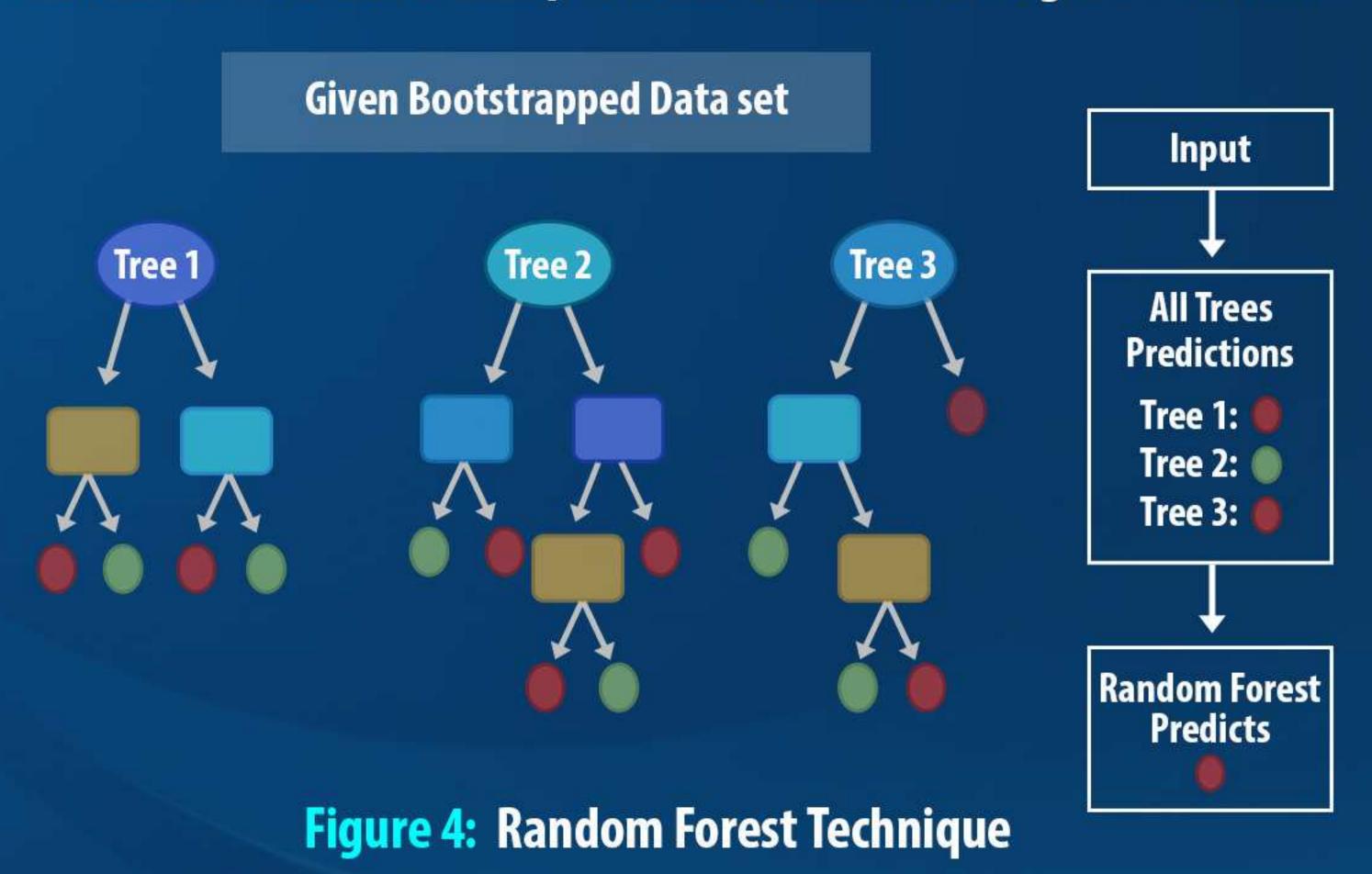
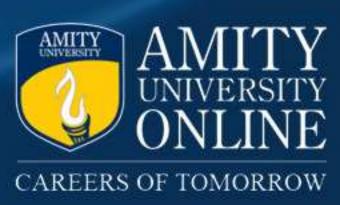
RANDOM FOREST TECHNIQUE

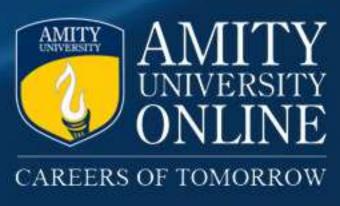
The Random Forest technique is illustrated in the Figure 4 below:





RANDOM FOREST CONSTRUCTION

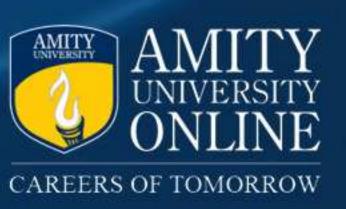
- The construction of Random forest consist of several steps as detailed below:
 - 1. Create Bootstrapped data set.
 - 2. Build Decision tree using the bootstrapped data set, but only use random subset of variables at each step.
 - 3. Repeat Step 1 and 2, i.e., make a new bootstrapped data set and build 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.



Original data set

Course	Marks	Height	Attendance	Feedback
The All State of the		Malada Jaka		
- Samulainininini				Yes
ile indolololololole	Edutation to the little of the			Yes
		F		No
				Yes
				Yes
» 				No
			()	No
			()	Yes
(Yes
				No
				Yes
X-statement and states	i adalah dalah dalah		Salaran and Assault	Yes

Figure 5: Sample Data set



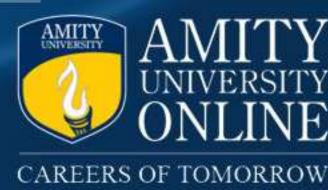
Step 1: Create Bootstrapped data set from original data set

Original data set

Bootstrapped data set

Course	Marks	Height	Attendance	Feedback	Course	Marks	Height	Attendance	Feedback	
	<u> </u>	<u> </u>		Yes	***********	·		8-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1	Yes	
				Yes					Yes	
				No				***************************************	Yes	ı
				Yes	: ::::::::::::::::::::::::::::::::::::	(properties)			Yes	4
				Yes	1		200000	71 -11-11-11-1	Yes	ı
				No	Tallahahahakakakak	Wednesdated and the later		1 Martin de	No	ı
		-	#41 New York Towns	No	(Market Market)	Western Williams	ENGINEERICA PE	(Entrancemental)	No	
***************************************		24 (4 (4 (4 (4 (4 (4 (4 (4 (4 (4 (4 (4 (4		Yes	STABBLE DOOR				Yes	1
				Yes					Yes	
				No	((a	No	
				Yes			11:11:11:		No	1
				Yes					Yes	

Figure 6: Creating Bootstrapped data set



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.

Bootstrapped data set

Course	Marks	Height	Attendance	Feedback
		******		Yes
				Yes
				Yes
Sala Cale				Yes
				Yes
				No
	*	T-AMARICANA T		No
				Yes
1334411111111				Yes
				No
		*******		No
	Service of the servic	National Control		Yes

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



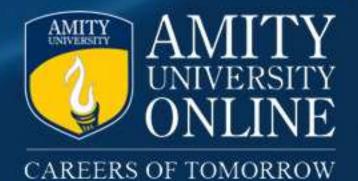


Figure 7: Deciding root node of Decision tree from random variables selection

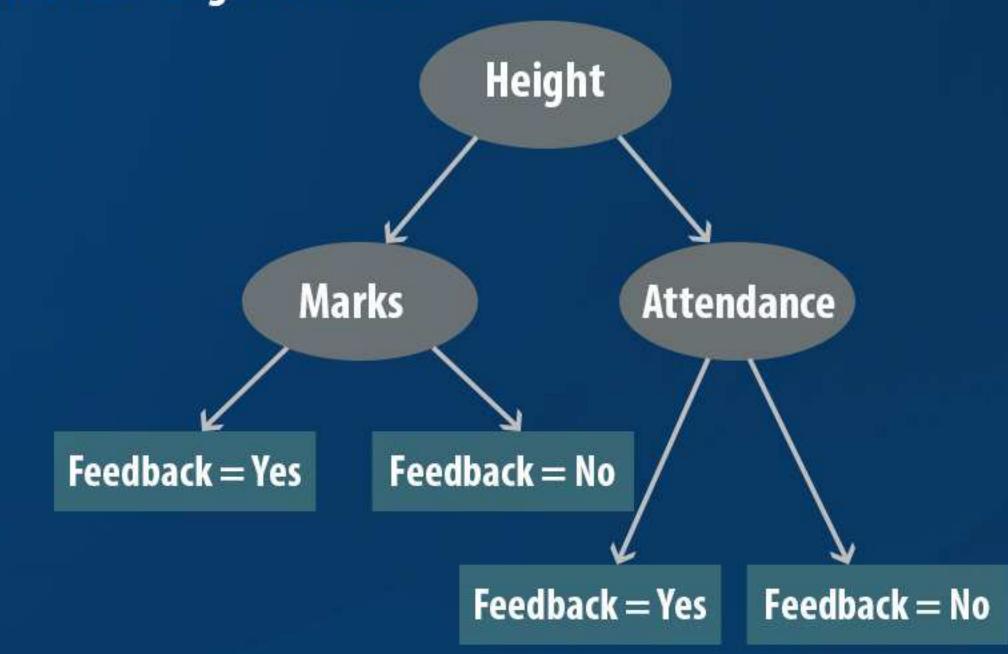
Step 2 (b): Build Decision tree using the bootstrapped data set, but only use random subset of variables at each step

Randomly select 2 variables excluding root.

Bootstrapped data set

Course	Marks	Height	Attendance	Feedback
		Substitution of the substi		Yes
				Yes
	Market Market Company	***************************************	72222000	Yes
	Name and Address of the Owner, where the Owner, which is the Owner, where the Owner, which is the Owner, where the Owner, which is the Owner, whic	Talana and a		Yes
			(accessors)	Yes
	******			No
	2222222		72000000	No
(Artistantino) Statement (State				Yes
	-		-	Yes
		(======		No
		/2000000		No
		PROPERTY.	<u> </u>	Yes

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



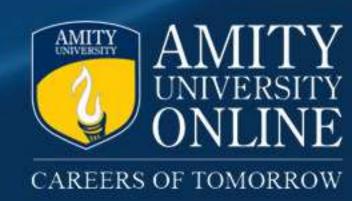


Figure 8: Decision tree from random variable selection

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
		(amazanta)		Yes
				Yes
				Yes
	AND ADDRESS OF THE PARTY OF THE			Yes
				Yes
				No
				No
				Yes
				Yes
				No
	Maria de la companio del companio de la companio de la companio del companio de la companio del companio de la companio de la companio de la companio del companio de la companio del companio de la companio de la companio de la companio del companio de la compan			No
Testino de la companio della compani	Total Control Control	Carata de la caración	- Inches	Yes





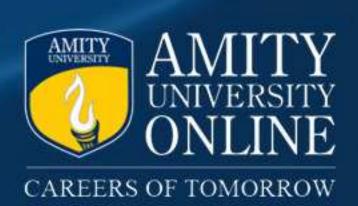


Figure 9: Deciding root node of Decision tree from random variables selection

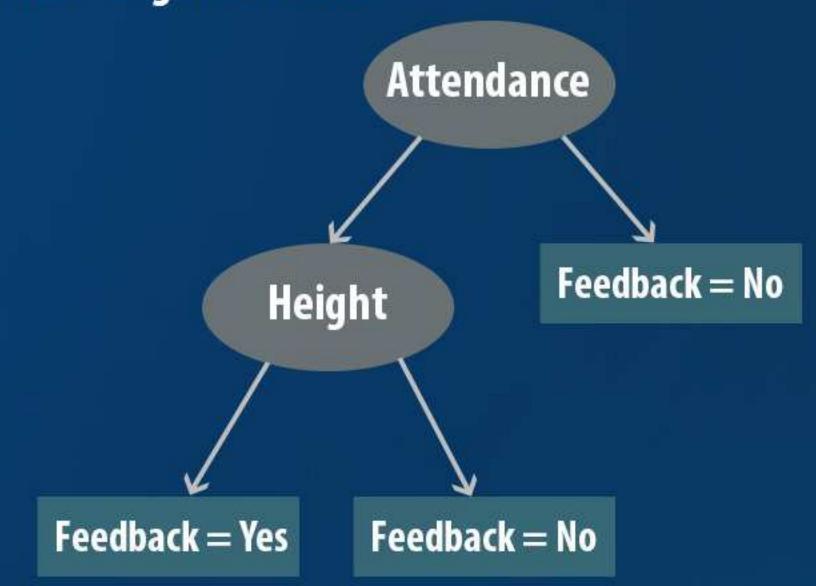
Step 2 (b): Build Decision tree using the bootstrapped data set, but only use random subset of variables at each step

Randomly select 2 variables excluding root.

Bootstrapped data set

Course	Marks	Height	Attendance	Feedback
				Yes
				Yes
(Computer)	(210)00-1202	(SHILLSHAME)	NOTE AND PROPERTY.	Yes
	Canada de la compansión		2-1-11-11-1	Yes
				Yes
(minutes)	:=Winte-Mile	(SHADWANA)		No
				No
				Yes
Santana -		(companies)		Yes
				No
Anna market	ie MWWwedth	:=W#Wexter	(PHANCE NAME)	No
				Yes

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



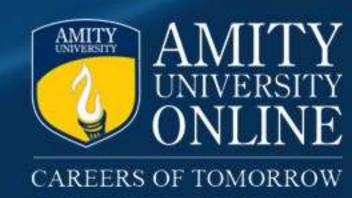
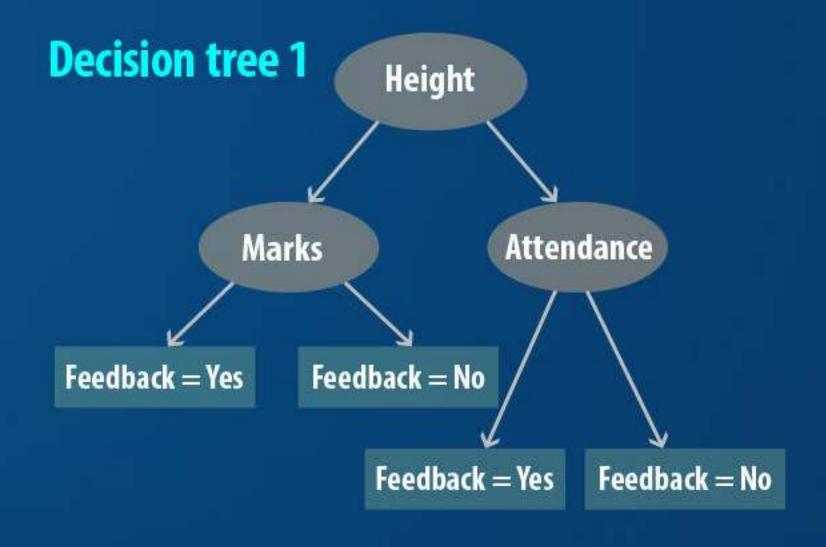
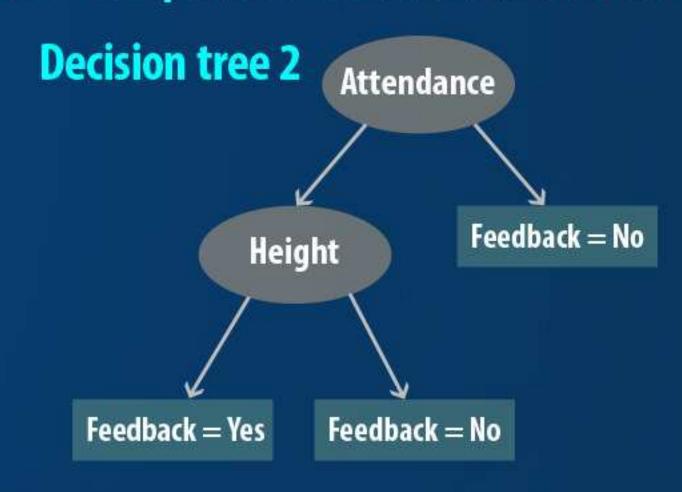


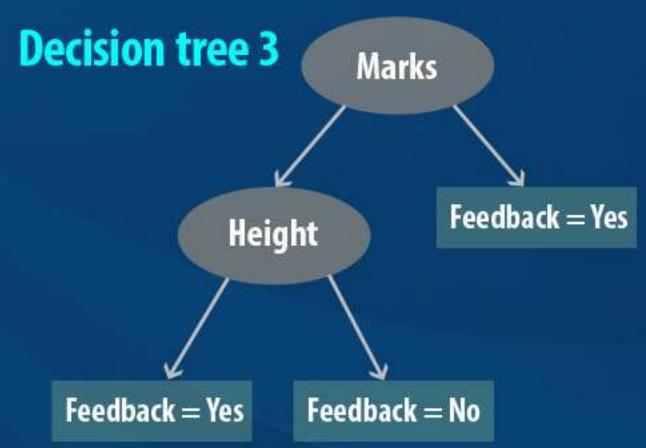
Figure 10: Decision tree from random variable selection



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







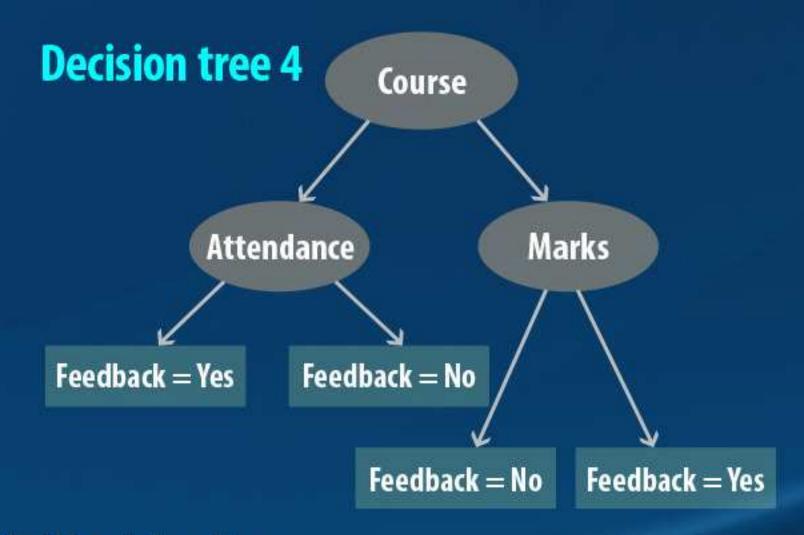
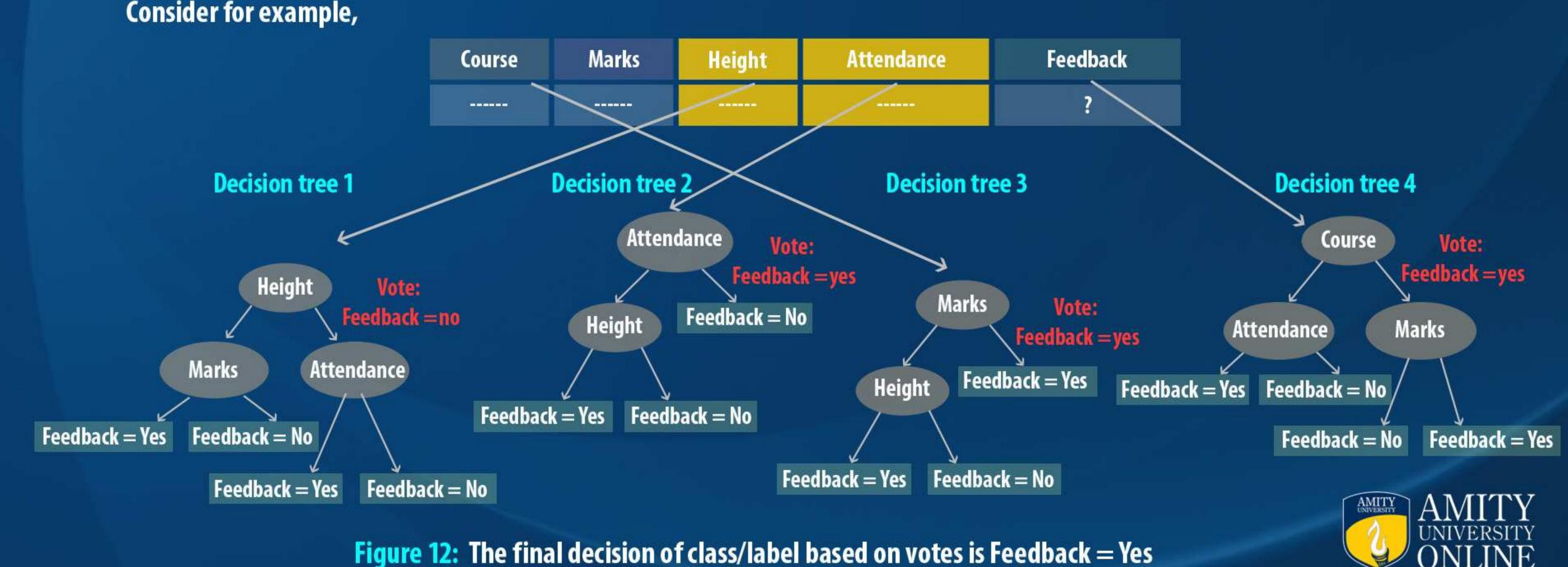




Figure 11: Random forest with 4 Decision Trees

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.



CAREERS OF TOMORROW

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.

