კლასიფიკაციის ამოცანა გადაწყვეტილების ხეებით

Decision Trees

მოცემული გვაქვს 30 სტუდენტისაგან შემდგარი ჯგუფი, რომლის თითოეული წევრი ხასიათდება 3 ცვლადით:

**Gender (Boy / Girl)**

**Class (IX / X)**

**Height (5 to 6 feet)**

ცნობილია რომ 30 სტუდენტიდან 15 თამაშობს კრიკეტს თავისუფალ დროს.

ამასთან ჩვენთვის ცნობილია თუ კონკრეტულად რომელი სტუდენტი თამაშობს კრიკეტს.

ჩვენი ამოცანაა ავაგოთ მოდელი რომელიც საშუალებას მოგვცემს არსებულ მონაცემებზე დაყრდნობით ვიწინასწარმეტყველოთ ითამაშებს თუ არა სტუდენტი (რომელიც არ განეკუთვნება ზემოთ მოცემული სტუდენტების ჯგუფს) კრიკეტს თავისუფალ დროს.

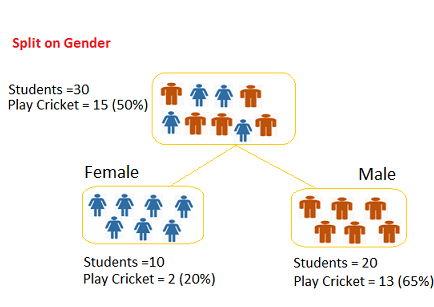
აღნიშნული ამოცანის გადასაწყვეტად ჩვენ შევეცდებით ავაგოთ სხვადასხვა სახის გადაწყვეტილების ხეები.

ამისათვის ჩვენ შევეცდებით სტუდენტების ჯგუფი დავყოთ სხვადასხვა კატეგორიებად.

**სტუდენტების დაყოფა სქესის სქესის მიხედვით:**

როგორც უკვე ვთქვით ჩვენ ვიცით კონკრეტულად რომელი სტუდენტი თამაშობს კრიკეტს და ვთქვათ აღმოჩნდა რომ კრიკეტის მოთამაშეებიდან 2 გოგონაა და 13 ბიჭი.

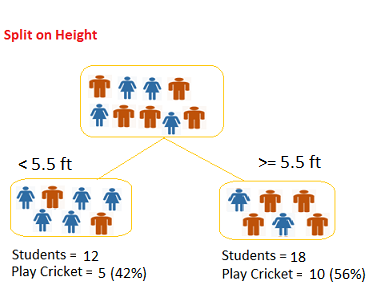
ასეთ შემთხვევაში დაყოფას სქესის მიხედვით და ალბათობებს იმისა თუ რომელ ჯგუფში რამდენია ალბათობა იმისა რომ ჯგუფის წევრი თამაშობს კრიკეტს ექნება შემდეგი სახე:



**სტუდენტების დაყოფა სიმაღლის მიხედვით:**

როგორც უკვე ვთქვით ჩვენ ვიცით კონკრეტულად რომელი სტუდენტი თამაშობს კრიკეტს და ვთქვათ აღმოჩნდა რომ კრიკეტის მოთამაშეებიდან 5 -ის სიმაღლე ნაკლებია 5.5 ft -ზე ხოლო 10-ის სიმაღლე მეტია 5.5 ft -ზე.

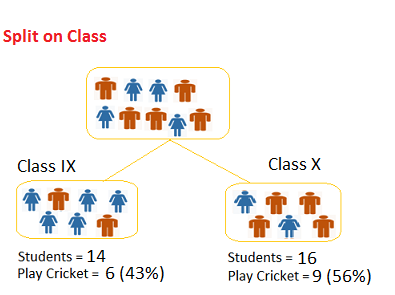
ასეთ შემთხვევაში დაყოფას სიმაღლის მიხედვით და ალბათობებს იმისა თუ რომელ ჯგუფში რამდენია ალბათობა იმისა რომ ჯგუფის წევრი თამაშობს კრიკეტს ექნება შემდეგი სახე:



**სტუდენტების დაყოფა კლასების მიხედვით:**

როგორც უკვე ვთქვით ჩვენ ვიცით კონკრეტულად რომელი სტუდენტი თამაშობს კრიკეტს და ვთქვათ აღმოჩნდა რომ კრიკეტის მოთამაშეებიდან 6 სწავლობს IX კლასში ხოლო 9 - X კლასში

ასეთ შემთხვევაში დაყოფას კლასების მიხედვით და ალბათობებს იმისა თუ რომელ ჯგუფში რამდენია ალბათობა იმისა რომ ჯგუფის წევრი თამაშობს კრიკეტს ექნება შემდეგი სახე:



Gini Index

Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure (p²+q²).

**Split on Gender:**

1. Calculate, Gini for sub-node Female = (0.2)\*(0.2)+(0.8)\*(0.8)=0.68
2. Gini for sub-node Male = (0.65)\*(0.65)+(0.35)\*(0.35)=0.55
3. Calculate weighted Gini for Split Gender = (10/30)\*0.68+(20/30)\*0.55 = **0.59**

**Split on Class:**

1. Gini for sub-node Class IX = (0.43)\*(0.43)+(0.57)\*(0.57)=0.51
2. Gini for sub-node Class X = (0.56)\*(0.56)+(0.44)\*(0.44)=0.51
3. Calculate weighted Gini for Split Class = (14/30)\*0.51+(16/30)\*0.51 = **0.51**

Above, you can see that Gini score for Split on Gender is higher than Split on Class, hence, the node split will take place on Gender

**Chi-Square**

It is an algorithm to find out the statistical significance between the differences between sub-nodes and parent node. We measure it by sum of squares of standardized differences between observed and expected frequencies of target variable.

1. It works with categorical target variable “Success” or “Failure”.
2. It can perform two or more splits.
3. Higher the value of Chi-Square higher the statistical significance of differences between sub-node and Parent node.
4. Chi-Square of each node is calculated using formula,
5. Chi-square = ((Actual – Expected)^2 / Expected)^1/2
6. It generates tree called CHAID (Chi-square Automatic Interaction Detector)

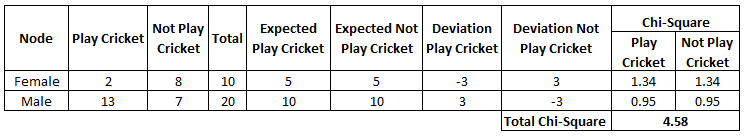
**Steps to Calculate Chi-square for a split:**

1. Calculate Chi-square for individual node by calculating the deviation for Success and Failure both
2. Calculated Chi-square of Split using Sum of all Chi-square of success and Failure of each node of the split

**Example:** Let’s work with above example that we have used to calculate Gini.

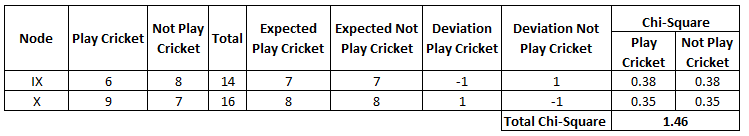
**Split on Gender:**

1. First we are populating for node Female, Populate the actual value for “**Play Cricket”** and**“Not Play Cricket”**, here these are 2 and 8 respectively.
2. Calculate expected value for “**Play Cricket”** and “**Not Play Cricket”**, here it would be 5 for both because parent node has probability of 50% and we have applied same probability on Female count(10).
3. Calculate deviations by using formula, Actual – Expected. It is for “**Play Cricket”** (2 – 5 = -3) and for “**Not play cricket”** ( 8 – 5 = 3).
4. Calculate Chi-square of node for “**Play Cricket**” and “**Not Play Cricket**” using formula with formula, **= ((Actual – Expected)^2 / Expected)^1/2**. You can refer below table for calculation.
5. Follow similar steps for calculating Chi-square value for Male node.
6. Now add all Chi-square values to calculate Chi-square for split Gender.



**Split on Class:**

**Perform similar steps of calculation for split on Class and you will come up with below tabl**e.

[](https://www.analyticsvidhya.com/blog/wp-content/uploads/2015/01/Decision_Tree_Chi_Square_2.png)

**Above, you can see that Chi-square also identify the Gender split is more significant compare to Class.**

**Steps to calculate entropy for a split:**

1. Calculate entropy of parent node
2. Calculate entropy of each individual node of split and calculate weighted average of all sub-nodes available in split.

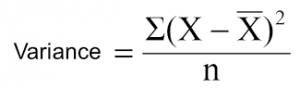
**Example:**Let’s use this method to identify best split for student example.

1. Entropy for parent node = -(15/30) log2 (15/30) – (15/30) log2 (15/30) = **1**. Here 1 shows that it is a impure node.
2. Entropy for Female node = -(2/10) log2 (2/10) – (8/10) log2 (8/10) = 0.72 and for male node,  -(13/20) log2 (13/20) – (7/20) log2 (7/20) = **0.93**
3. Entropy for split Gender = Weighted entropy of sub-nodes = (10/30)\*0.72 + (20/30)\*0.93 = **0.86**
4. Entropy for Class IX node, -(6/14) log2 (6/14) – (8/14) log2 (8/14) = 0.99 and for Class X node,  -(9/16) log2 (9/16) – (7/16) log2 (7/16) = 0.99.
5. Entropy for split Class =  (14/30)\*0.99 + (16/30)\*0.99 = **0.99**

Above, you can see that entropy for *Split on Gender* is the lowest among all, so the tree will split on *Gender*. We can derive information gain from entropy as **1- Entropy.**

**Reduction in Variance**

Till now, we have discussed the algorithms for categorical target variable. Reduction in variance is an algorithm used for continuous target variables (regression problems). This algorithm uses the standard formula of variance to choose the best split. The split with lower variance is selected as the criteria to split the population:

[](https://www.analyticsvidhya.com/blog/wp-content/uploads/2015/01/Varince.png)

Above X-bar is mean of the values, X is actual and n is number of values.

**Steps to calculate Variance:**

1. Calculate variance for each node.
2. Calculate variance for each split as weighted average of each node variance.

**Example:-** Let’s assign numerical value 1 for play cricket and 0 for not playing cricket. Now follow the steps to identify the right split:

1. Variance for Root node, here mean value is (15\*1 + 15\*0)/30 = 0.5 and we have 15 one and 15 zero. Now variance would be ((1-0.5)^2+(1-0.5)^2+….15 times+(0-0.5)^2+(0-0.5)^2+…15 times) / 30, this can be written as (15\*(1-0.5)^2+15\*(0-0.5)^2) / 30 = **0.25**
2. Mean of Female node =  (2\*1+8\*0)/10=0.2 and Variance = (2\*(1-0.2)^2+8\*(0-0.2)^2) / 10 = 0.16
3. Mean of Male Node = (13\*1+7\*0)/20=0.65 and Variance = (13\*(1-0.65)^2+7\*(0-0.65)^2) / 20 = 0.23
4. Variance for Split Gender = Weighted Variance of Sub-nodes = (10/30)\*0.16 + (20/30) \*0.23 = **0.21**
5. Mean of Class IX node =  (6\*1+8\*0)/14=0.43 and Variance = (6\*(1-0.43)^2+8\*(0-0.43)^2) / 14= 0.24
6. Mean of Class X node =  (9\*1+7\*0)/16=0.56 and Variance = (9\*(1-0.56)^2+7\*(0-0.56)^2) / 16 = 0.25
7. Variance for Split Gender = (14/30)\*0.24 + (16/30) \*0.25 = **0.25**

Above, you can see that Gender split has lower variance compare to parent node, so the split would take place on *Gender* variable.

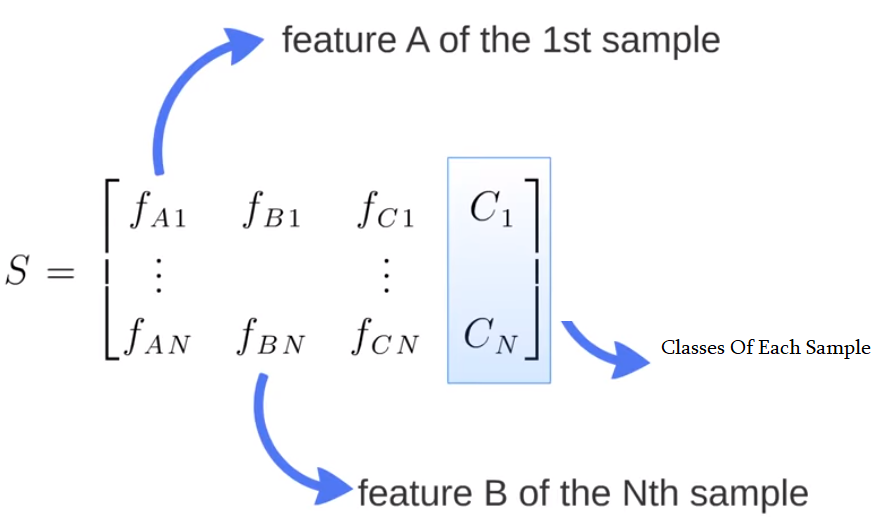
Until here, we learnt about the basics of decision trees and the decision making process involved to choose the best splits in building a tree model. As I said, decision tree can be applied both on regression and classification problems. Let’s understand these aspects in detail.

Random Forest

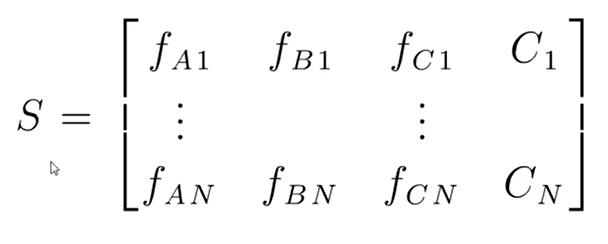
<https://www.youtube.com/watch?v=loNcrMjYh64>

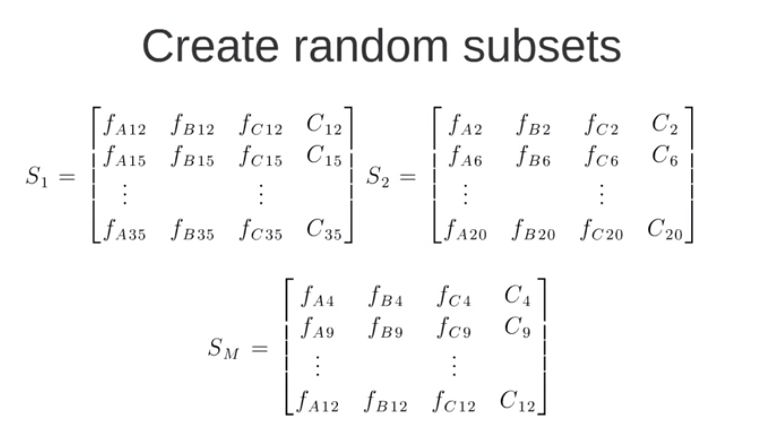
Random Forest

ვთქვათ მოცემული გვაქვს რაიმე ობიექტები რომელთაც გააჩნიათ მახასიათებლები (თვისებები) და რომელთათვისაც განსაზღვრულია კლასი რომელთაც ისინი განეკუთვნებიან:

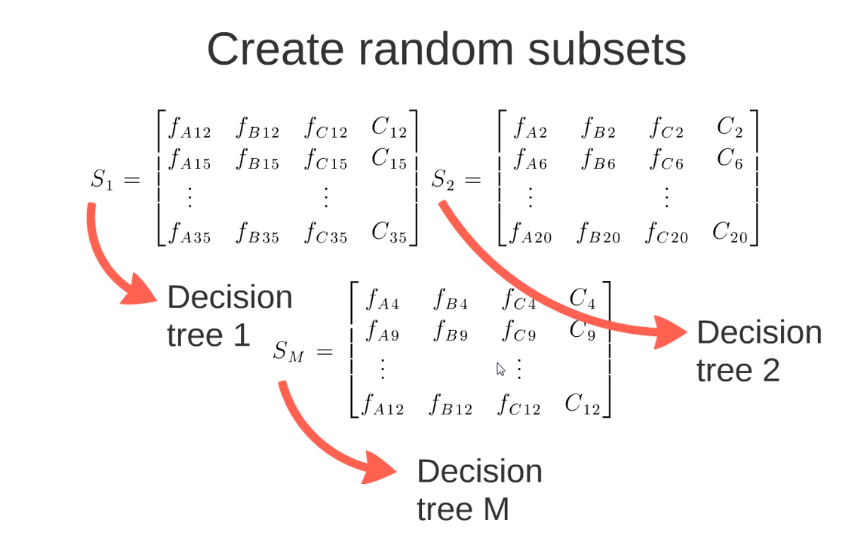


For The Following Matrix

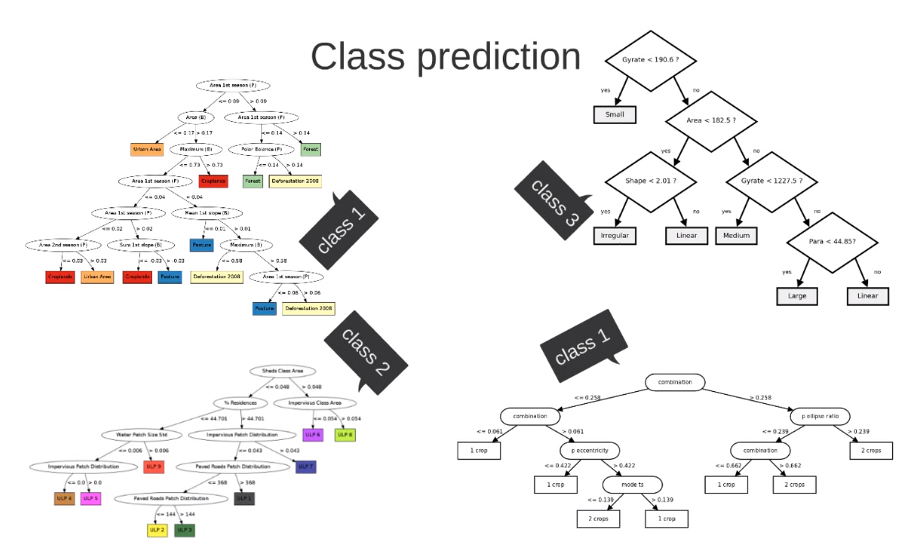




ამის შემდეგ ჩვენ თითოეული ზემოთ მოცემული მატრიცისათვის ავაგებთ გადაწყვეტილების ხეს.



ამის შემდგომ ვაგებთ გადაწყვეტილების ხეებს იმ მეთოდებით რომლებიც დასაწყისში აღვწერეთ



და შემთხვევითი ობიექტის კლასიფიკაციას ვახდენთ იმის მიხედვით თუ რომელი კლასითაც დაკლასიფიცირდება აღნიშნული ობიექტი ზემოთ მოყვანილი გადაწყვეტილების ხეები. ჩვენს მაგალითში გამოდის რომ ობიექტი დაკლასიფიცირდა Class 1-ით

**მაგალითი პითონზე:**

