Data Science PROPHET

UNDERSTANDING THE MATHEMATICS BEHIND THE DECISION TREE ALGORITHM (PART – I)

Vrutti Tanna

Jan 31 , 2020

BACK

Introduction

From classrooms to corporate, one of the first lessons in machine learning involves decision trees. Mathematics behind decision tree is very easy to understand compared to other machine learning algorithms. Decision tree is also easy to interpret and understand compared to other ML algorithms.

If you are just getting started with machine learning, it's very easy to pick up decision trees. In this tutorial, you'll learn:

Data Science PROPHET

3. Decision tree types

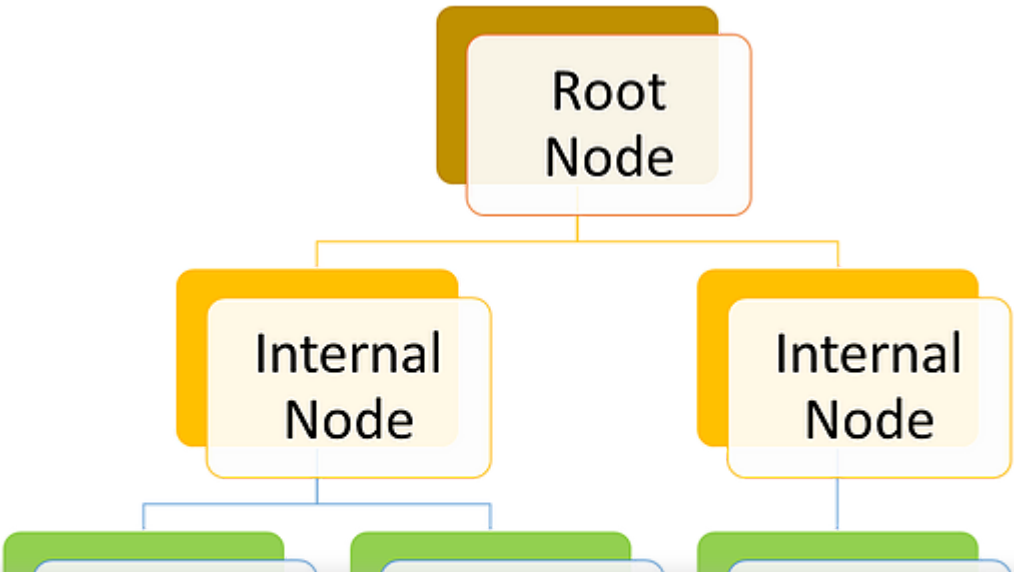
- Classification and regression tree – CART
- ID 3

What is a decision tree?

Decision tree is a supervised learning algorithm that works for both categorical and continuous input and output variables that is we can predict both categorical variables (classification tree) and a continuous variable (regression tree).

Its graphical representation makes human interpretation easy and helps in decision making.

Decision tree is a flow chart like structure with the if-else condition. The top node is the root node, you start question from root node then move through the tree branches according to which groups you belong to until you reach a leaf node.



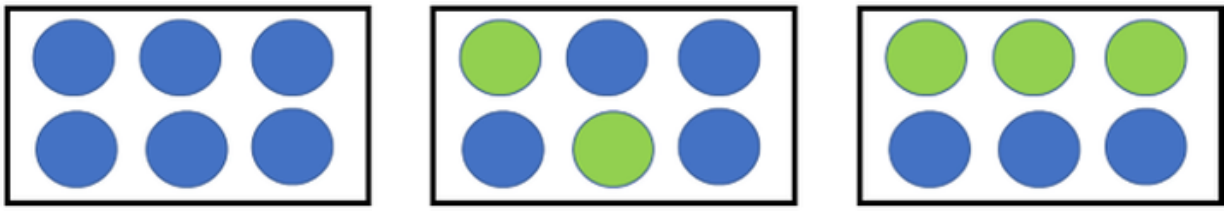
Data Science PROPHET

Algorithms used in constructing decision tree

For understanding decision tree algorithm better, let us first understand “impurity” and types of measures of impurity.

What is impurity?

Let us understand impurity from the below image.



Square 1 Square 2 Square 3

So, there are two round objects of blue and green color inside a square and there are 3 squares. So now how much information I need in each square to accurately identify the color of round objects. So, square 1 needs less information as all objects are blue, square 2 needs little more information than square 1 to tell accurately the color of the object and square 3 requires maximum information as both blue and green objects are equal in number.

As information is a measure of purity, square 1 is a pure node, square 2 is less impure and square 3 is more impure.

So how to measure impurity in data?

In this article, we are going to look at two such impurity measure

So, square 1 has the lowest entropy, square 2 has more entropy and square 3 has the highest entropy.

Mathematically it is written as:

$$Entropy = -\sum_{i=1}^{n} p_i * \log(p_i)$$

Gini index/Gini impurity

It measures impurity in the node. It has a value between 0 and 1. So the Gini index of value 0 means sample are perfectly homogeneous and all elements are similar, whereas, Gini index of value 1 means maximal inequality among elements. It is sum of the square of the probabilities of each class. It is illustrated as,

$$Gini\,index = 1 - \sum_{i=1}^{n} p_i^2$$

So, (Im)purity measures homogeneity in data and if data is homogenous then it belongs to the same class and decision tree splits on homogeneity.

Decision tree types

- In this article, we will talk about CART and ID3 that are mostly used in the industry.

CART

It is used for both classification and regression. Let us understand classification (CA) and regression tree (RT)

1. Classification tree

- A decision tree where target variable is categorical
- The algorithm classifies the class within which the target variable would most likely to fall
- It uses Gini index as metric/cost function to evaluate split in feature selection
- Example like predicting who will or who will not order food, whether the weather will be rainy, sunny or cool

2. Regression tree

- A decision tree where the target variable is continuous/discrete
- Algorithm predicts value
- It uses least square / standard deviation reduction as a metric to select features in case of the Regression tree.
- Example like predicting the price of a house, predicting the sell of crops

Let's start with the classification tree.

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

Data Science PROPHET

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) = $1-(2/5)^2+(3/5)^2 = 1- (0.16+0.36) = 0.48$

Data Science PROPHET

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

Data Science PROPHET

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) = $1-(3/6)^2+(3/6)^2= 0.5$

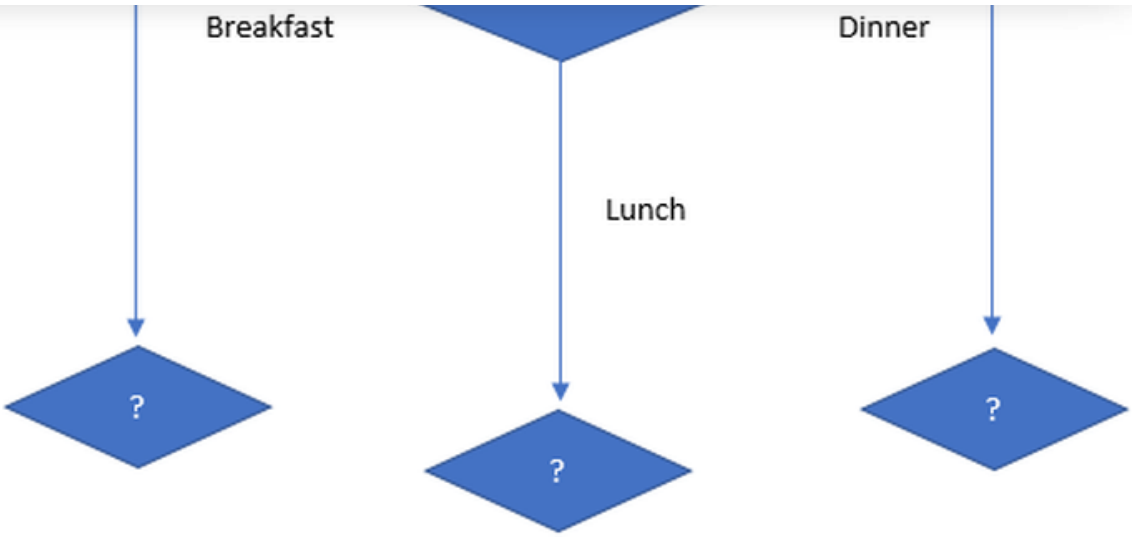
Now, the weighted sum of the Gini index for Packed features can be calculated as,

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

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

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) = $1-(1/2)^2+(1/2)^2 = 0.5$

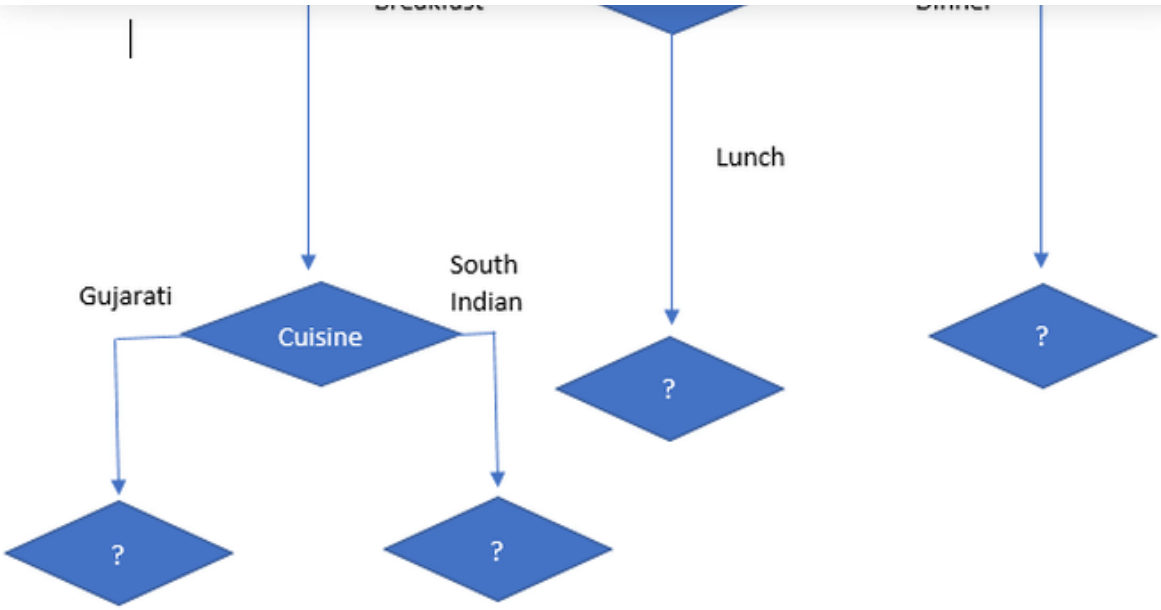
Now, the weighted sum of Gini index for wind on sunny outlook features can be calculated as,

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

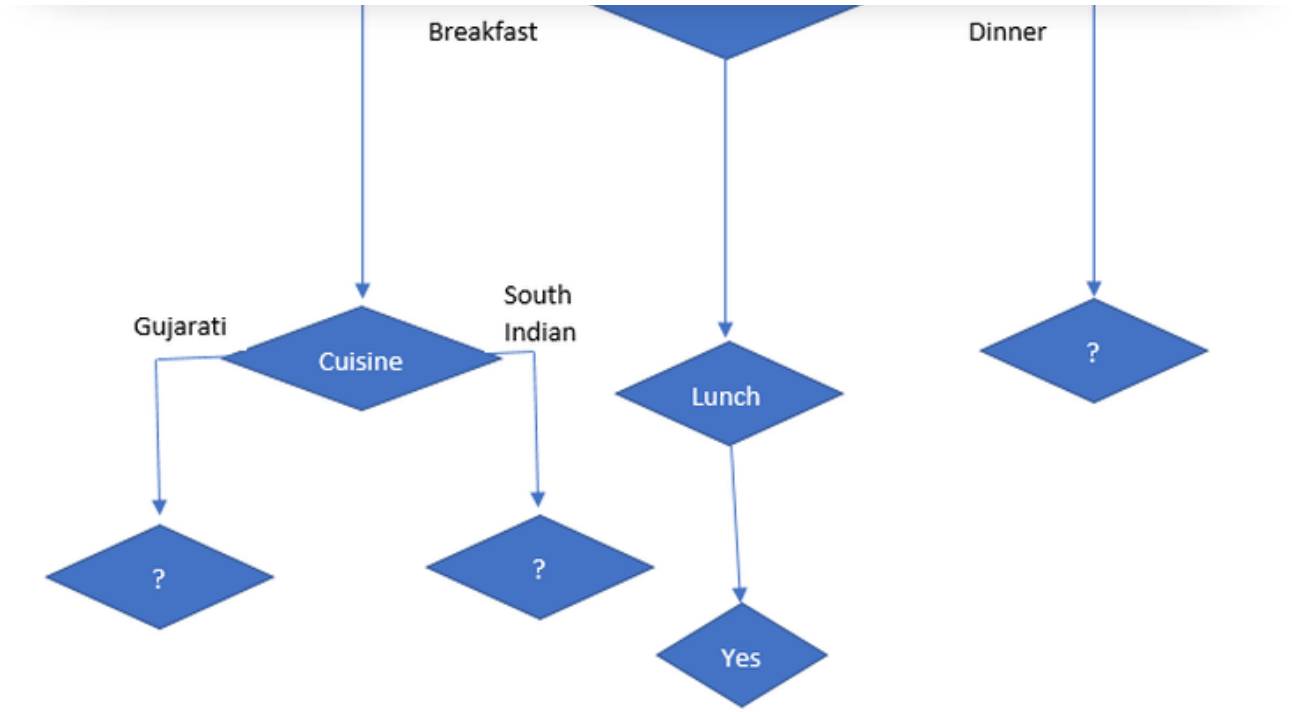
As we can see for the breakfast meal type, the cuisine has the lowest Gini value that is highly homogenous and highest pure amongst other features, so we can conclude that the next node will be cuisine. So, the tree will be like:



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

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

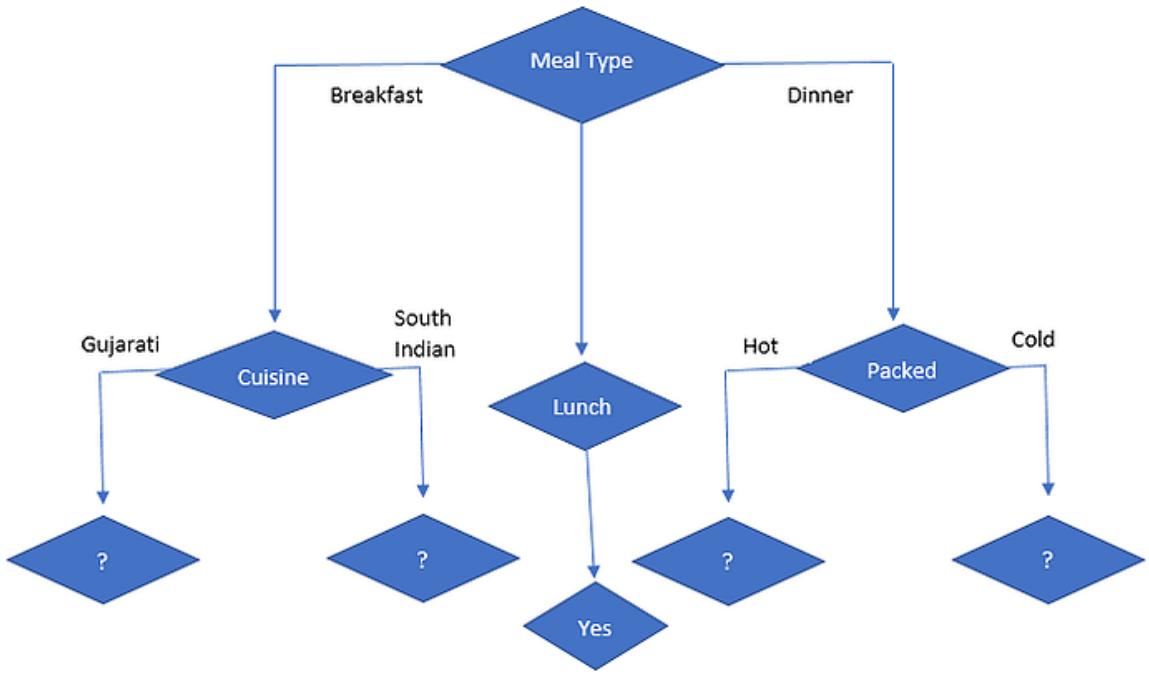
Gini index for Packed on meal type = Dinner

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

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

Data Science PROPHET

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

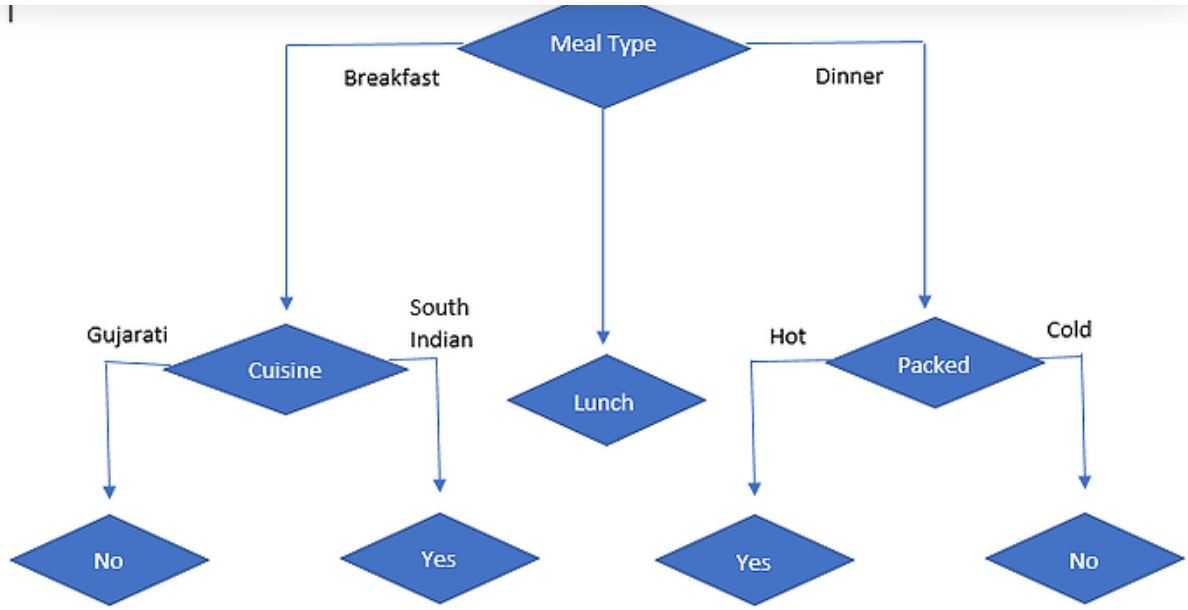
And when Meal Type = Breakfast and Cuisine = South Indian then the decision is always Yes.

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

Data Science PROPHET



We will continue with the regression tree and ID3 tree in the next articles.

Happy Learning.

19 Likes

Leave a comment

You must be **logged in** to post a comment.

QUICK LINKS

About Us

Data Science PROPHET

Statistics

predictive modelling

AI

Machine Learning

Terms of use

Sitemap

Subscribe