the magic of classification trees

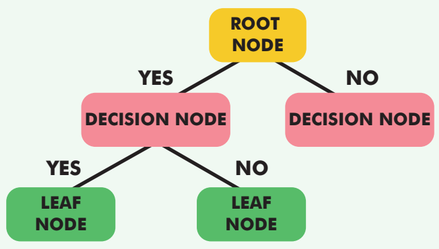
How can machines make accurate predictions? Classification trees can be used to predict possible outcomes to a decision based on observations of a certain item's features.
For example, a classification tree could be used to predict the drinkability of new samples from a water source by comparing certain qualities (e.g., color, pH, and hardness) against a database of previous samples.

Join Wilfred the wizard while he prepares a potion. To finish it, he needs more magical mushrooms. By analyzing the individual features of mushrooms in his vast catalogue of specimens, Wilfred wants to create a classification tree to identify if new mushrooms are magical.

training phase

1 first things first

The goal of a classification tree is to develop a model to predict the category (in this case: magic or non-magic), of an element (mushroom) based on its features or variables (cap color, stem color, and their height) by learning rules inferred by previous data.



The classification starts with a root node and then the algorithm will divide the mushrooms into smaller groups by asking a series of yes or no questions on their features.

2 shroomy dataset

To build the classification tree, we use the following data from the wizard's catalogue.

While most of the data will be used to build the tree, **a portion are kept aside** to be used to check the accuracy of the mushroom magic analyses during the testing phase later on.

cap colour	stem colour	stem height	is it magic?
pink	pink	tall	no
pink	green	tall	yes
green	pink	tall	no
yellow	green	short	yes
pink	green	short	yes
pink	pink	short	no
pink	yellow	short	no
yellow	green	tall	yes
yellow	pink	tall	yes
pink	yellow	tall	no
green	yellow	tall	yes
green	green	tall	yes
yellow	pink	short	yes
yellow	yellow	tall	no

3 make it binary

The model can't understand categorical data (e.g., pink, green, or yellow stem colour) so it must be transformed into **binary information**, by asking yes or no questions.

This way we can prepare Wilfred's data from his observations on the features of his mushrooms for use.

For example, is the cap pink?

cap colour	binary answer
pink	yes
green	no
yellow	no
pink	yes

4 how to split the data?

How does the algorithm decide where to divide the mushrooms? **The algorithm tries to create groups that are as pure as possible** - ideally groups of exclusively magical and non-magical mushrooms.

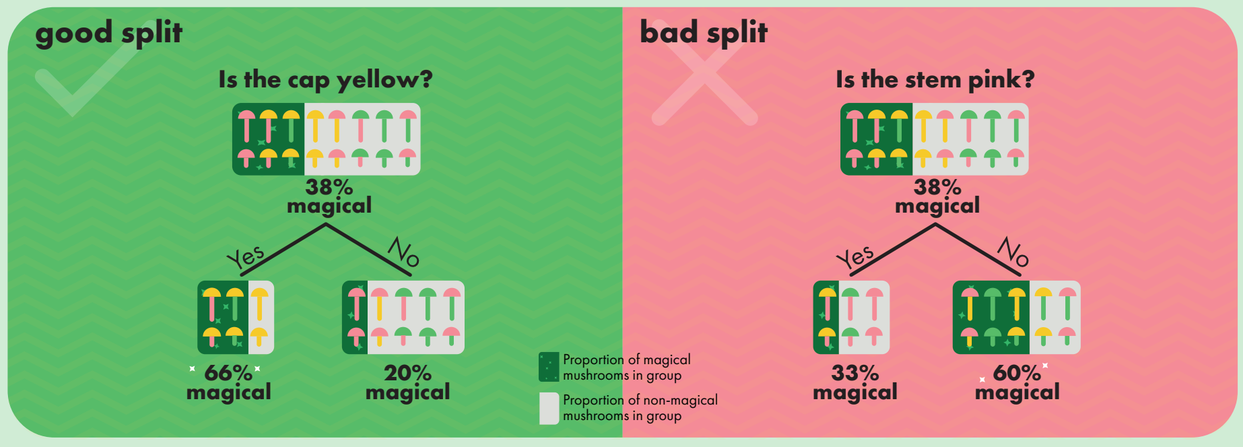
A good split would put all the magical mushrooms in one node and all non-magical mushrooms in another.

A bad split would divide the mushrooms but keep the same ratio of magical and non-magical mushrooms in each group.

By using this logic, the whole tree is created.

Of these two questions, this one gives a better divide.

This is because there are almost exclusively magical or non-magical mushrooms on either side, meaning that a yellow cap is a good indicator of a magical mushroom.



testing phase

5 checking the accuracy

To ensure its accuracy, we must test our tree. To do this, we use the testing mushrooms we put aside at the very start. We want to see how successfully the tree sorts these known mushrooms.

Ideally, the testing mushrooms would give a similar accuracy to the training mushrooms:

100%
TRAINING ACCURACY
Accuracy of predicting initial sample that built tree

95%
TESTING ACCURACY
Accuracy of predicting testing sample of known magicness

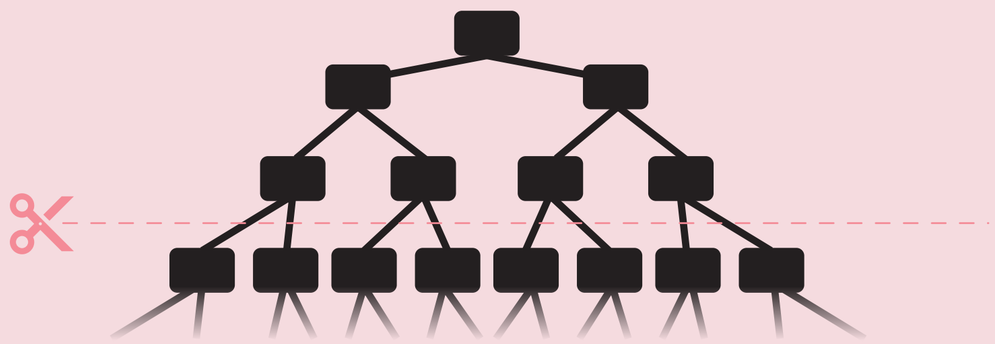
If the testing accuracy is significantly lower than the training accuracy, it means it is **overfit**. This is a problem because it means that the tree is too much adapted to the initial database and won't predict well when the wizard wants to identify a new mushroom.

100%
TRAINING ACCURACY
Accuracy of predicting initial sample that built tree

60%
TESTING ACCURACY
Accuracy of predicting testing sample of known magicness

6 pruning

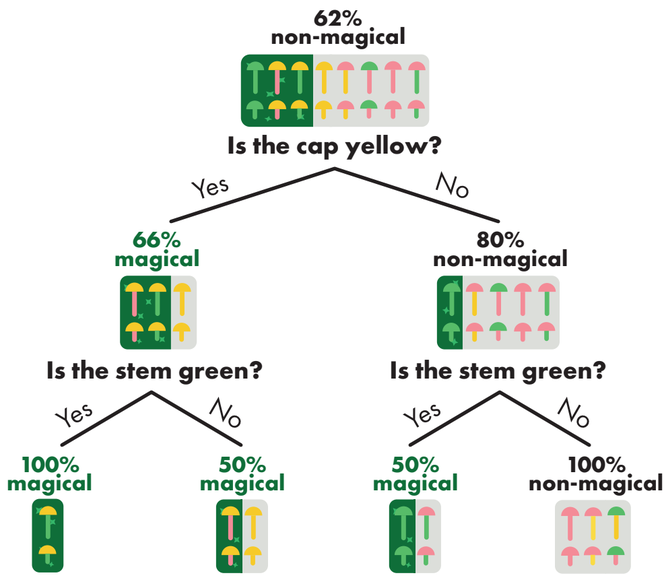
If the tree is overfit, we can fix this by pruning it. That means that we cut the questions that do not help the tree classify information, and helps reduce its size and complexity.



the result

The tree created using the sample data

Proportion of magical mushrooms in group
Proportion of non-magical mushrooms in group

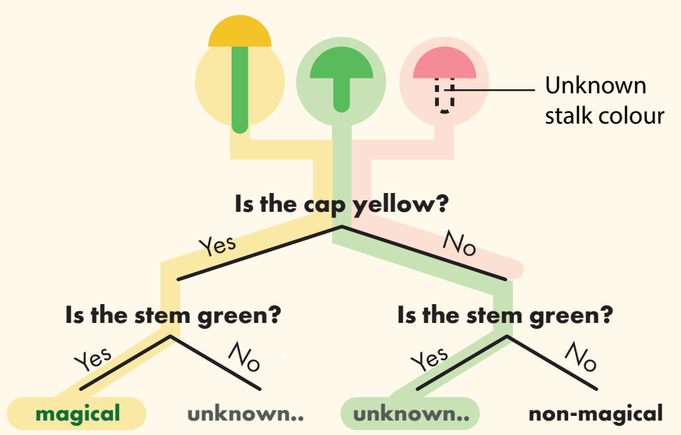


prediction phase

we've found some new mushrooms!

Once the pruning is done, the algorithm is more accurate and is ready to be applied to new unknown mushrooms.

Wilfred wants to discover if his new discoveries are magical or not.



Most likely magical! As the original group had 100% magical mushrooms.

Impossible to tell if magical or not, as the original group only had 50% magical mushrooms.

Most likely non-magical. Classification trees can be used with incomplete data, as the sample can follow the tree for as long as it can. In the original tree, this group had 80% non-magical.

As our original sample was very small, the reliability of this tree would be low. A successful tree needs many more samples.

VISUAL EXPLANATIONS OF STATISTICAL MODELS

Classification Trees

AUTHORS

Marina Fernández
Andrea Llamas
Amanda Cestaro
Leo Gamberini
Regina Salviato
Qi Yu
Renata Martínez

FACULTY

Michele Mauri
Ángeles Briones
Gabriele Colombo
Simone Vantini
Salvatore Zingale

TEACHING ASSISTANTS

Andrea Benedetti
Tommaso Elli
Beatrice Gobbo

Final Synthesis Design Studio
Sect. C3

LM in Communication Design
A.A. 2021/2022

DEN-SITY GN+

POLITECNICO MILANO 1863