G-Statistics Method in Decision Tree

Instructor : Chan Sophal

Group 03

Our Team

1. Khon Yin Sakal	e20200425
2. Sao Samarth	e20200084
3. Chou Vandy	e20200664
4. Hong Kimleng	e20200766
5. Hok Kimleang	e20200637
6. Chorn Seyhak	e20201099
7. Hok Ratanak	e20201106
8. Rith Chanthyda	e20200612
9. Vann Visal	e20200537
10. Oun Vikreth	e20200485
11. Kry Senghort	e20200706

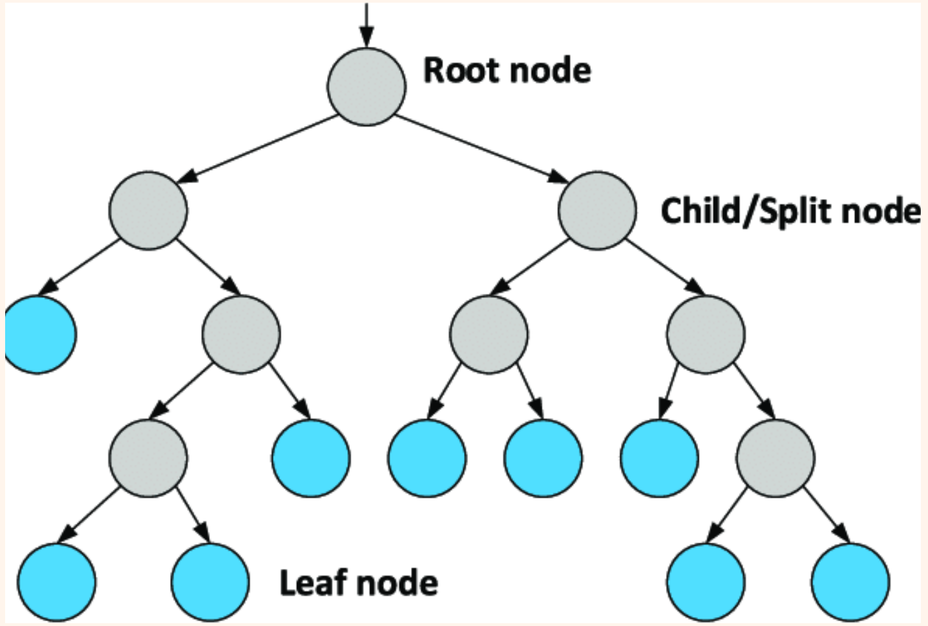
Table of Content

Introduction

Decision trees are often used while implementing machine learning algorithms. The hierarchical structure of a decision tree leads us to the final outcome by traversing through the nodes of the tree. Each node consists of an attribute or feature which is further split into more nodes as we move down the tree. But how do we decide:

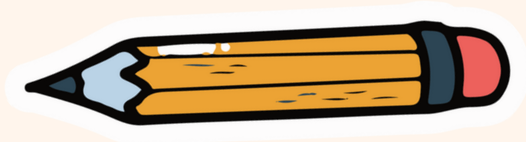
- Which attribute/feature should be placed at the root node?
- Which features will act as internal nodes or leaf nodes?

To decide this, and how to split the tree, we use splitting measures like Gini Index, Information Gain, etc. In today's topic, we will learn all about the Gini Index, including the use of the Gini Index to split attribute in a decision tree.



What is Gini Index ?

In a decision tree model, G-statistics (also known as Gini impurity or Gini index) is a method used in decision tree algorithms to assess the impurity of splits and guide the tree construction process. It helps identify the optimal features and split points that result in more homogeneous child nodes.



Characteristics

1. The G-statistics, or Gini index, measures the impurity of a node in a decision tree

But what is actually meant by ‘impurity’?

The degree of the Gini Index varies between 0 and 1 where:

- '0' denotes that all elements belong to a certain class or there exists only one class (pure).
- '1' denotes that the elements are randomly distributed across various classes (impure).
- A Gini Index of '0.5 'denotes equally distributed elements into some classes.

A lower Gini_Index value indicates higher purity and easier classification, while a higher GIni-Index value suggests lower purity and more difficulty in classification.

The formula for Gini Index

$$Gini = 1 - \sum_{i=1}^{j} P(i)^2$$

Where j represents the no. of classes in the target variable — Pass and Fail in our example

P(i) represents the ratio of Pass/Total no. of observations in node.

How does it work?

Steps involved in using the Gini index as an attribute selection method in decision trees:

Step 1: The Gini index is calculated for each potential split point across all features in the dataset

Step 2: The feature and split point with the lowest Gini index are selected as the best-split point for the decision tree.

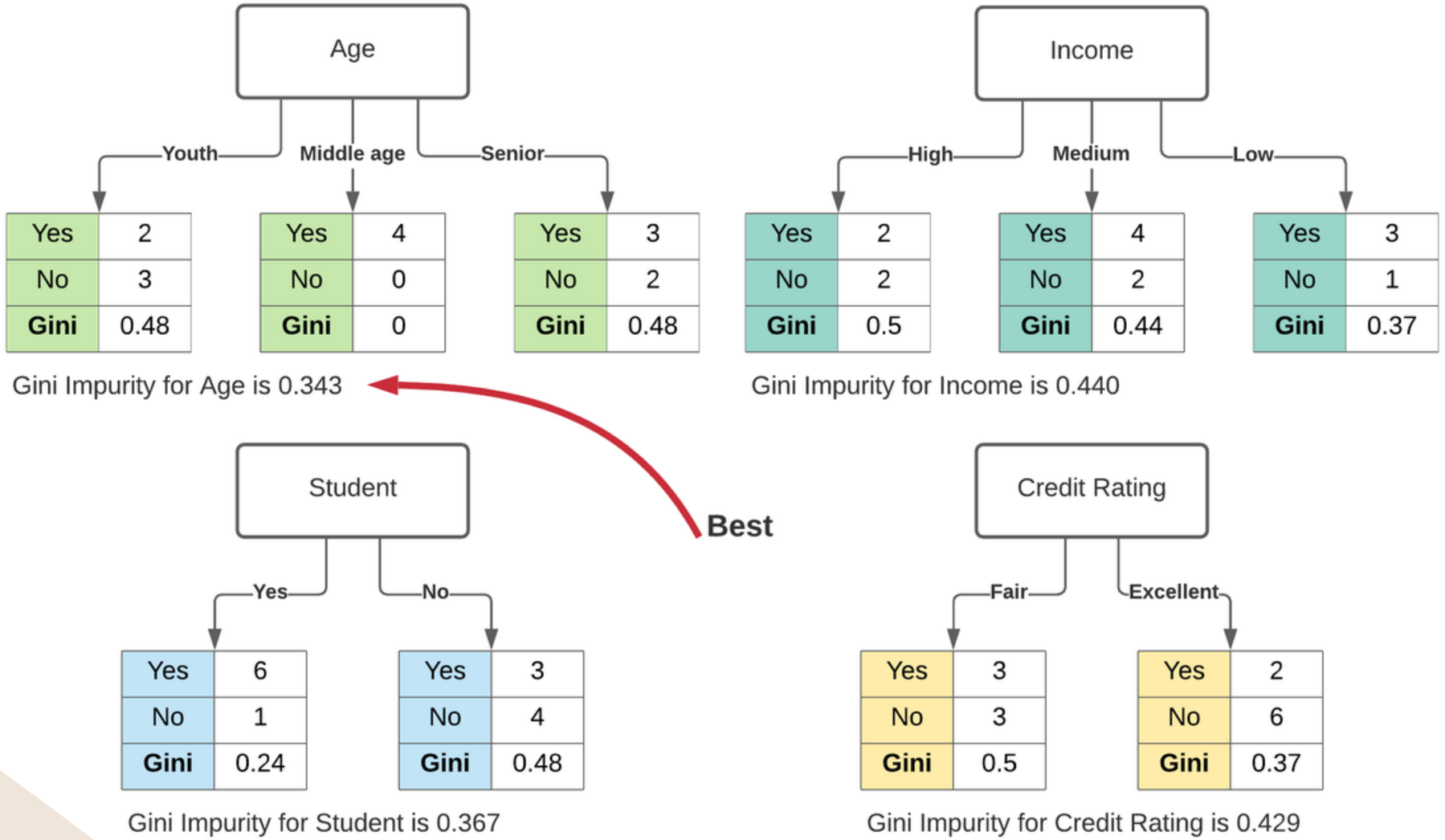
Step 3: This selected split point divides the data into two subsets, maximizing the homogeneity of the resulting child nodes based on the Gini index.

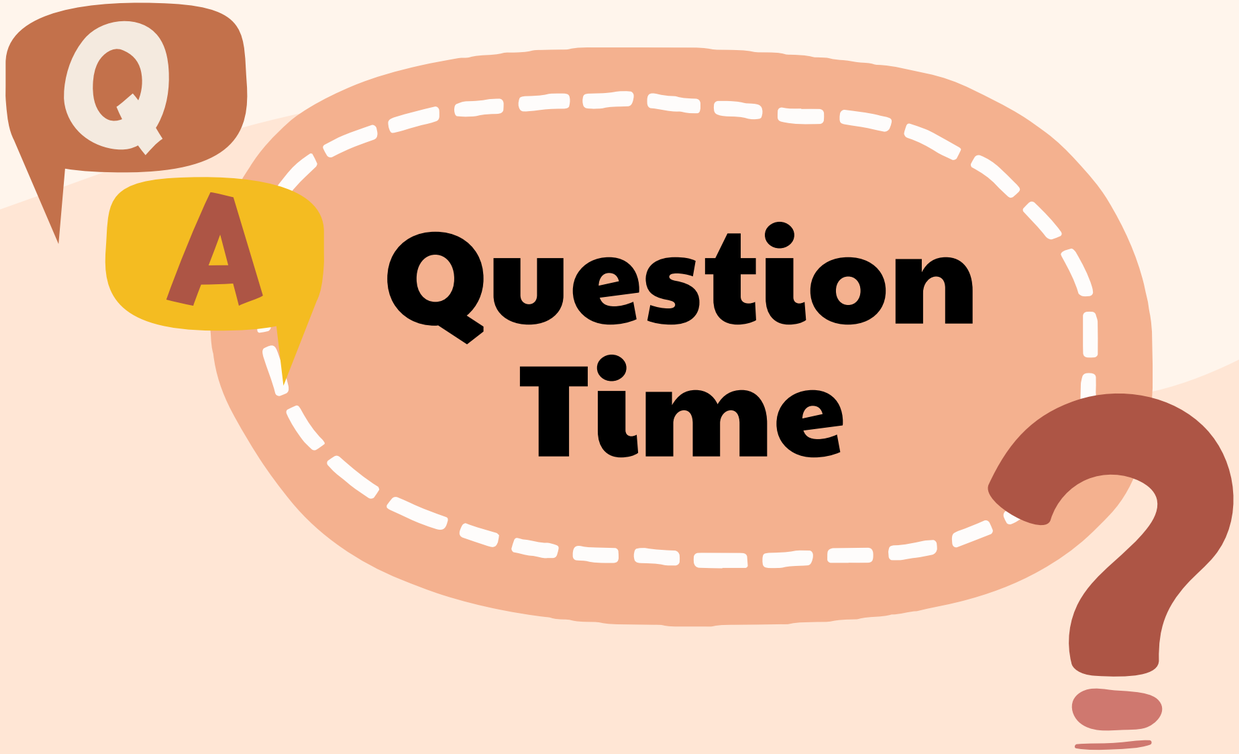
Step 4: The decision tree construction process continues recursively, repeating these steps for each child node until a stopping criterion is met or the tree is fully grown.

Example:

	age	income	student	credit_rate	default
0	youth	high	no	fair	no
1	youth	high	no	excellent	no
2	middle_age	high	no	fair	yes
3	senior	medium	no	fair	yes
4	senior	low	yes	fair	yes
5	senior	low	yes	excellent	no
6	middle_age	low	yes	excellent	yes
7	youth	medium	no	fair	no
8	youth	low	yes	fair	yes
9	senior	medium	yes	fair	yes
10	youth	medium	yes	excellent	yes
11	middle_age	medium	no	excellent	yes
12	middle_age	high	yes	fair	yes
13	senior	medium	no	excellent	no

Example:


Age
Youth
Middle age
Senior
Yes 2
No 3
Gini 0.48
Yes 4
No 0
Gini 0
Yes 3
No 2
Gini 0.48
Gini Impurity for Age is 0.343
Income
High
Medium
Low
Yes 2
No 2
Gini 0.5
Yes 4
No 2
Gini 0.44
Yes 3
No 1
Gini 0.37
Gini Impurity for Income is 0.440
Best
Student
Yes
No
Yes 6
No 1
Gini 0.24
Yes 3
No 4
Gini 0.48
Gini Impurity for Student is 0.367
Credit Rating
Fair
Excellent
Yes 3
No 3
Gini 0.5
Yes 2
No 6
Gini 0.37
Gini Impurity for Credit Rating is 0.429



Question Time

THANK YOU SO MUCH!