**Decision Tree example**

In the below example response variable has only 2 classes whether to play tennis or not. But based on various conditions recorded on various days below table has been compiled. Now our task is to find out what are the variables resulting significantly the output YES or NO. This example comes under Classification tree

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Day** | **Outlook** | **Temperature** | **Humidity** | **Wind** | **Play Tennis** |
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

Taking **Humidity** example:

**CHAID:**  Humidity has 2 categories and our expected values should be evenly distributed in order to calculate how distinguishing the variable is

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Humidity Category** | Play Tennis | | Expected | | Difference | |
|  | No | Yes | No | Yes | No | Yes |
| High | 4 | 3 | 2.5 | 4.5 | **1.5** | **-1.5** |
| Normal | 1 | 6 | 2.5 | 4.5 | **-1.5** | **1.5** |
|  | **5** | **9** | **5** | **9** |  |  |

Calculating χ2 (Chi-square) value:

= 2.8

Calculating degrees of freedom = (r-1) \* (c-1)

Where r = number of row components/number of variable categories

C = number of response variables

Here row categories are 2 (High & Normal) and column categories are 2 (No & Yes)

Hence = (2-1) \* (2-1) = 1

P-value for Chi-square 2.8 with 1 d.f = 0.0942

P-value can be obtained with the following excel formulae

= CHIDIST (2.8, 1) = 0.0942

In similar way, we will calculate p-value for all variables and select the best variable with **small p-value**

**ENTROPY:**

Entropy = **- Σ P \* Log 2 P**

Normal

High

Humidity S [9 +, 5 -]

S [3+, 4-]

S [6+, 1-]

Entropy =

= 0.4097 + 0.5305 = 0.9402

In case if you forgot how to calculate = Log (9/14) / Log 2 = - 0.6374

Entropy High =

= 0.523 + 0.4613 = 0.9851

Entropy Normal =

= 0.4010 + 0.1906 = 0.5916

Information gain =

= 0.1518

In similar way, we will calculate **information gain** for all variables and select the best variable with **highest information gain**

**GINI:**

Gini = 1**- Σ p 2**

Normal

High

Humidity S [9 +, 5 -]

S [3+, 4-]

S [6+, 1-]

Gini = = 0.459

Gini High = = 0.489

Gini Normal = = 0.2448

Expected Gini = = 0.3669

In similar way, we will calculate **expected Gini** for all variables and select the best with **lowest expected value**

For better understanding purpose, we are doing the similar calculations for **WIND** variable

**CHAID:**  Wind has 2 categories and our expected values should be evenly distributed in order to calculate how distinguishing the variable is

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Wind Category** | Play Tennis | | Expected | | Difference | |
|  | No | Yes | No | Yes | No | Yes |
| Weak | 2 | 6 | 2.5 | 4.5 | **-0.5** | **1.5** |
| Strong | 3 | 3 | 2.5 | 4.5 | **0.5** | **-1.5** |
|  | **5** | **9** | **5** | **9** |  |  |

= 1.2

P-value = 0.2733

**ENTROPY:**

Strong

Weak

Wind S [9 +, 5 -]

S [6+, 2-]

S [3+, 3-]

Entropy Weak =

= 0.3112 + 0.5 = 0.8112

Entropy Strong =

= 0.5 + 0.5 = 1

Information gain =

= 0.0482

**GINI:**

Gini Weak = = 0.375

Gini Strong = = 0.5

Expected Gini = = 0.4285

Now we are comparing both variables for all 3 metrics so that we can understand better

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | CHAID   (P-value) | Entropy Information Gain | Gini Expected value |
| Humidity | 0.0942 | 0.1518 | 0.3669 |
| Wind | 0.2733 | 0.0482 | 0.4285 |
| Better | Low value | High value | Low Value |

For all 3 calculations, **Humidity** is better classifier than **Wind.** Hence we can confirm that **a**ll methods are telling similar story