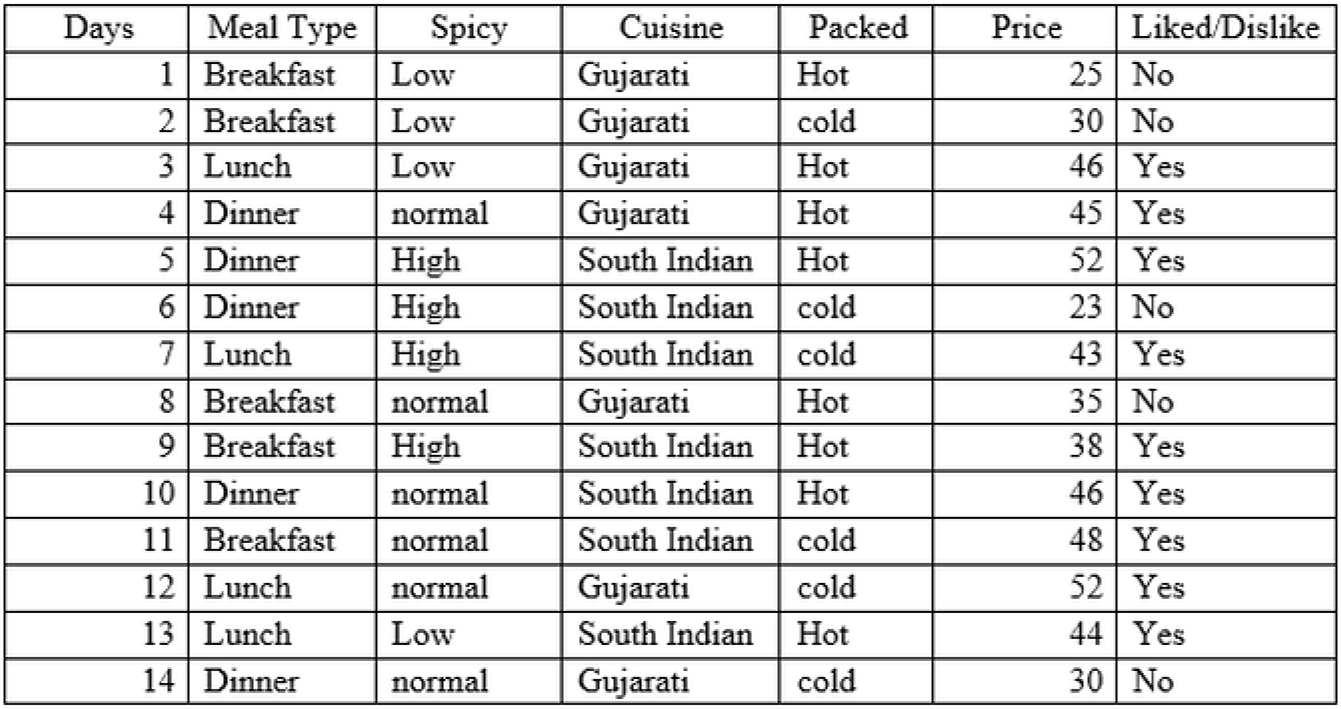
Understanding Decision Tree With A Numerical Example

Let’s start with a dataset that is hypothetical, where the target variable is whether the customer liked the food or not.



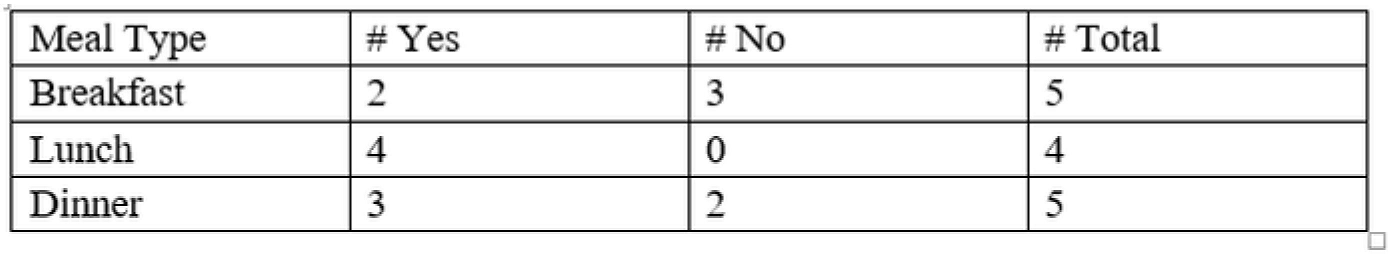
Days	Meal Type	Spicy	Cuisine	Packed	Price	Liked/Dislike
1	Breakfast	Low	Gujarati	Hot	25	No
2	Breakfast	Low	Gujarati	cold	30	No
3	Lunch	Low	Gujarati	Hot	46	Yes
4	Dinner	normal	Gujarati	Hot	45	Yes
5	Dinner	High	South Indian	Hot	52	Yes
6	Dinner	High	South Indian	cold	23	No
7	Lunch	High	South Indian	cold	43	Yes
8	Breakfast	normal	Gujarati	Hot	35	No
9	Breakfast	High	South Indian	Hot	38	Yes
10	Dinner	normal	South Indian	Hot	46	Yes
11	Breakfast	normal	South Indian	cold	48	Yes
12	Lunch	normal	Gujarati	cold	52	Yes
13	Lunch	Low	South Indian	Hot	44	Yes
14	Dinner	normal	Gujarati	cold	30	No

From the above data, Meal type, Spicy, Cuisine and packed are the inputs/features of data and liked/dislike is the target variable.

Now let’s start building tree having Gini index as im(purity) measure.

Meal Type

Meal Type is a nominal data that has 3 values Breakfast, Lunch and Dinner. Let’s classify the instances on basis of liked/dislike.



Meal Type	# Yes	# No	# Total
Breakfast	2	3	5
Lunch	4	0	4
Dinner	3	2	5

Gini index (Meal Type = Breakfast) = $1-(2/5)^2+(3/5)^2 = 1-(0.16+0.36) = 0.48$

Gini index (Meal Type = Lunch) = 1- (4/4) ^{2}+(0/4) ^{2} = 1- (1+ 0) = 0

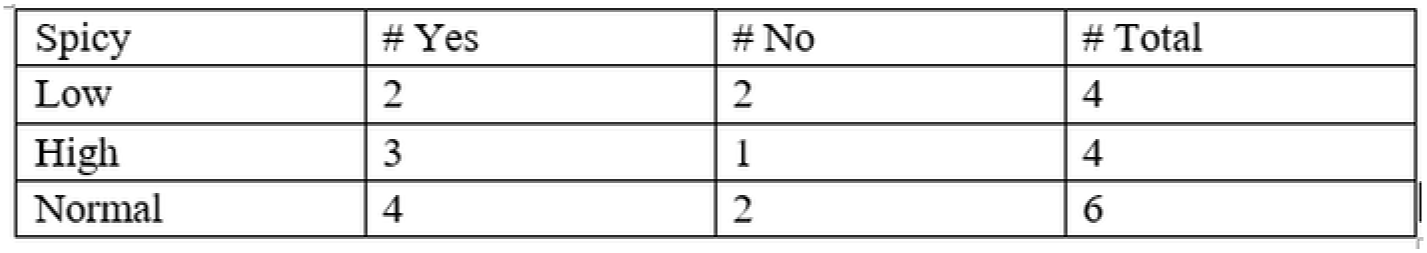
Gini index (Meal Type = Dinner) = 1- (3/5) ^{2} +(2/5) ^{2} = 1- (0.36+0.16) = 0.48

Now, we will calculate the weighted sum of Gini index for Meal Type features,

Gini (Meal Type) = (5/14) *0.48 + (4/14) *0 + (5/14) *0.48 = 0.342

Spicy

Spicy is a nominal data that has 3 values Low, Normal and High. Let’s classify the instances on basis of liked/dislike.



Spicy	# Yes	# No	# Total
Low	2	2	4
High	3	1	4
Normal	4	2	6

Gini (Spicy = Low) = 1-(2/4) ^{2}+(2/4) ^{2} = 0.5

Gini (Spicy = High) = 1-(3/4) ^{2}+(1/4) ^{2} = 0.375

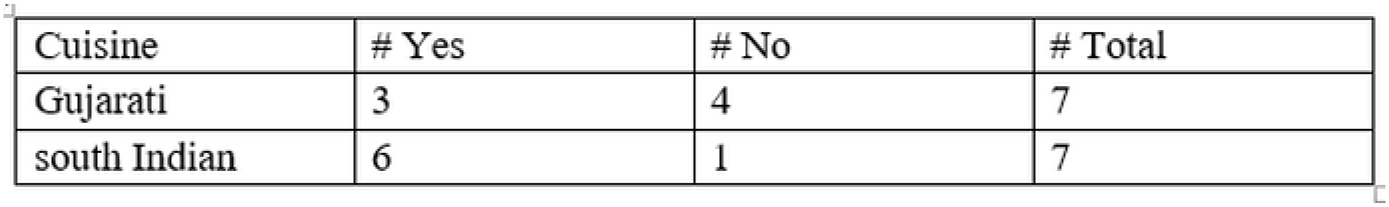
Gini (Spicy = Normal) = 1-(4/6) ^{2}+(2/6) ^{2} = 0.445

Now, the weighted sum of Gini index for Spicy features can be calculated as,

Gini (Spicy)= (4/14) *0.5 + (4/14) *0.375 + (6/14) *0.445 =0.439

Cuisine

The cuisine is a binary data that has 2 values Gujarati and south Indian. Let’s classify the instances on the basis of liked/dislike.



Cuisine	# Yes	# No	# Total
Gujarati	3	4	7
south Indian	6	1	7

Gini (Cuisine = Gujarati) = 1-(3/7) ^{2}+(4/7) ^{2} = 0.489

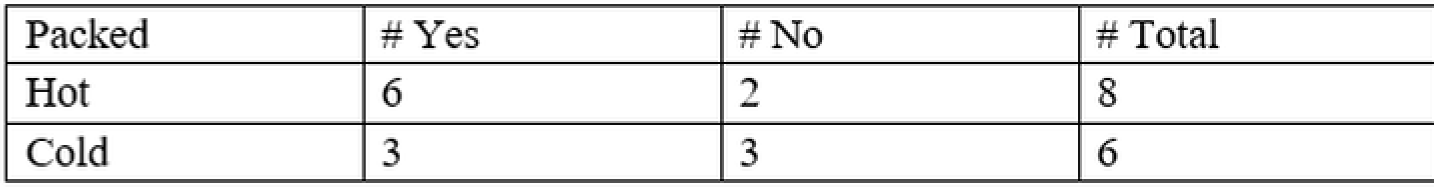
Gini (Cuisine =south Indian) = 1-(6/7) ^{2}+(1/7) ^{2} = 0.244

Now, the weighted sum of the Gini index for Cuisine features can be calculated as,

Gini (Cuisine) = (7/14) *0.489 + (7/14) *0.244=0.367

Packed

Packed is a binary data that has 2 values Hot and cold. Let’s classify the instances on the basis of liked/dislike.



Packed	# Yes	# No	# Total
Hot	6	2	8
Cold	3	3	6

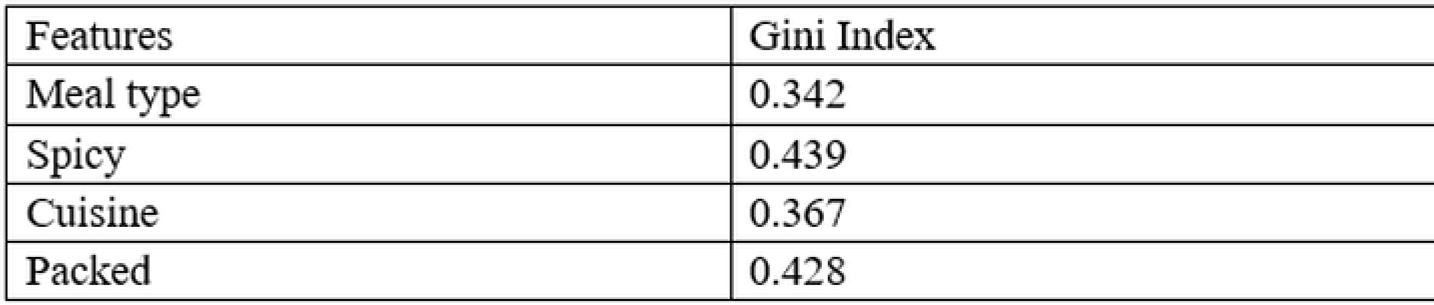
Gini (Packed = Hot) = 1-(6/8) ^{2}+(2/8) ^{2} = 0.375

Gini (Packed = Cold) = 1-(3/6) ^{2}+(3/6) ^{2}= 0.5

Now, the weighted sum of the Gini index for Packed features can be calculated as,

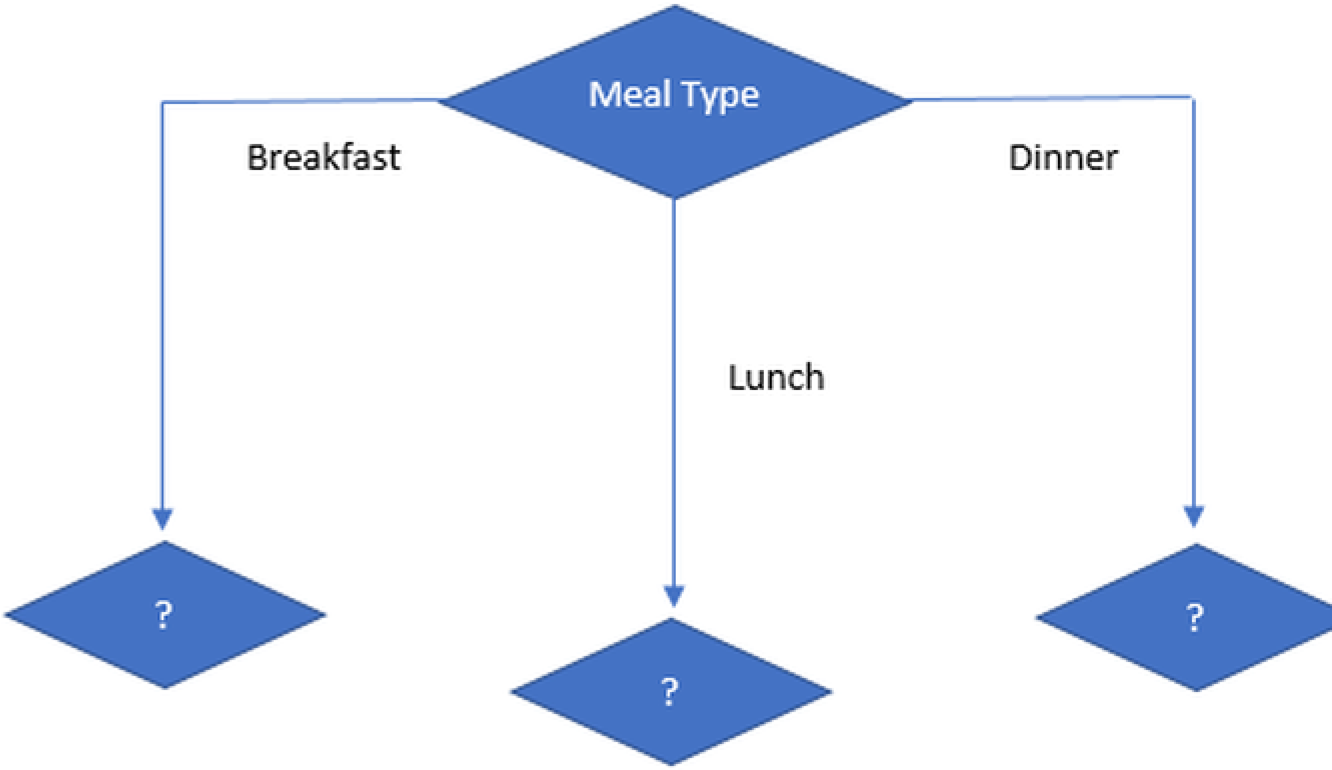
Gini (Packed) = (8/14) *0.375 + (6/14) *0.5=0.428

So, the Gini index for all the feature is:



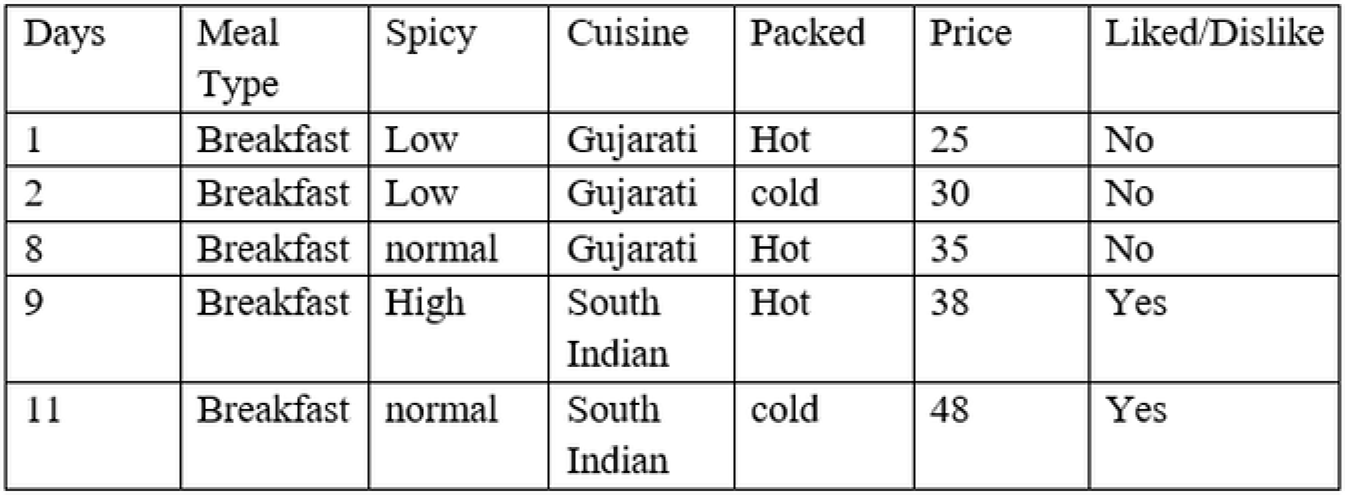
Features	Gini Index
Meal type	0.342
Spicy	0.439
Cuisine	0.367
Packed	0.428

So, we can conclude that the lowest Gini index is of “Meal Type” and a lower Gini index means the highest purity and more homogeneity. So, our root node is “Meal type”. So, our tree looks like



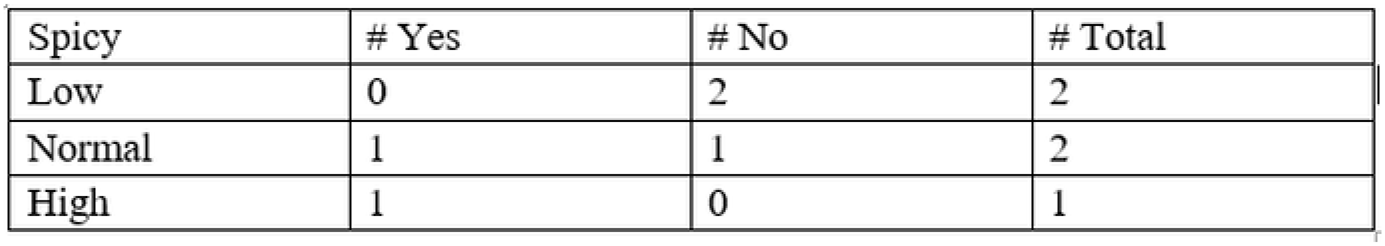
Let’s calculate the next split with the Gini index on the sub data set for the Meal Type feature, we will use the same method as above to find the next split.

Let’s find the Gini index of spicy, cuisine and packed on sub-data of Meal type = Breakfast.



Days	Meal Type	Spicy	Cuisine	Packed	Price	Liked/Dislike
1	Breakfast	Low	Gujarati	Hot	25	No
2	Breakfast	Low	Gujarati	cold	30	No
8	Breakfast	normal	Gujarati	Hot	35	No
9	Breakfast	High	South Indian	Hot	38	Yes
11	Breakfast	normal	South Indian	cold	48	Yes

Gini index for Spicy on breakfast meal type



Spicy	# Yes	# No	# Total
Low	0	2	2
Normal	1	1	2
High	1	0	1

Gini (Meal type = Breakfast & Spicy = Low) = $1-(0/2)^2+(2/2)^2 = 0$

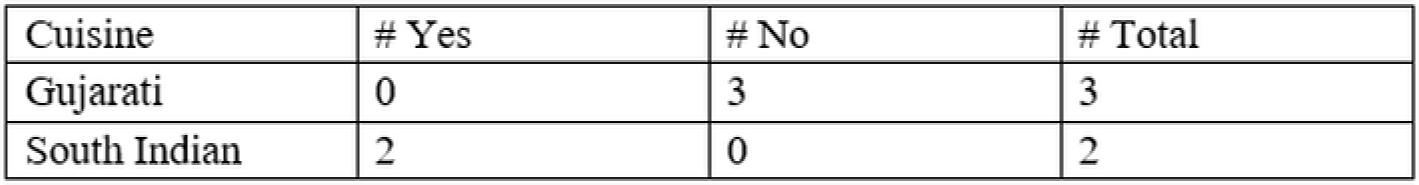
Gini (Meal type = Breakfast & Spicy = High) = $1-(1/1)^2+(0/1)^2 = 0$

Gini (Meal type = Breakfast & Spicy = Normal) = $1-(1/2)^2+(1/2)^2 = 0.5$

Now, the weighted sum of Gini index for temperature on sunny outlook features can be calculated as,

Gini (Meal type = Breakfast & Spicy) = $(2/5) *0 + (1/5) *0+ (2/5) *0.5 =0.2$

Gini index for cuisine on breakfast meal type



Cuisine	# Yes	# No	# Total
Gujarati	0	3	3
South Indian	2	0	2

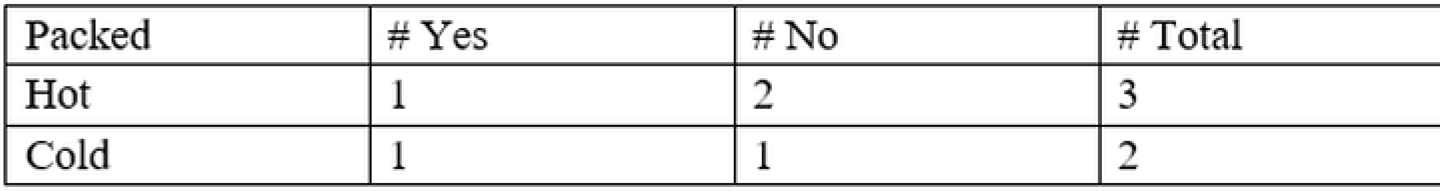
Gini (Meal type = Breakfast & Cuisine = Gujarati) = $1-(0/3)^2+(3/3)^2 = 0$

Gini (Meal type = Breakfast & Cuisine = South Indian) = $1-(2/2)^2+(0/2)^2 = 0$

Now, the weighted sum of Gini index for humidity on sunny outlook features can be calculated as,

Gini (Meal type = Breakfast & Cuisine) = $(3/5) *0 + (2/5) *0=0$

Gini index for packed on breakfast meal type



Packed	# Yes	# No	# Total
Hot	1	2	3
Cold	1	1	2

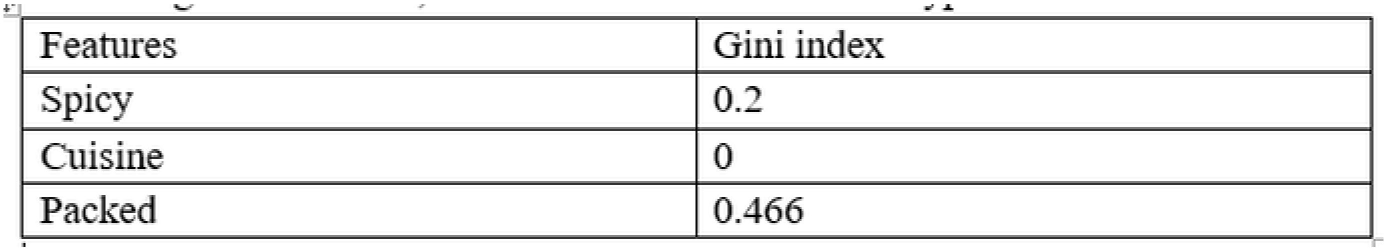
Gini (Meal type = Breakfast & Packed = hot) = $1-(1/3)^2+(2/3)^2 = 0.44$

Gini (Meal type = Breakfast & Packed = cold) = $1-(1/2)^2+(1/2)^2 = 0.5$

Now, the weighted sum of Gini index for wind on sunny outlook features can be calculated as,

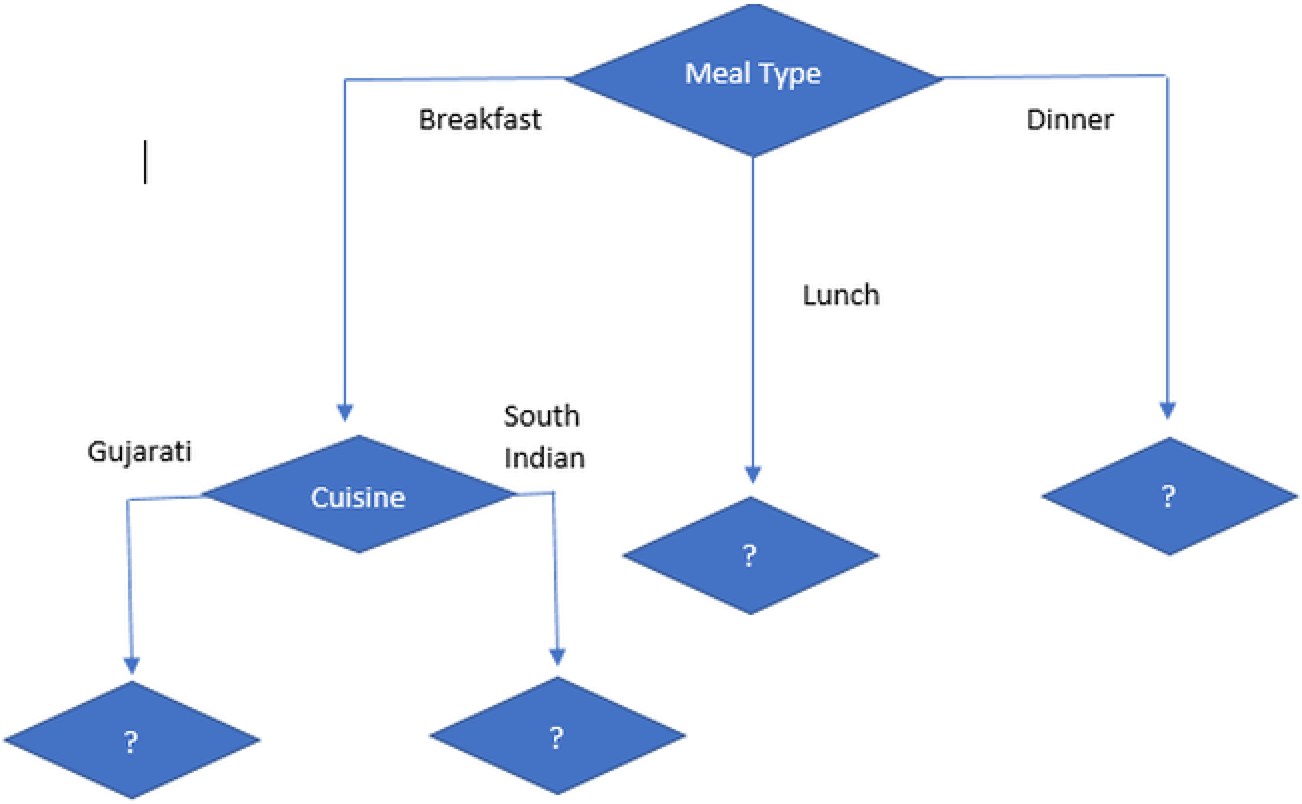
Gini (Meal type = Breakfast and Packed) = (3/5) *0.44 + (2/5) *0.5=0.266+0.2= 0.466

According to the Gini index, Decision on Breakfast Meal Type is:

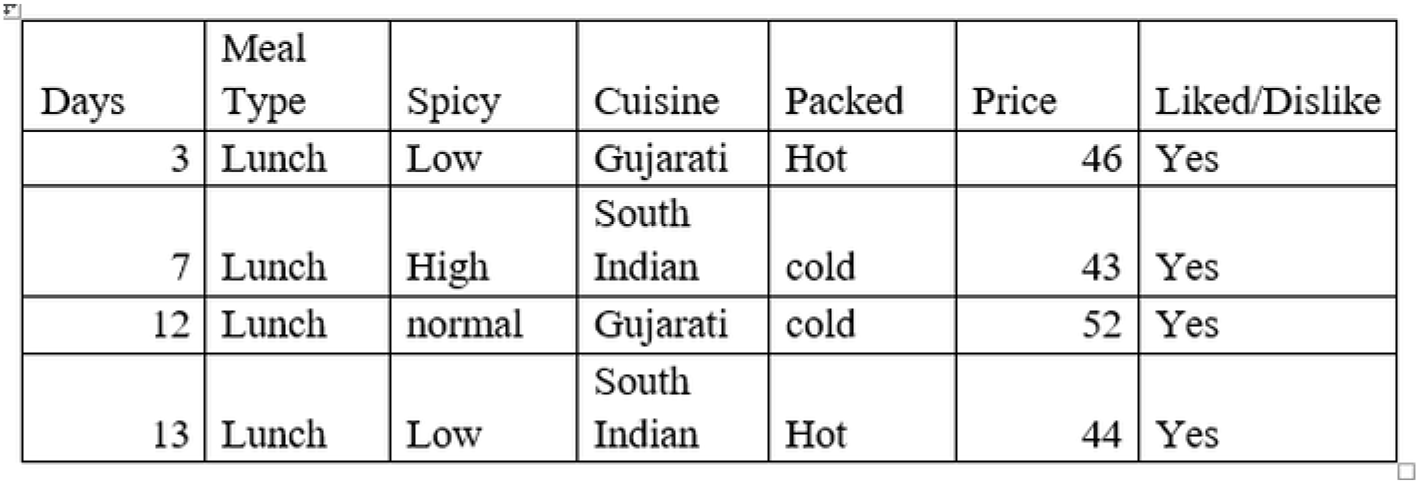


Features	Gini index
Spicy	0.2
Cuisine	0
Packed	0.466

As we can see for the breakfast meal type, the cuisine has the lowest Gini value that is highly homogenous and highest pure amongst other features, so we can conclude that the next node will be cuisine. So, the tree will be like:

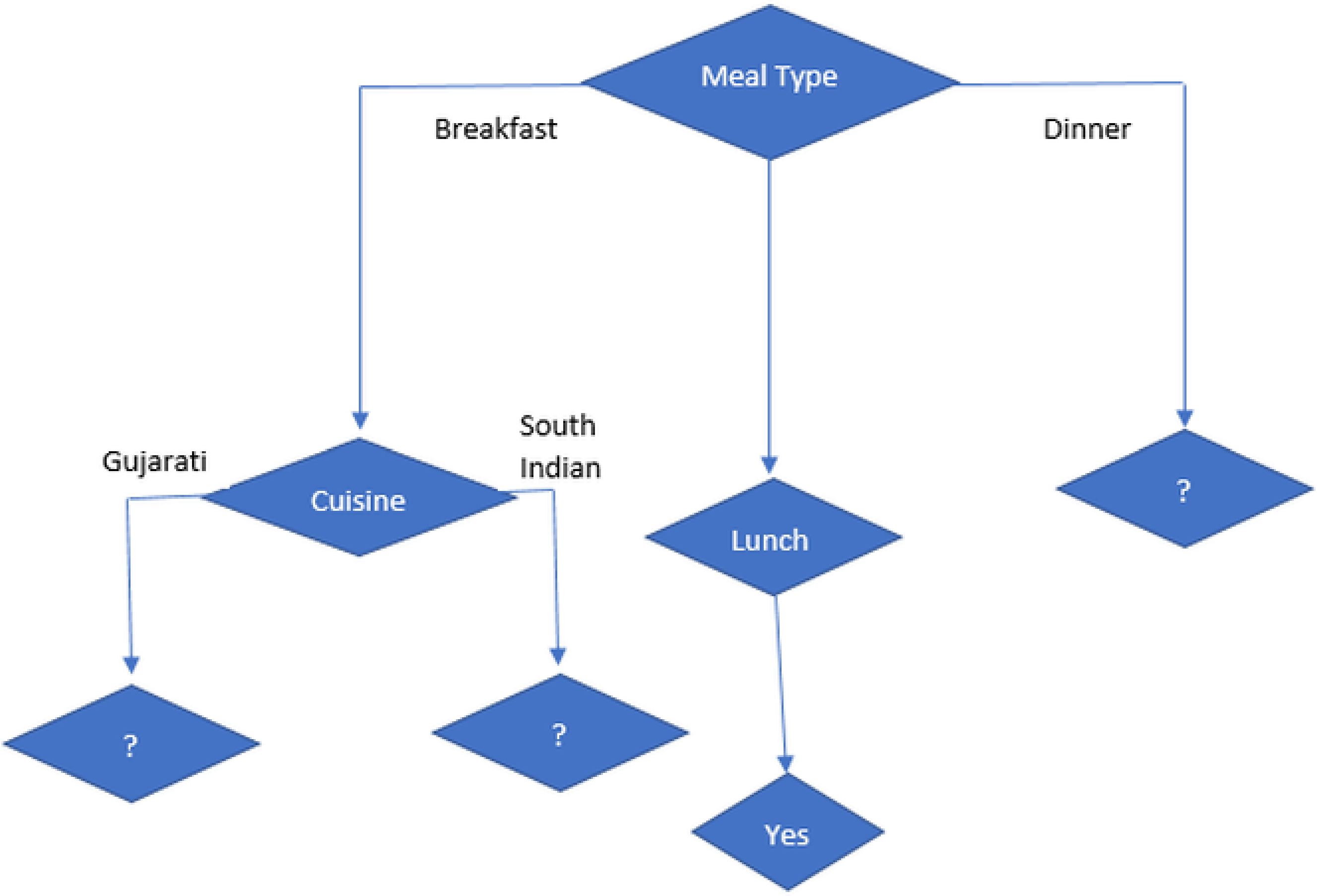


Now let’s focus on sub-data of Meal Type = Lunch

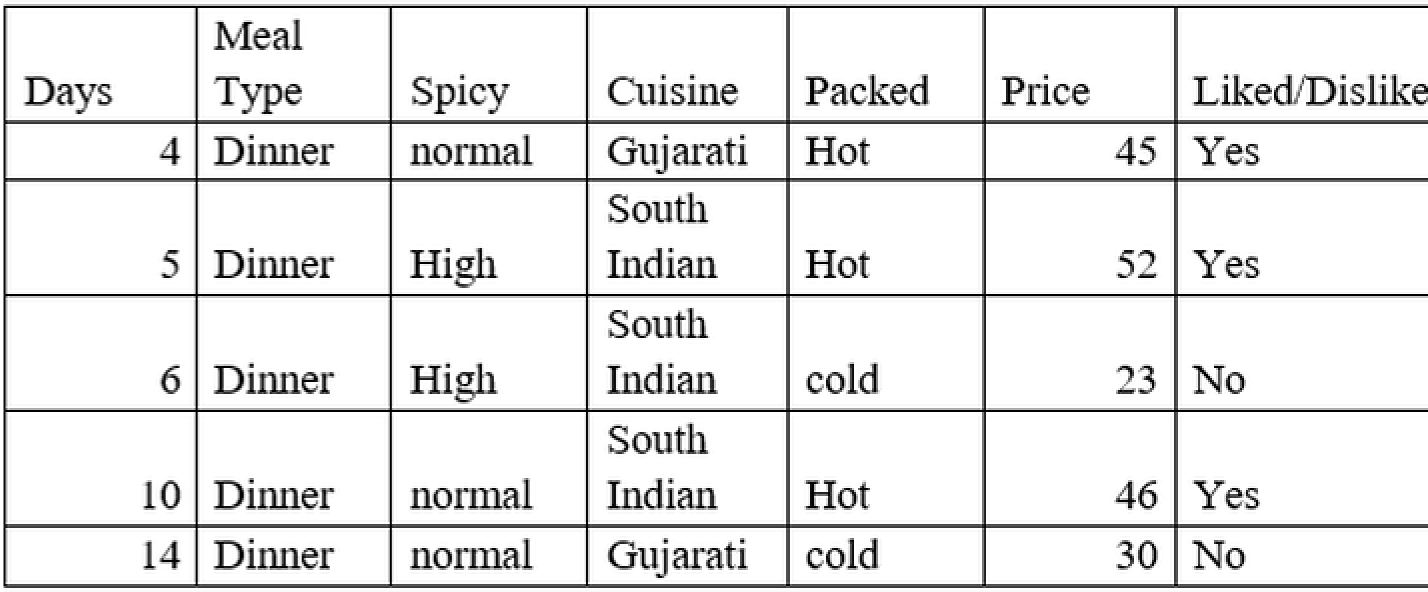


Days	Meal Type	Spicy	Cuisine	Packed	Price	Liked/Dislike
3	Lunch	Low	Gujarati	Hot	46	Yes
7	Lunch	High	South Indian	cold	43	Yes
12	Lunch	normal	Gujarati	cold	52	Yes
13	Lunch	Low	South Indian	Hot	44	Yes

As we can see for Meal Type = Lunch, the target variable is “Yes” for all so the Gini index is 0 that is there is no impurity and it is highly homogenous. So, it’s a leaf node.

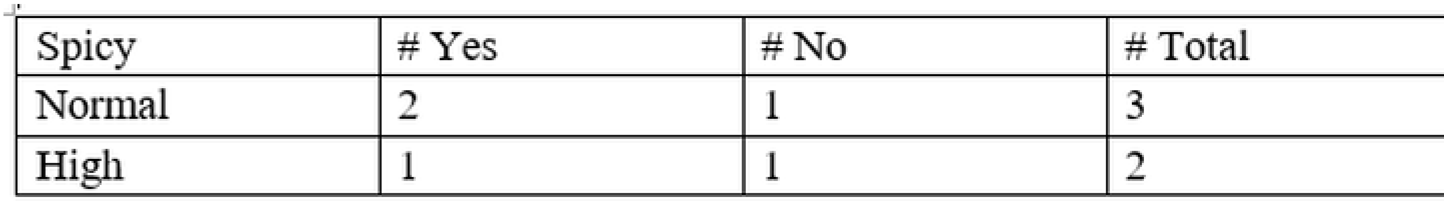


Now, let’s focus on Meal Type = Dinner and find the Gini index for spicy, cuisine and packed.



Days	Meal Type	Spicy	Cuisine	Packed	Price	Liked/Dislike
4	Dinner	normal	Gujarati	Hot	45	Yes
5	Dinner	High	South Indian	Hot	52	Yes
6	Dinner	High	South Indian	cold	23	No
10	Dinner	normal	South Indian	Hot	46	Yes
14	Dinner	normal	Gujarati	cold	30	No

Gini index for spicy on meal type = Dinner



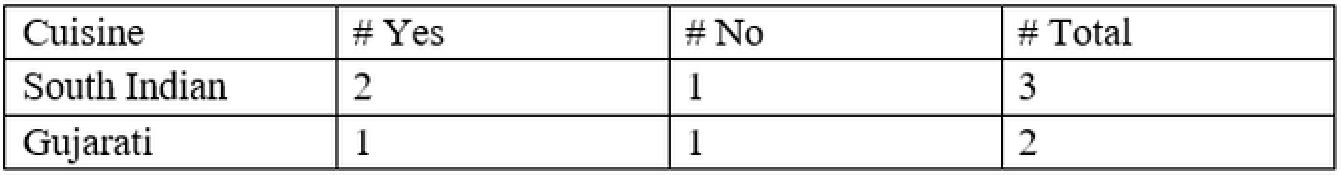
Spicy	# Yes	# No	# Total
Normal	2	1	3
High	1	1	2

Gini (meal type = Dinner and Spicy= High) = 1 – (1/2)2 + (1/2)2 = 0.5

Gini (meal type = Dinner and Spicy = Normal) = 1 – (2/3)2 + (1/3)2 = 0.444

Gini (meal type = Dinner and Spicy) = (2/5) *0.5 + (3/5) *0.444 = 0.466

Gini index for cuisine on meal type = Dinner



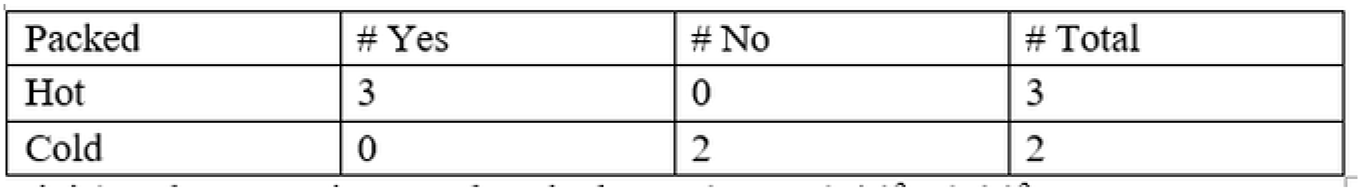
Cuisine	# Yes	# No	# Total
South Indian	2	1	3
Gujarati	1	1	2

Gini (meal type = Dinner and Cuisine = Gujarati) = 1 – (1/2)2 + (1/2)2 = 0.5

Gini (meal type = Dinner and Cuisine = South Indian) = 1 – (2/3)2 + (1/3)2 = 0.444

Gini (meal type = Dinner and Cuisine) = (2/5) *0.5 + (3/5) *0.444 = 0.466

Gini index for Packed on meal type = Dinner



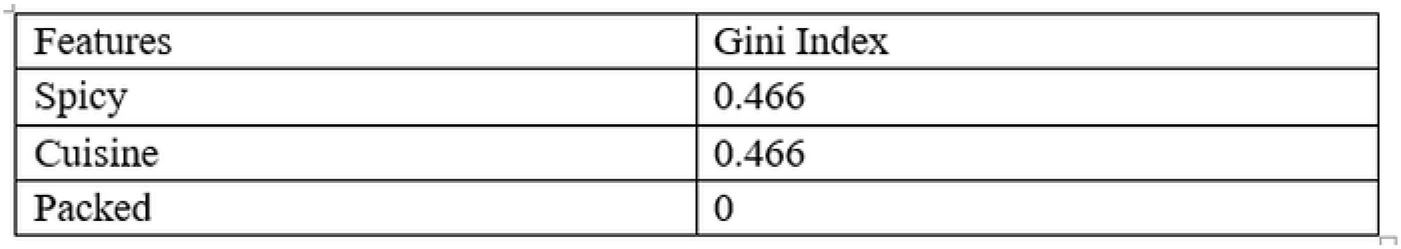
Packed	# Yes	# No	# Total
Hot	3	0	3
Cold	0	2	2

Gini (meal type = Dinner and Packed = Hot) = 1 – (3/3)2 + (0/3)2 = 0

Gini (meal type = Dinner and Packed = Cold) = 1 – (0/2)2 + (2/2)2 = 0

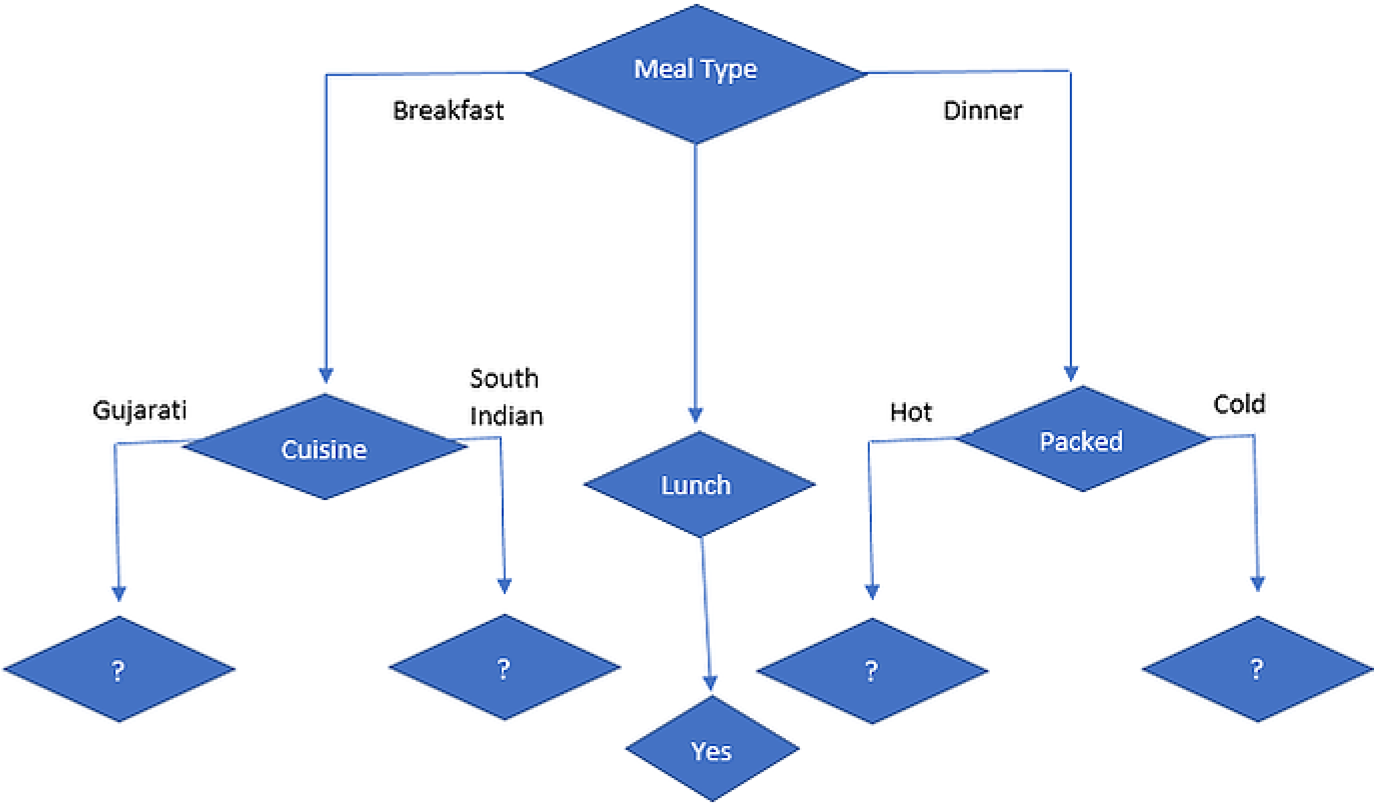
Gini (meal type = Dinner and Packed) = (3/5) *0 + (2/5) *0 = 0

The decision on Meal Type = Dinner



Features	Gini Index
Spicy	0.466
Cuisine	0.466
Packed	0

So, packed has the lowest Gini value, so the next node will be packed and the following is a decision tree.



Now, let’s focus on sub-data of:

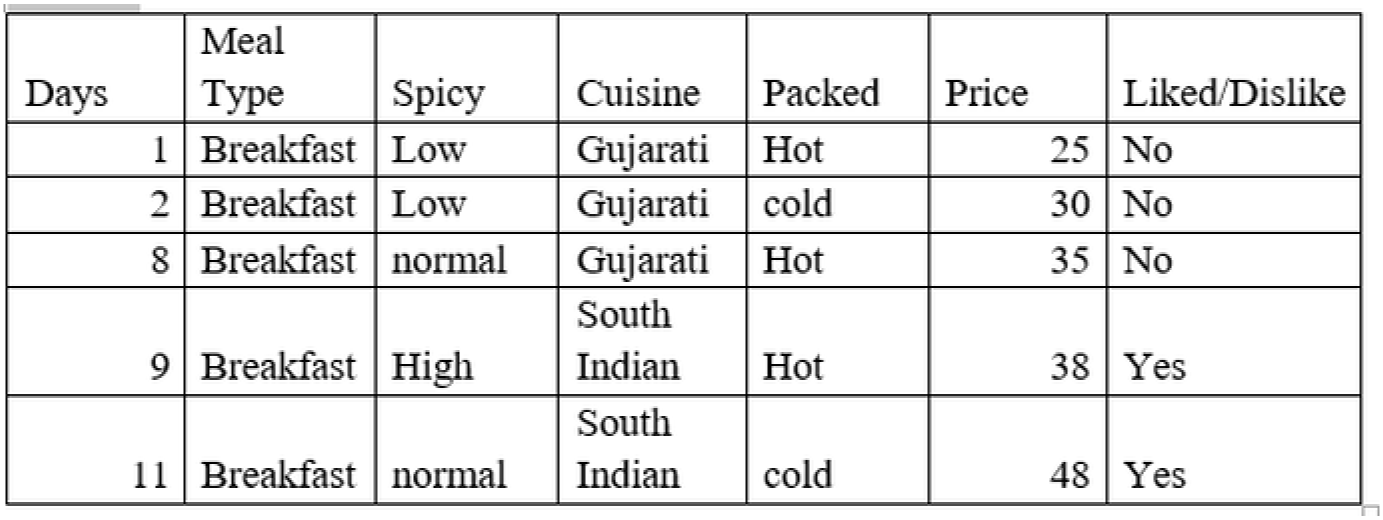
1. Cuisine

   - Gujarati
   - South Indian

2. Packed

   - Hot
   - Cold

First, we will focus on Meal Type= Breakfast and Gujarati and south Indian cuisine



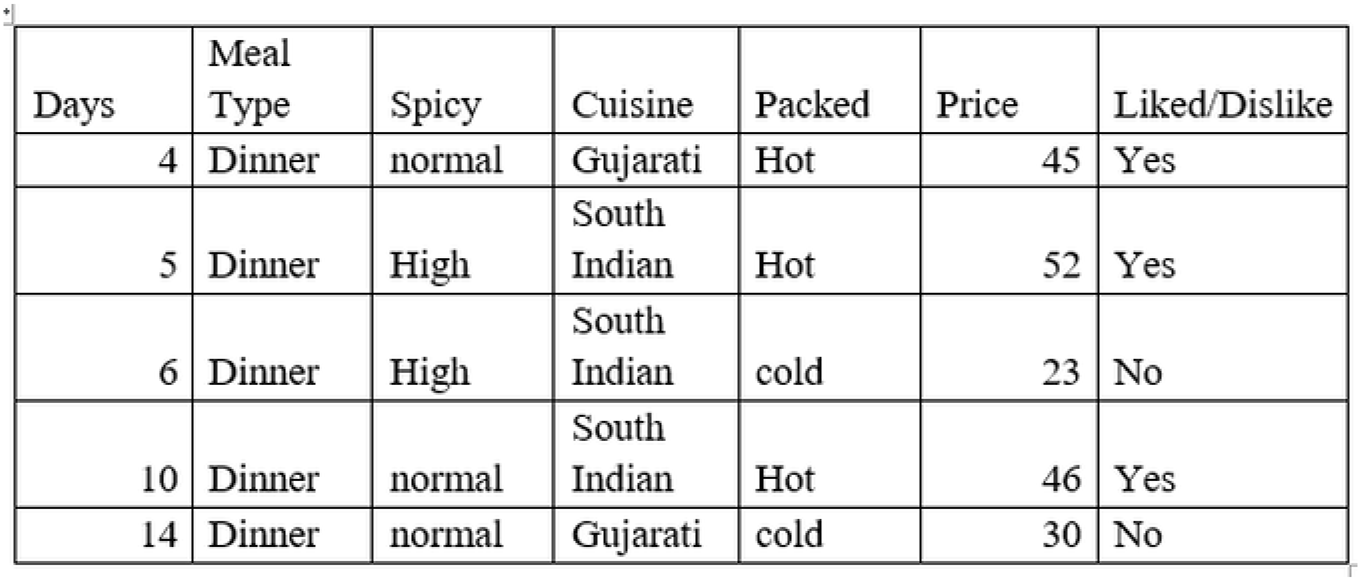
Days	Meal Type	Spicy	Cuisine	Packed	Price	Liked/Dislike
1	Breakfast	Low	Gujarati	Hot	25	No
2	Breakfast	Low	Gujarati	cold	30	No
8	Breakfast	normal	Gujarati	Hot	35	No
9	Breakfast	High	South Indian	Hot	38	Yes
11	Breakfast	normal	South Indian	cold	48	Yes

As we can see that when Meal Type = Breakfast and Cuisine = Gujarati then the decision is always No

And when Meal Type = Breakfast and Cuisine = South Indian then the decision is always Yes.

So we got the leaf nodes.

Now we will focus on Meal Type= Dinner and hot and cold packed



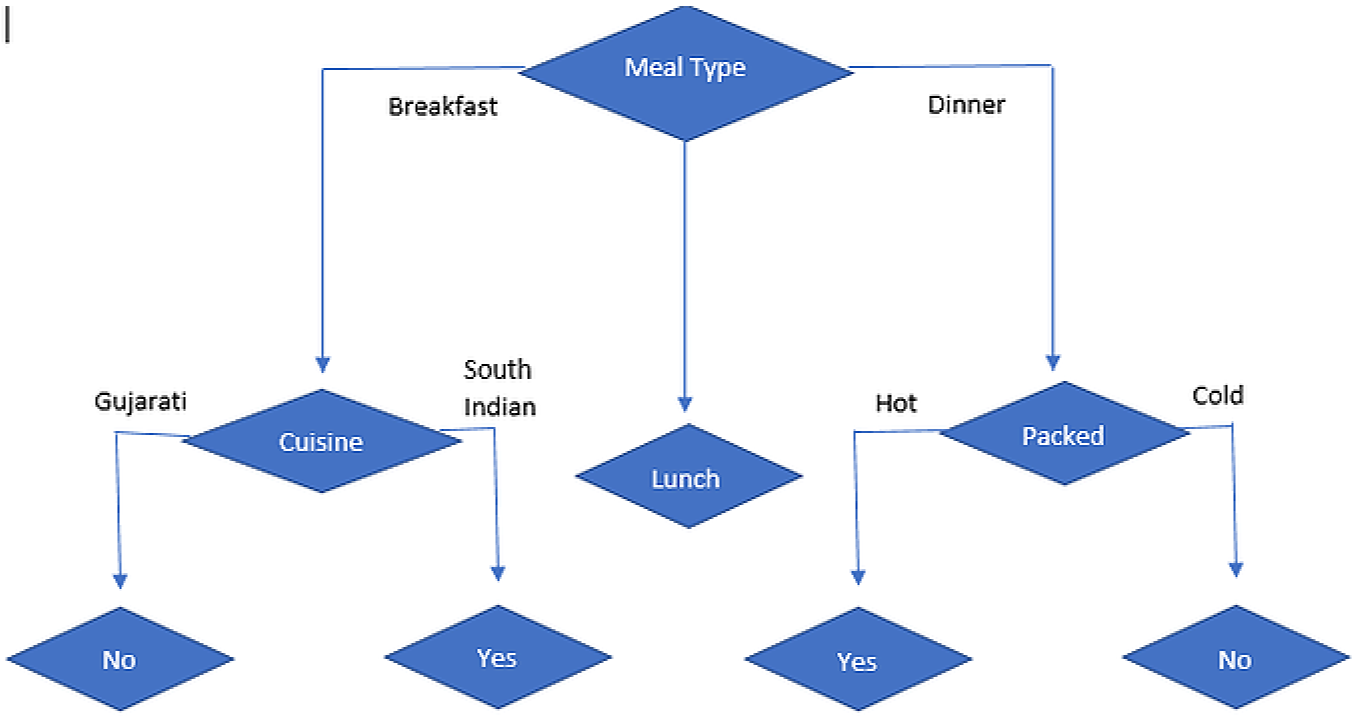
Days	Meal Type	Spicy	Cuisine	Packed	Price	Liked/Dislike
4	Dinner	normal	Gujarati	Hot	45	Yes
5	Dinner	High	South Indian	Hot	52	Yes
6	Dinner	High	South Indian	cold	23	No
10	Dinner	normal	South Indian	Hot	46	Yes
14	Dinner	normal	Gujarati	cold	30	No

As we can see that when Meal Type = Dinner and Packed = Hot then the decision is always Yes

And when Meal Type = Dinner and Packed = Cold then the decision is always No.

So, we got the leaf nodes.

The following is our final classification decision tree:


Meal Type
Breakfast
Dinner
Gujarati
Cuisine
South Indian
Lunch
Hot
Packed
Cold
No
Yes
Yes
No