Decision Trees - Decision Trees - 1

One should look for what is and not what he thinks should be. (Albert Einstein)

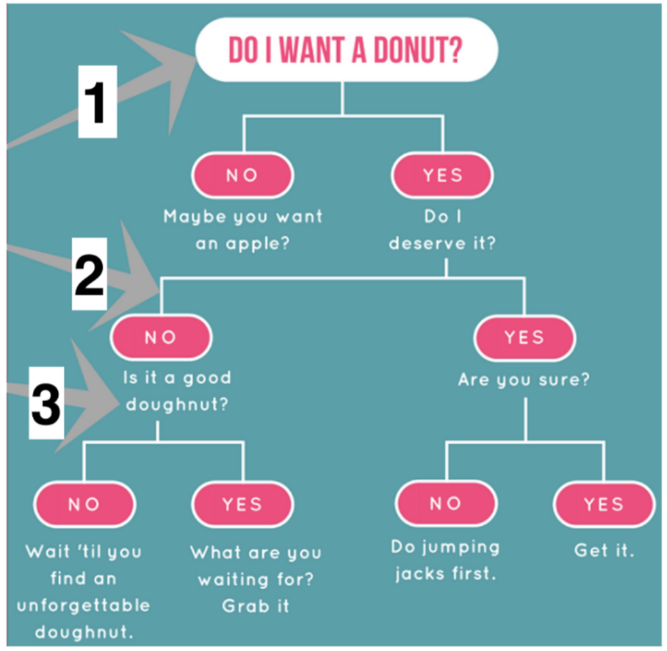
Decision Trees: Topic introduction

In this part of the course, we will cover the following concepts:

- Decision Trees use cases and the theory behind them
- Data transformation necessary for decision trees
- Implementation of decision trees on a dataset
- Model performance evaluation and tuning

Decision Tree terms

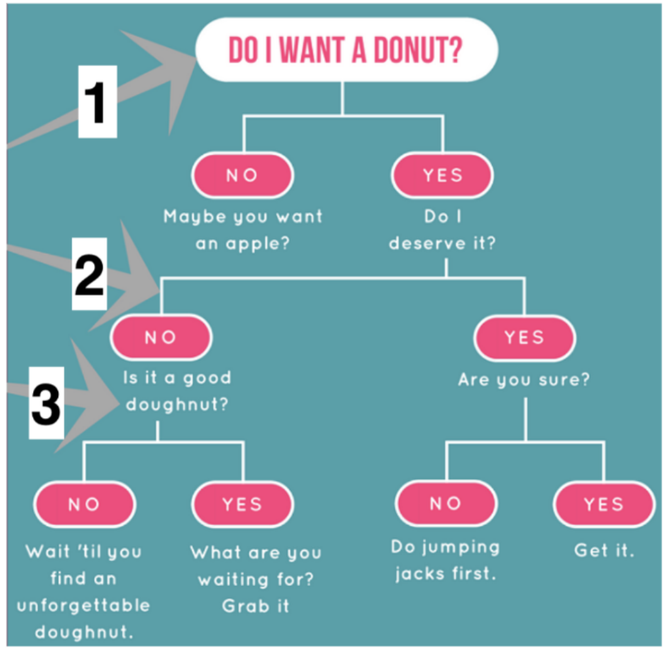
- This course focuses on a kind of advanced classification method called a **decision tree**
- Every decision tree starts with a specific decision called the **root node**
- The root and **leaf nodes** hold questions you have to answer
- **Branches** are lines that connect the nodes



Source

Decision Tree terms

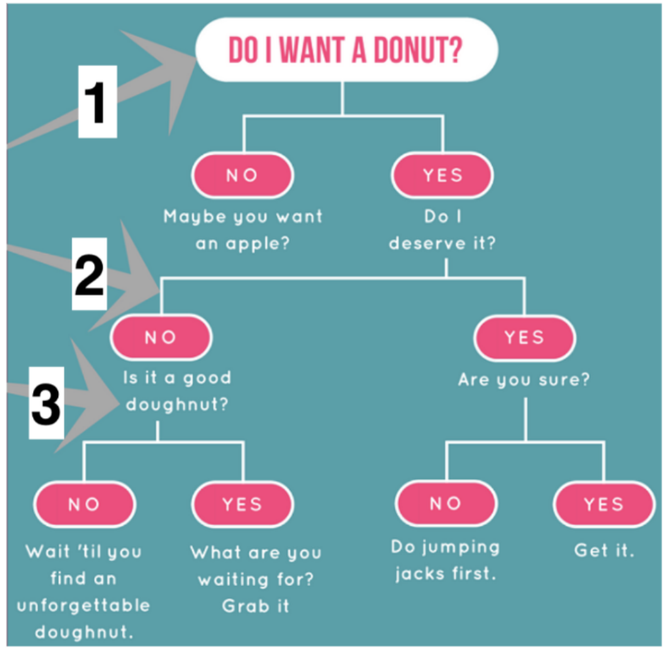
- Which number on the diagram corresponds to the **leaf** node?
- Share your response in the chat box



Source

Decision Tree terms

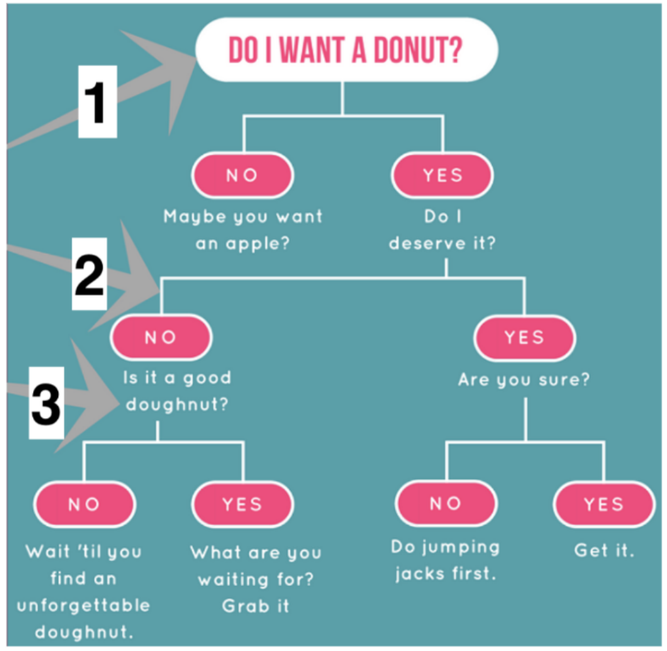
- How about the **root** node?
- Share your response in the chat box



Source

Decision Tree terms

- Finally, what is the **name** of **number 2**?
- Share your response in the chat box



Source

Decision Trees: what are they?

- Decision Trees are one of the supervised machine learning models used to perform Classification
  - They have a **tree-like** structure with a root node, branches, and a set of leaf nodes
  - They are easy to **walk through and explain to stakeholders who are not familiar with data science**
  - They are **intuitive** and popular because they **provide explicit rules for Classification**
  - They **cope well with heterogeneous data, missing data, and nonlinear effects**
  - They **predict** the target value of an item by **mapping observations about the item**

Classification: general use cases

- These are some examples of how you would apply classification algorithms in a health setting

Question	Example
What is this object like?	Selecting similar medicines with similar purposes
Who is this person like?	Anticipating behavior or preferences of a person based on her similarities with others
What category is this in?	Anticipating if your patient is high risk, has an illness, will develop symptoms, etc.
What is the probability that something is in a given category?	Determining the probability that a drug is in a particular category; determining the probability that someone will contract an illness

Module completion checklist

Objectives	Complete
Discuss use cases for Decision Trees	
Summarize the concepts and math behind Decision Trees	

Decision Trees: pros and cons

- Decision Trees are **great** when used for:
  - Classification and Regression
  - Handling numerical and categorical data
  - Handling data with missing values
  - Handling data with nonlinear relationships between parameters

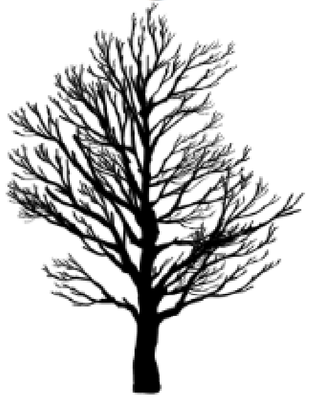
- Decision Trees are **not very good** at:
  - **Generalization**: they are known for overfitting
  - **Robustness**: small variations in data can result in a different tree
  - **Mitigating bias**: if some classes dominate, trees may be unbalanced and biased

Module completion checklist

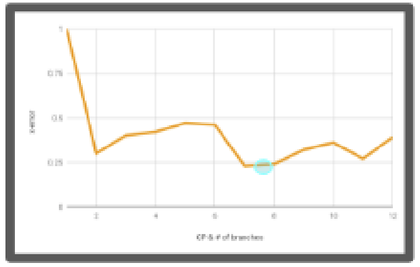
	Complete
Discuss use cases for Decision Trees	✔
Summarize the concepts and math behind Decision Trees	

Decision Trees: process

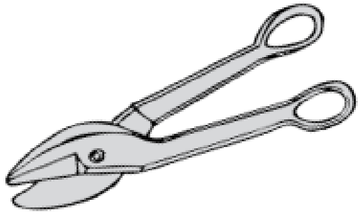
Step 1:
Grow tree on training data



Step 2:
Examine Model output



Step 3:
Prune Tree



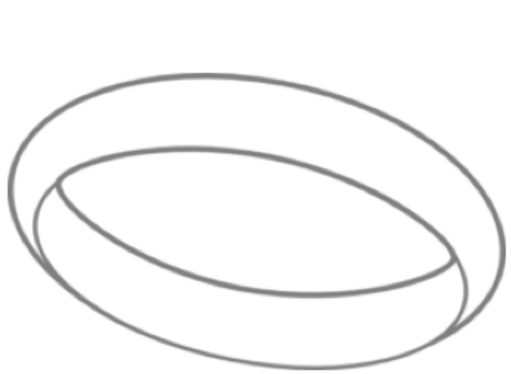
Step 4:
Check performance on test data

	Act +	Act -	
Pred +			
Pred -			

Growing Decision Trees: find the most important question

Which question is more important on a date?

The most relevant question

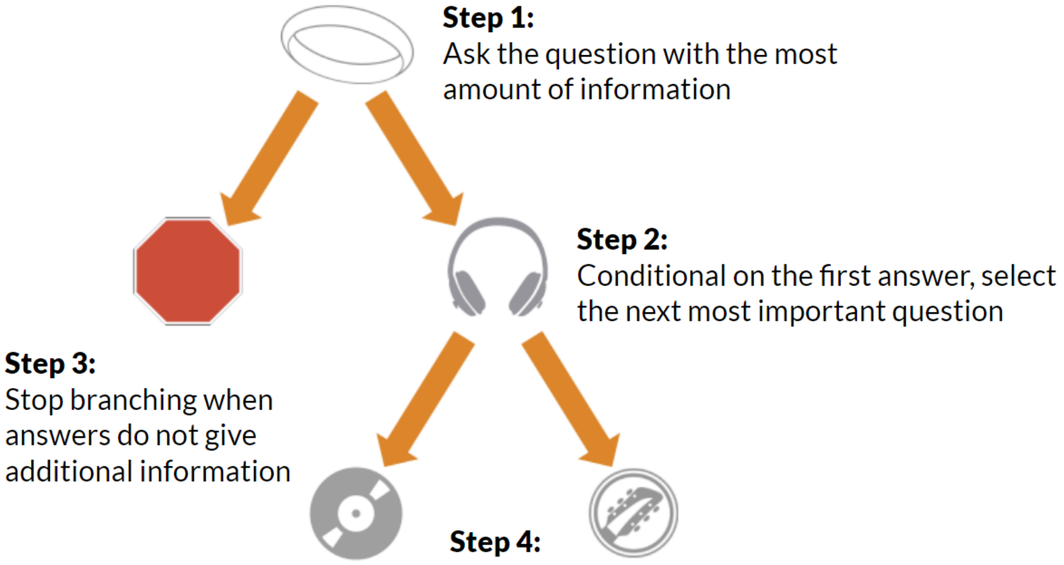


Are you married?

>

What music do you like?

Growing Decision Trees: four steps



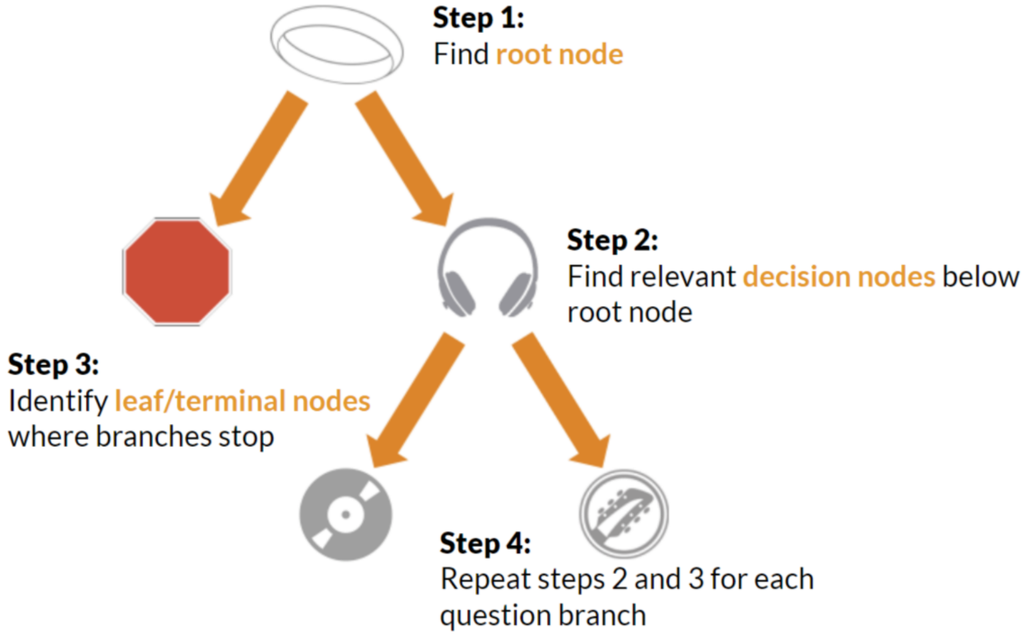
Step 1:
Ask the question with the most amount of information

Step 2:
Conditional on the first answer, select the next most important question

Step 3:
Stop branching when answers do not give additional information

Step 4:
Repeat steps 2 and 3 for each question branch

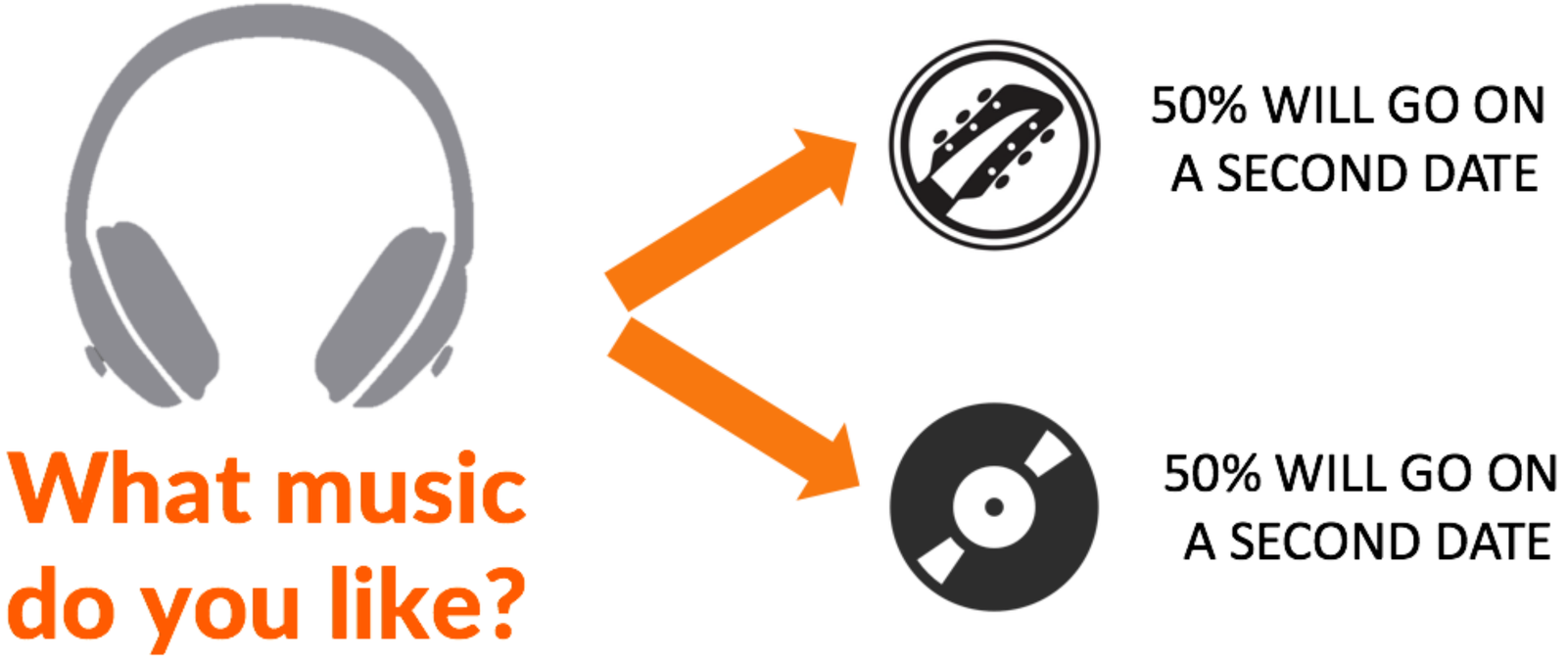
Growing decision trees: steps with vocabulary



Growing Decision Trees: when to stop

When should you stop asking questions?

When the answer no longer provides additional relevant information



Growing Decision Trees: the math of the splits

How do we decide which node to split and how to split it?

- There are two **impurity** functions that are most commonly used with tree-based models
  - Gini
  - Entropy
- The **`sklearn.tree` algorithm** uses Gini, so this is the method we will focus on in this module

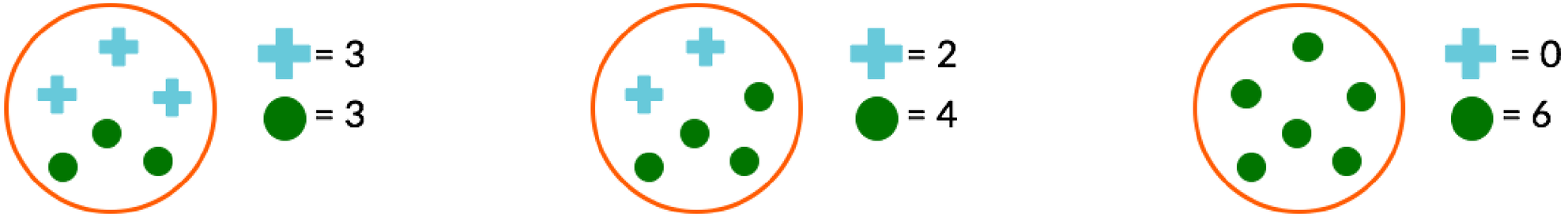
Growing Decision Trees: Gini

Gini measures the **probability of misclassification** in the model for each branch of a decision tree. Gini ranges from 0 to 1.

$$Gini(E) = 1 - \sum_{j=1}^{c} p_j^2$$

P*i* = the probability that a random selection would have state *i*

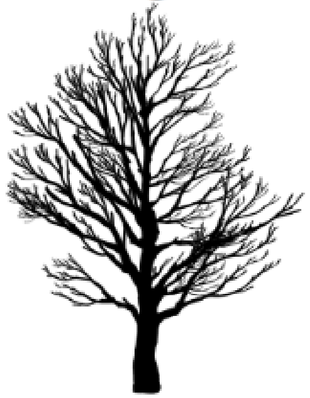
$$\text{Gini impurity} = 1 - \text{sum}[\ (P_i)^2\ ]$$



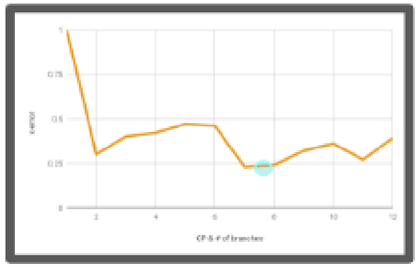
$1-(3/6)^2-(3/6)^2$ $\quad$ $1-(4/6)^2-(2/6)^2$ $\quad$ $1-(6/6)^2-(0/6)^2$

Decision Trees: process

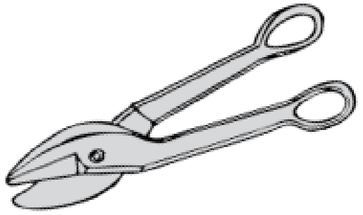
Step 1: Grow tree on training data



Step 2: Examine Model output



Step 3: Prune Tree



Step 4: Check performance on test data

	Act +	Act -	
Pred +			
Pred -			

SO far, we learned about growing the tree and making the choice on how to proceed regarding step 2. In order to get to step 3 and 4, we will need to work with a dataset next.

Knowledge check

Module completion checklist

Objectives	Complete
Discuss use cases for Decision Trees	✔
Summarize the concepts and math behind Decision Trees	✔

Congratulations on completing this module!