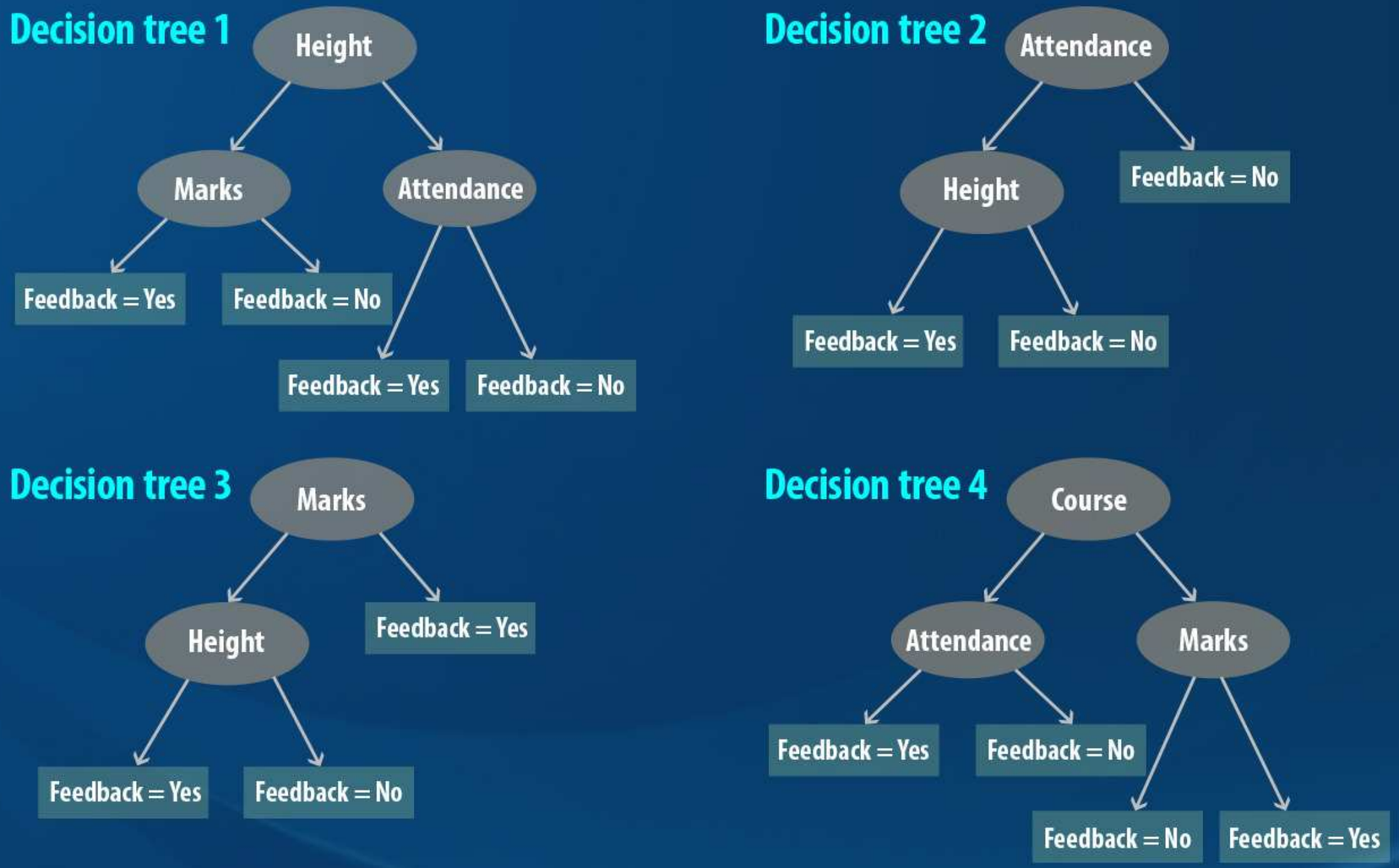

Figure 11: Random forest with 4 Decision Trees

RANDOM FOREST CONSTRUCTION - EXAMPLE ILLUSTRATION

Step 4: Apply Bagging

For a given new/unknown observation, get the predictive result from each decision tree in the forest. The final predictive result will be based on the maximum votes collected from all decision trees in the forest for a particular class/label.

Consider for example,

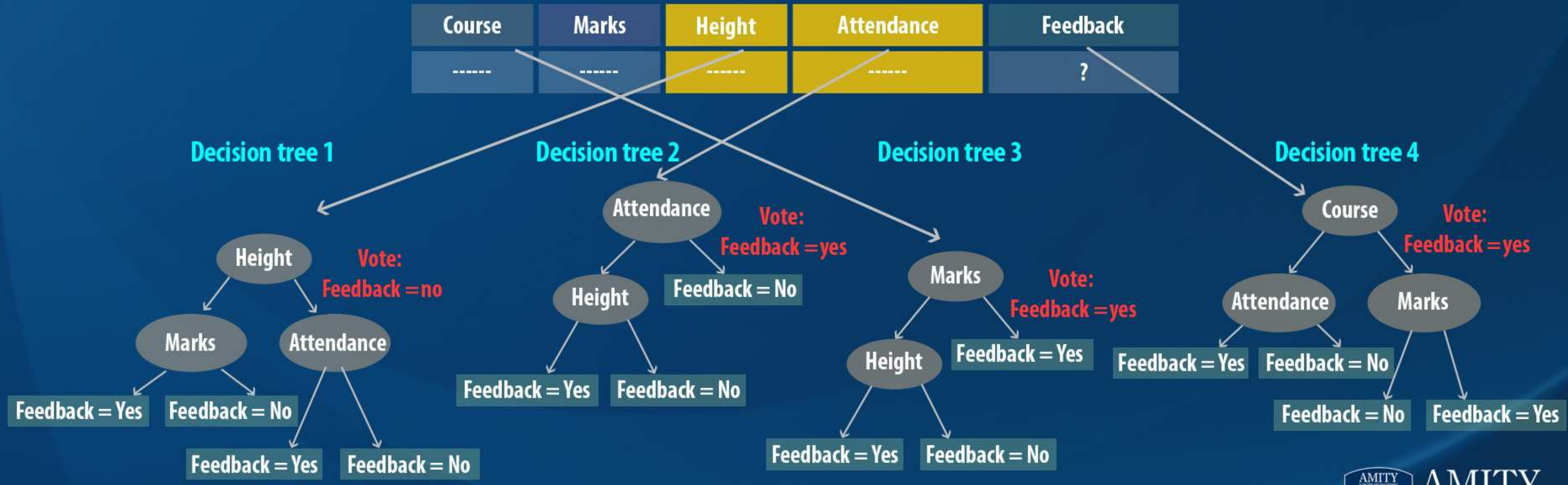

Figure 12: The final decision of class/label based on votes is Feedback = Yes

PARAMETERS IN RANDOM FOREST CONSTRUCTION

1. How to select number of random variables in construction of decision tree?

   Use $\sqrt{n}$, where, n is number of features present in the data set.

2. How many trees in the forest?

   Having more trees actually strengthens the final estimate. The default setting is 10 in most of the built in libraries but, it can be set based on user's choice.

STRENGTHS AND WEAKNESS OF RANDOM FOREST

- **Strengths**
  - Random forest can be used as classification or regression model
  - Reduction in overfitting
  - More stable (low variance model)

- **Weakness**
  - It is a black box model. Difficult to visualise the model or understand why it predicted something.
  - Computationally expensive.