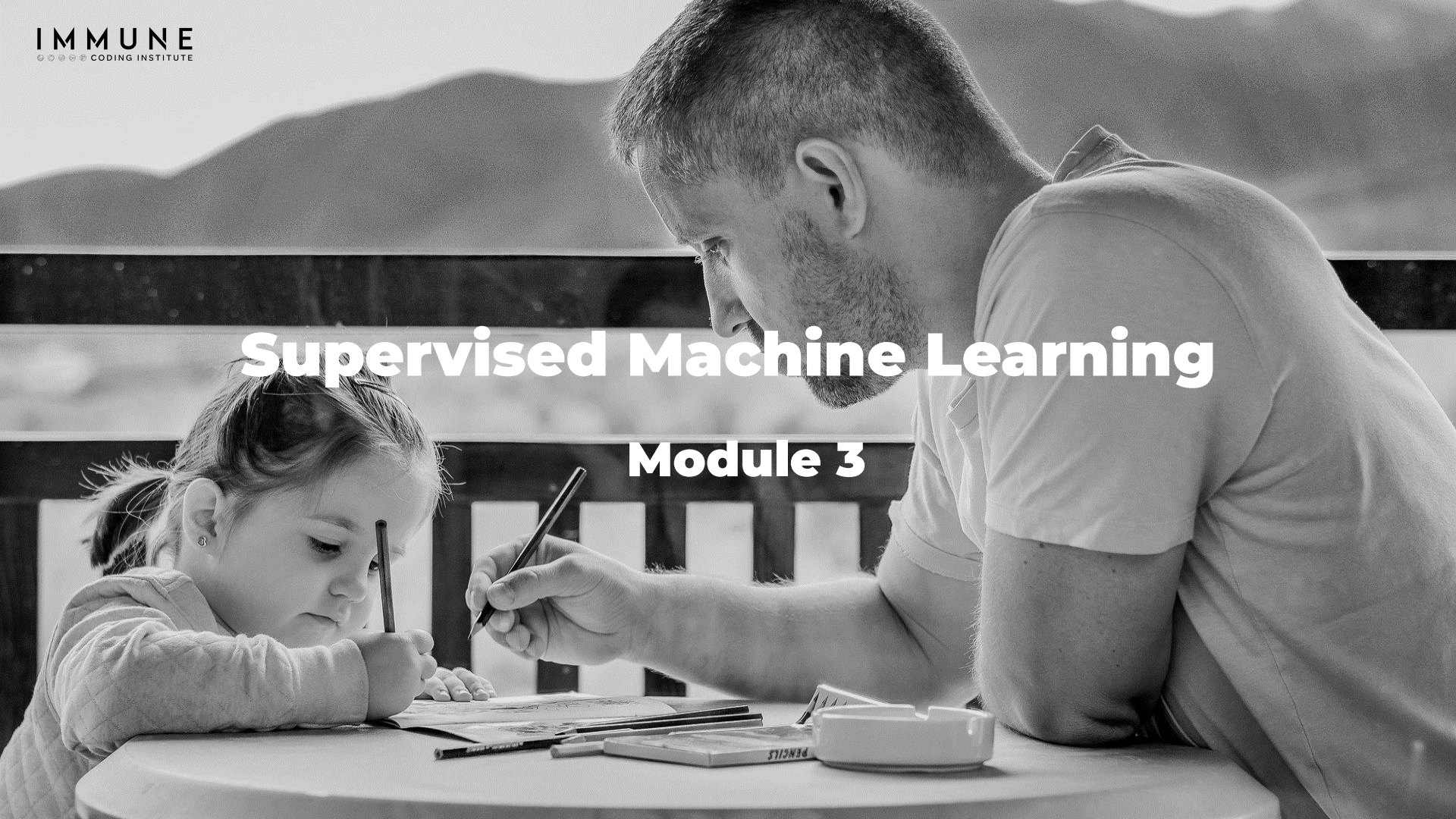
IMMUNE
CODING INSTITUTE



Supervised Machine Learning

Module 3

IMMUNE
CODING INSTITUTE

Module 3 Summary

SESSION	TITLE	TEACHER
1	ML Foundations	Juan
2	Regression Introduction and Practice	Juan
3	Classification Introduction and Practice	Carlos
4	Feature Engineering and Selection for ML	Carlos
5	**Advanced Supervised Models 1**	**Carlos**
6	Advanced Supervised Models 2	Carlos
7	Hands-on Practice	Carlos

IMMUNE
CODING INSTITUTE

Introduction to Tree Methods

The Wanaka Tree, New Zealand

Outline

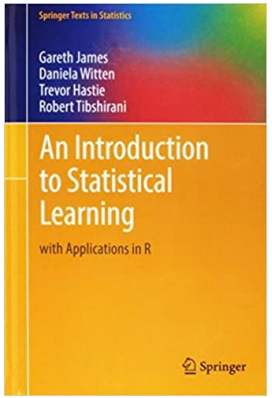
- Basic of Decision Trees
- Classification Trees
- Regression Trees
- Tree Pruning
- Bagging
- Random Forest
- Boosting

Classification and Regression Trees (CART)

CART, commonly known as Decision trees, can be applied to both regression and classification problems and can be represented as binary trees.

One of the most important aspects of CART is its interpretability.

- Chapter 8 of **Introduction to Statistical Learning** (Gareth James, et al.)



Let's start with a practical example

Let's say that every Sunday we meet some friends to play Padel.

Sometimes Rubén shows up and sometimes he doesn't.

For him it depends on a variety of factors, such as: temperature, humidity, wind, weather, family, etc.

So, we start keeping track of these events and its features to understand the pattern.

Our Data

Temp	Humidity	Outlook	Football	Rubén?
Mild	80	Sunny	No	Yes
Hot	75	Sunny	Yes	**No**
Hot	77	Overcast	No	Yes
Cool	70	Rain	No	Yes
Cool	72	Overcast	Yes	Yes
Mild	77	Sunny	No	**No**
Cool	70	Sunny	No	Yes
Mild	69	Rain	No	Yes
Mild	65	Sunny	Yes	Yes
Mild	77	Overcast	Yes	Yes
Hot	74	Overcast	No	Yes
MIld	77	Rain	Yes	**No**
Cool	73	Rain	Yes	**No**
Mild	78	Rain	No	Yes

Our Data

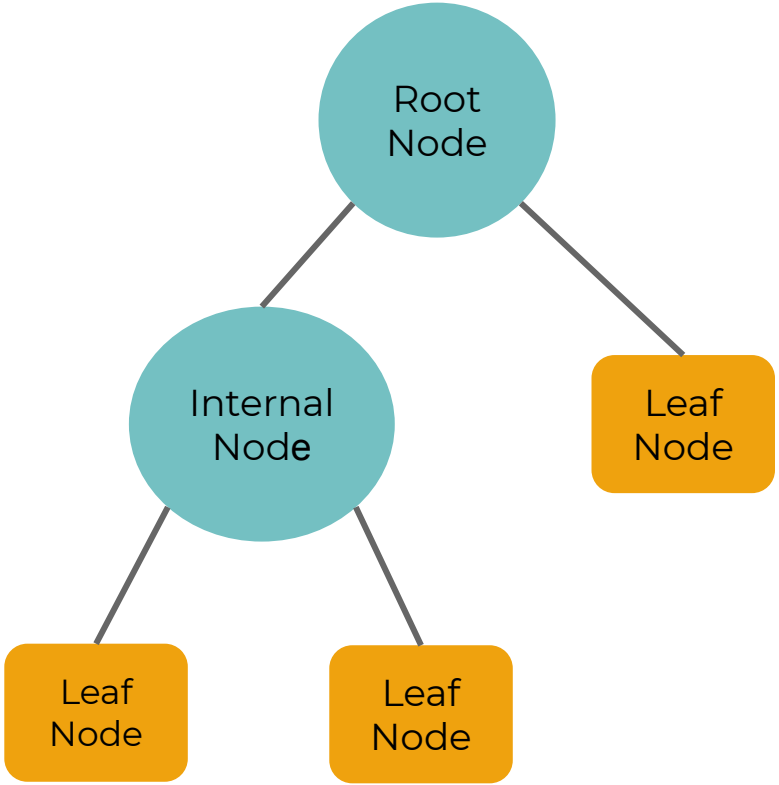
Temp	Humidity	Outlook	Football	Rubén?
Hot	70	Sunny	No	**??**

Temp	Humidity	Outlook	Football	Rubén?
Mild	80	Sunny	No	Yes
Hot	75	Sunny	Yes	**No**
Hot	77	Overcast	No	Yes
Cool	70	Rain	No	Yes
Cool	72	Overcast	Yes	Yes
Mild	77	Sunny	No	**No**
Cool	70	Sunny	No	Yes
Mild	69	Rain	No	Yes
Mild	65	Sunny	Yes	Yes
Mild	77	Overcast	Yes	Yes
Hot	74	Overcast	No	Yes
MIld	77	Rain	Yes	**No**
Cool	73	Rain	Yes	**No**
Mild	78	Rain	No	Yes

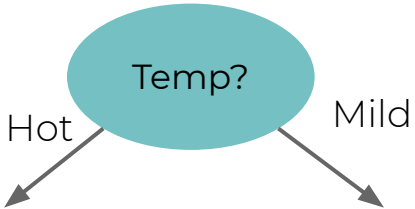
In a tree we have:

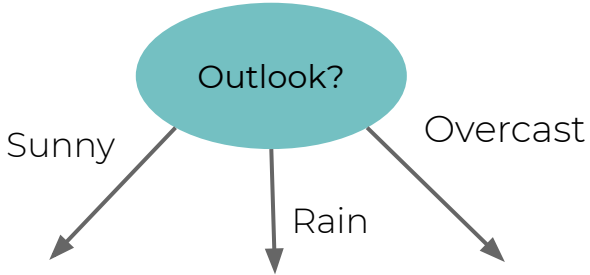
Root:
- The node that performs the first split

Leaves:
- Terminal nodes that predict the outcome



Which feature should be the Root Node?





Suppose a Data with 3 features (X, Y, and Z) with two possible classes:

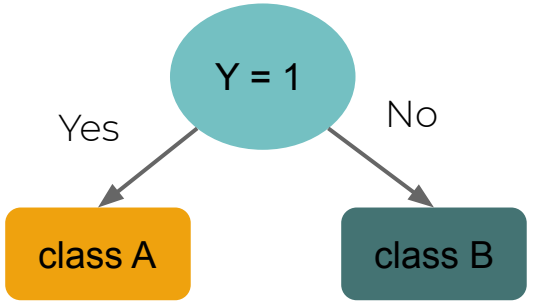
X	Y	Z	Class
1	1	1	A
1	1	0	A
0	0	1	B
1	0	0	B

Suppose a Data with 3 features (X, Y, and Z) with two possible classes:

X	Y	Z	Class
1	1	1	A
1	1	0	A
0	0	1	B
1	0	0	B

Suppose a Data with 3 features (X, Y, and Z) with two possible classes:

X	Y	Z	Class
1	1	1	A
1	1	0	A
0	0	1	B
1	0	0	B

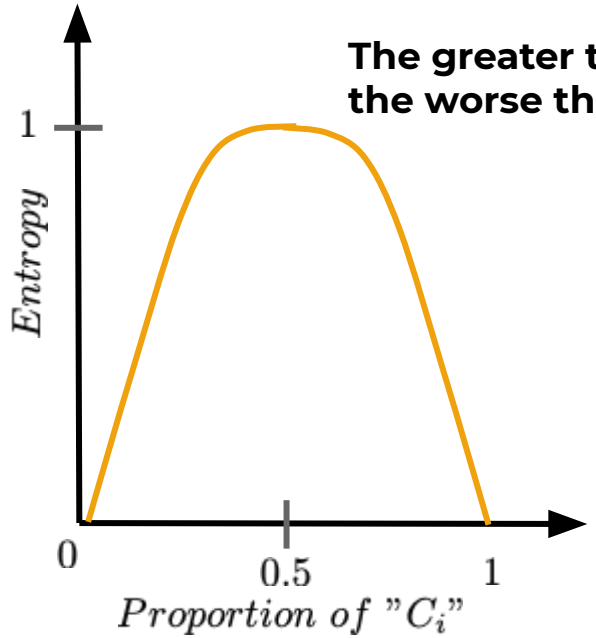


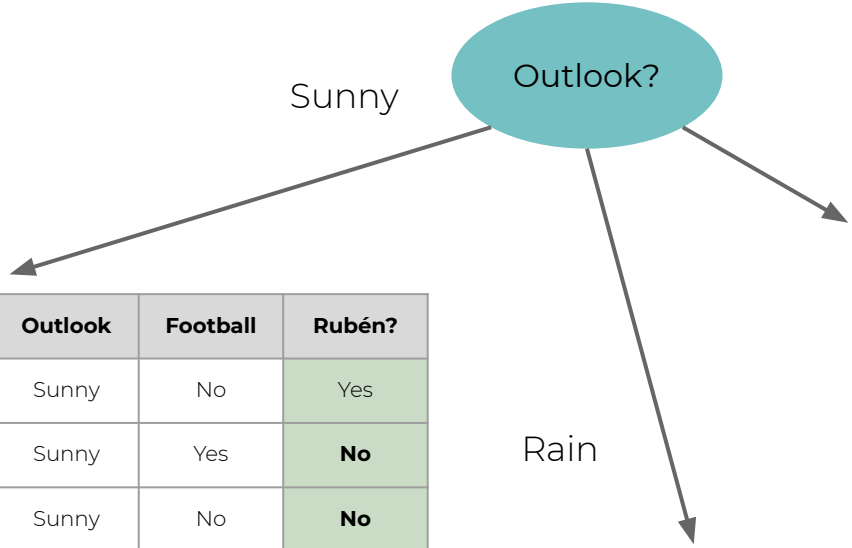
How do we know which variable to split by?

Entropy and **Information Gain** are the Mathematical Methods of choosing the best split.

$$H(P) = -\sum_{Ci} p_{C_i} log_2(p_{C_i})$$

We will use a measure that gets the best value when the attribute gives me partitions that are as homogeneous as possible, on average





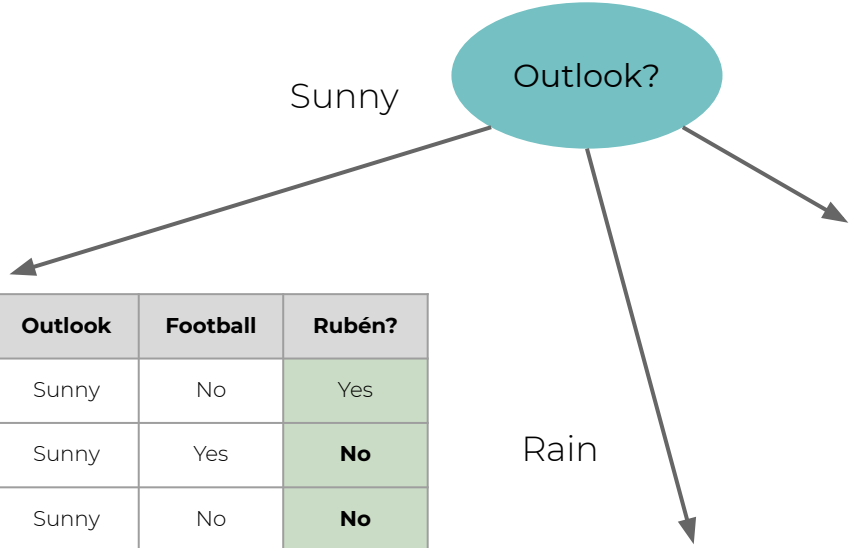
Temp	Humidity	Outlook	Football	Rubén?
Mild	80	Sunny	No	Yes
Hot	75	Sunny	Yes	**No**
Mild	77	Sunny	No	**No**
Cool	70	Sunny	No	Yes
Mild	65	Sunny	Yes	Yes

Temp	Humidity	Outlook	Football	Rubén?
Hot	77	Overcast	No	Yes
Cool	72	Overcast	Yes	Yes
Mild	77	Overcast	Yes	Yes
Hot	74	Overcast	No	Yes

Temp	Humidity	Outlook	Football	Rubén?
Cool	70	Rain	No	Yes
Mild	69	Rain	No	Yes
MIld	77	Rain	Yes	**No**
Cool	73	Rain	Yes	**No**
Mild	78	Rain	No	Yes

IMMUNE CODING INSTITUTE

$$H(P) = -\sum_{Ci} p_{C_i} log_2(p_{C_i})$$



Temp	Humidity	Outlook	Football	Rubén?
Mild	80	Sunny	No	Yes
Hot	75	Sunny	Yes	**No**
Mild	77	Sunny	No	**No**
Cool	70	Sunny	No	Yes
Mild	65	Sunny	Yes	Yes

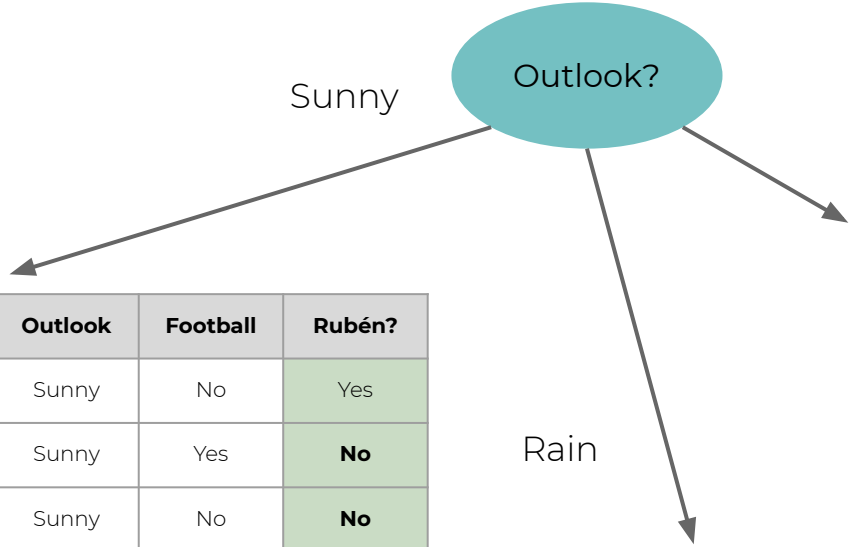
$$H(P_1) = -((3/5) * log_2(3/5) + (2/5) * log_2(2/5)) = 0.97$$

Temp	Humidity	Outlook	Football	Rubén?
Hot	77	Overcast	No	Yes
Cool	72	Overcast	Yes	Yes
Mild	77	Overcast	Yes	Yes
Hot	74	Overcast	No	Yes

Temp	Humidity	Outlook	Football	Rubén?
Cool	70	Rain	No	Yes
Mild	69	Rain	No	Yes
MIld	77	Rain	Yes	**No**
Cool	73	Rain	Yes	**No**
Mild	78	Rain	No	Yes

IMMUNE CODING INSTITUTE

$$H(P) = -\sum_{Ci} p_{C_i} log_2(p_{C_i})$$



Temp	Humidity	Outlook	Football	Rubén?
Mild	80	Sunny	No	Yes
Hot	75	Sunny	Yes	**No**
Mild	77	Sunny	No	**No**
Cool	70	Sunny	No	Yes
Mild	65	Sunny	Yes	Yes

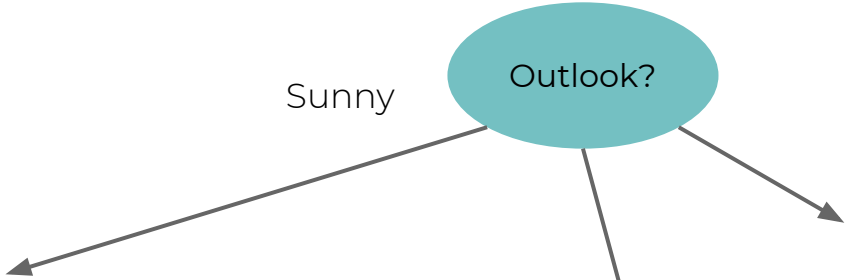
$$H(P_1) = -((3/5) * log_2(3/5) + (2/5) * log_2(2/5)) = 0.97$$

Temp	Humidity	Outlook	Football	Rubén?
Hot	77	Overcast	No	Yes
Cool	72	Overcast	Yes	Yes
Mild	77	Overcast	Yes	Yes
Hot	74	Overcast	No	Yes

Temp	Humidity	Outlook	Football	Rubén?
Cool	70	Rain	No	Yes
Mild	69	Rain	No	Yes
MIld	77	Rain	Yes	**No**
Cool	73	Rain	Yes	**No**
Mild	78	Rain	No	Yes

$$H(P_2) = -((3/5) * log_2(3/5) + (2/5) * log_2(2/5)) = 0.97$$

$$H(P) = -\sum_{Ci} p_{C_i} log_2(p_{C_i})$$



Temp	Humidity	Outlook	Football	Rubén?
Mild	80	Sunny	No	Yes
Hot	75	Sunny	Yes	**No**
Mild	77	Sunny	No	**No**
Cool	70	Sunny	No	Yes
Mild	65	Sunny	Yes	Yes

$$H(P_1) = -((3/5) * log_2(3/5) + (2/5) * log_2(2/5)) = 0.97$$

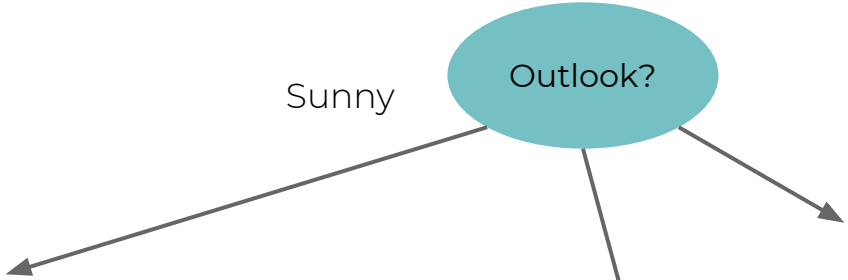
Temp	Humidity	Outlook	Football	Rubén?
Hot	77	Overcast	No	Yes
Cool	72	Overcast	Yes	Yes
Mild	77	Overcast	Yes	Yes
Hot	74	Overcast	No	Yes

$$H(P_3) = -((4/4) * log_2(4/4) + (0/4) * log_2(0/4)) = 0$$

Temp	Humidity	Outlook	Football	Rubén?
Cool	70	Rain	No	Yes
Mild	69	Rain	No	Yes
MIld	77	Rain	Yes	**No**
Cool	73	Rain	Yes	**No**
Mild	78	Rain	No	Yes

$$H(P_2) = -((3/5) * log_2(3/5) + (2/5) * log_2(2/5)) = 0.97$$

$$H(P) = -\sum_{Ci} p_{C_i} log_2(p_{C_i})$$



Temp	Humidity	Outlook	Football	Rubén?
Mild	80	Sunny	No	Yes
Hot	75	Sunny	Yes	**No**
Mild	77	Sunny	No	**No**
Cool	70	Sunny	No	Yes
Mild	65	Sunny	Yes	Yes

$$H(P_1) = -((3/5) * log_2(3/5) + (2/5) * log_2(2/5)) = 0.97$$

$$H_{mean} = \frac{5}{14} * 0.97 + \frac{5}{14} * 0.97 + \frac{4}{14} * 0 = 0.69$$

Temp	Humidity	Outlook	Football	Rubén?
Hot	77	Overcast	No	Yes
Cool	72	Overcast	Yes	Yes
Mild	77	Overcast	Yes	Yes
Hot	74	Overcast	No	Yes

$$H(P_3) = -((4/4) * log_2(4/4) + (0/4) * log_2(0/4)) = 0$$

Temp	Humidity	Outlook	Football	Rubén?
Cool	70	Rain	No	Yes
Mild	69	Rain	No	Yes
MIld	77	Rain	Yes	**No**
Cool	73	Rain	Yes	**No**
Mild	78	Rain	No	Yes

$$H(P_2) = -((3/5) * log_2(3/5) + (2/5) * log_2(2/5)) = 0.97$$

What about continuous variables?

Humidity	Rubén?
80	Yes
75	**No**
77	Yes
70	Yes
72	Yes
77	**No**
70	Yes
69	Yes
65	Yes
77	Yes
74	Yes
77	**No**
73	**No**
78	Yes

IMMUNE CODING INSTITUTE

What about continuous variables?

65 - **Yes**, 69 - **Yes**, 70 - **Yes Yes**, 72 - **Yes**, 73 - **Yes**, 74 - **Yes**, 75 - **No**, 77 - **Yes Yes No No**, 78 - **Yes**, 80 - **Yes**,

Humidity	Rubén?
80	Yes
75	**No**
77	Yes
70	Yes
72	Yes
77	**No**
70	Yes
69	Yes
65	Yes
77	Yes
74	Yes
77	**No**
73	**No**
78	Yes

What about continuous variables?

$$H(P) = -\sum_{Ci} p_{C_i} log_2(p_{C_i})$$

65 - **Yes**, 69 - **Yes**, 70 - **Yes Yes**, 72 - **Yes**, 73 - **Yes**, 74 - **Yes**, | 75 - **No**, 77 - **Yes Yes No No**, 78 - **Yes**, 80 - **Yes**,

7 Yes, 0 No

$X = 74,5$

4 Yes, 3 No

What about continuous variables?

$$H(P) = -\sum_{Ci} p_{C_i} log_2(p_{C_i})$$

65 - **Yes**, 69 - **Yes**, 70 - **Yes Yes**, 72 - **Yes**, 73 - **Yes**, 74 - **Yes**, | 75 - **No**, 77 - **Yes Yes No No**, 78 - **Yes**, 80 - **Yes**,

$H1_{mean} = 0.49$

7 Yes, 0 No

$X = 74,5$

4 Yes, 3 No

IMMUNE CODING INSTITUTE

What about continuous variables?

$$H(P) = -\sum_{Ci} p_{C_i} log_2(p_{C_i})$$

65 - **Yes**, 69 - **Yes**, 70 - **Yes Yes**, 72 - **Yes**, 73 - **Yes**, 74 - **Yes**, 75 - **No**, 77 - **Yes Yes No No**, 78 - **Yes**, 80 - **Yes**,

9 Yes, 4 No

$X = 77,5$

2 Yes, 0 No

What about continuous variables?

$$H(P) = -\sum_{Ci} p_{C_i} log_2(p_{C_i})$$

65 - **Yes**, 69 - **Yes**, 70 - **Yes Yes**, 72 - **Yes**, 73 - **Yes**, 74 - **Yes**, 75 - **No**, 77 - **Yes Yes No No**, 78 - **Yes**, 80 - **Yes**,

$H2_{mean} = 0.67$

9 Yes, 4 No

$X = 77,5$

2 Yes, 0 No

IMMUNE CODING INSTITUTE

What about continuous variables?

$$H(P) = -\sum_{Ci} p_{C_i} log_2(p_{C_i})$$

65 - **Yes**, 69 - **Yes**, 70 - **Yes Yes**, 72 - **Yes**, 73 - **Yes**, 74 - **Yes**, | 75 - **No**, 77 - **Yes Yes No No**, 78 - **Yes**, 80 - **Yes**,

7 Yes, 0 No

$X = 74,5$

4 Yes, 3 No

$$H1_{mean} = 0.49$$

65 - **Yes**, 69 - **Yes**, 70 - **Yes Yes**, 72 - **Yes**, 73 - **Yes**, 74 - **Yes**, 75 - **No**, 77 - **Yes Yes No No**, | 78 - **Yes**, 80 - **Yes**,

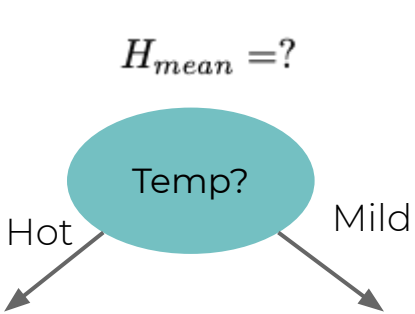
9 Yes, 4 No

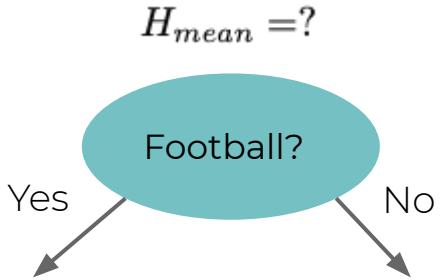
$X = 77,5$

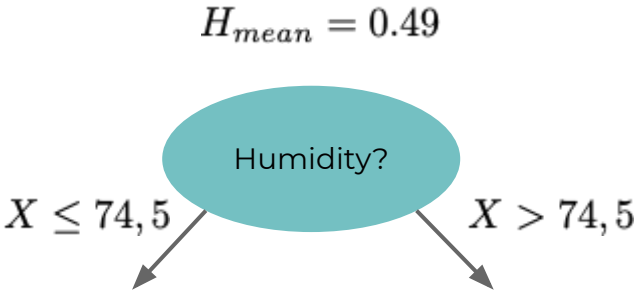
2 Yes, 0 No

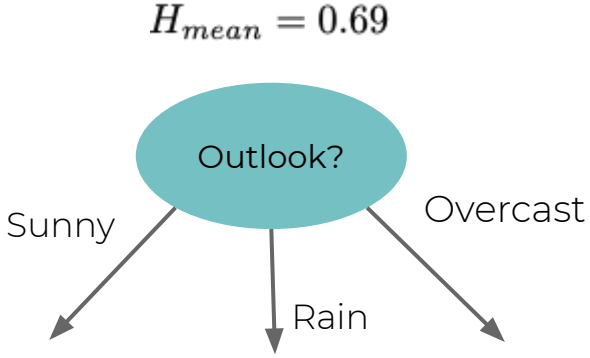
$$H2_{mean} = 0.67$$

Which feature should be the Root Node?

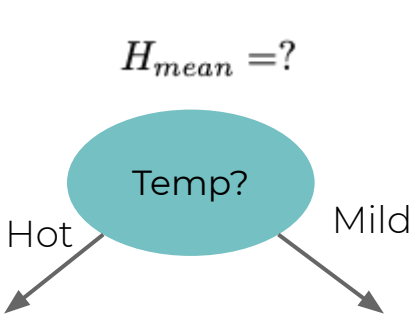


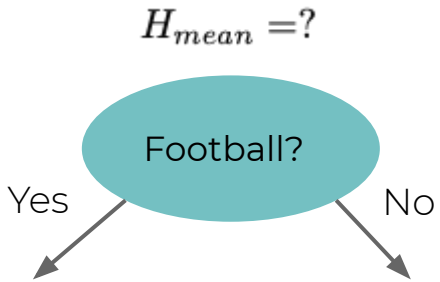


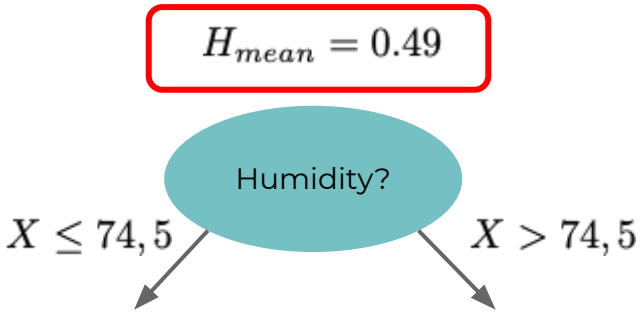


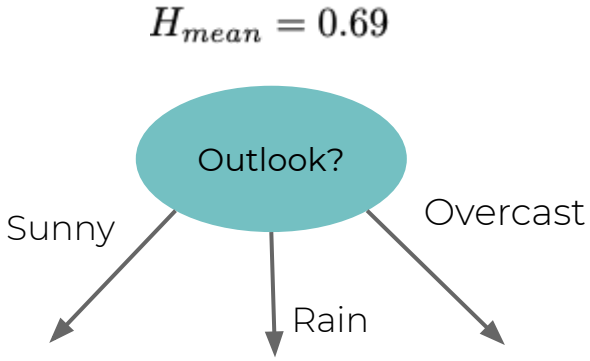


Which feature should be the Root Node?

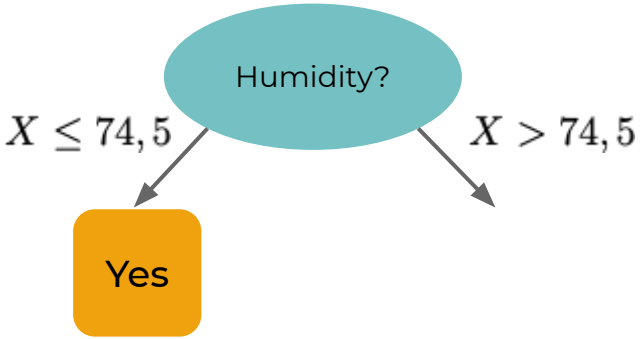








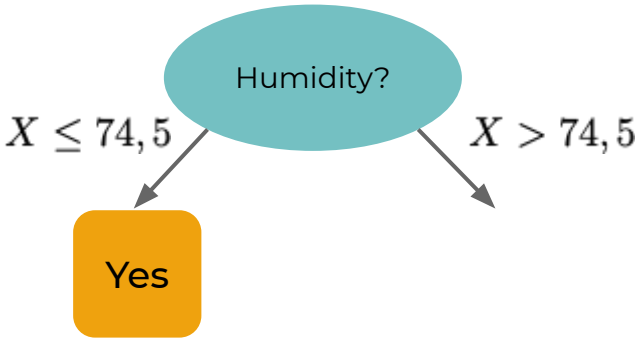
Recursive tree construction



Now that we have created the root node, the process continues recursively. We have to build sub-trees with the remaining data.

Temp	Humidity	Outlook	Football	Rubén?
Mild	80	Sunny	No	Yes
Hot	75	Sunny	Yes	**No**
Hot	77	Overcast	No	Yes
Cool	70	Rain	No	Yes
Cool	72	Overcast	Yes	Yes
Mild	77	Sunny	No	**No**
Cool	70	Sunny	No	Yes
Mild	69	Rain	No	Yes
Mild	65	Sunny	Yes	Yes
Mild	77	Overcast	Yes	Yes
Hot	74	Overcast	No	Yes
MIld	77	Rain	Yes	**No**
Cool	73	Rain	Yes	**No**
Mild	78	Rain	No	Yes

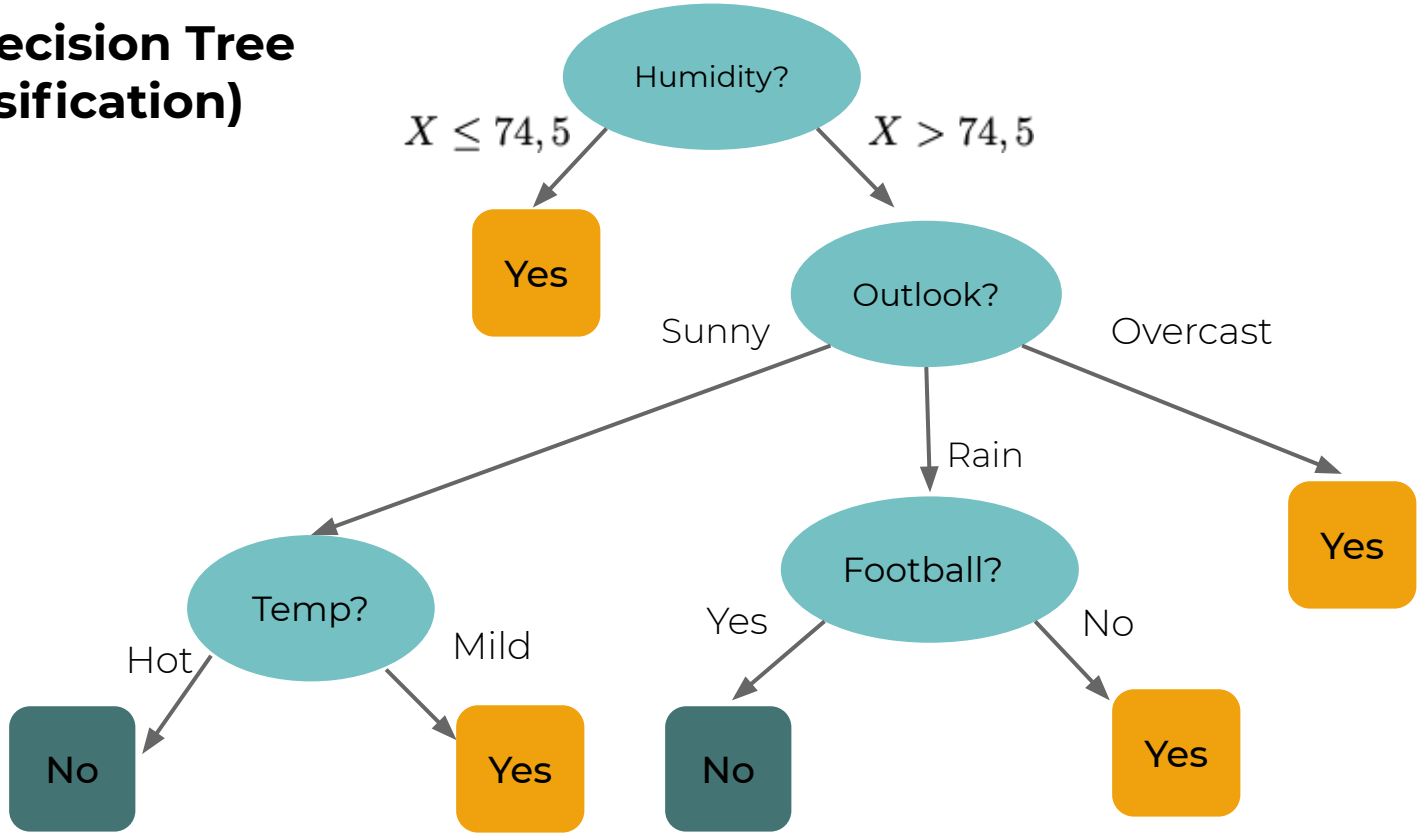
Recursive tree construction



Temp	Humidity	Outlook	Football	Rubén?
Mild	80	Sunny	No	Yes
Hot	75	Sunny	Yes	**No**
Hot	77	Overcast	No	Yes
Mild	77	Sunny	No	**No**
Mild	77	Overcast	Yes	Yes
MIld	77	Rain	Yes	**No**
Mild	78	Rain	No	Yes

Now that we have created the root node, the process continues recursively. We have to build sub-trees with the remaining data.

Final Decision Tree (Classification)



Regression Trees

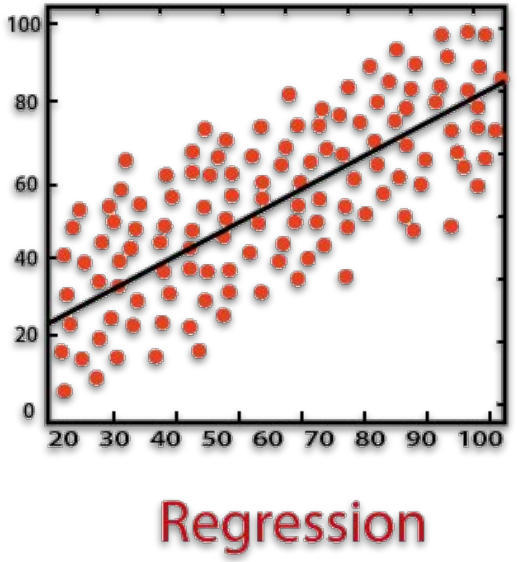
What if the output is continuous?

Instead of using entropy reduction we use variance reduction.

We can create two types of trees:

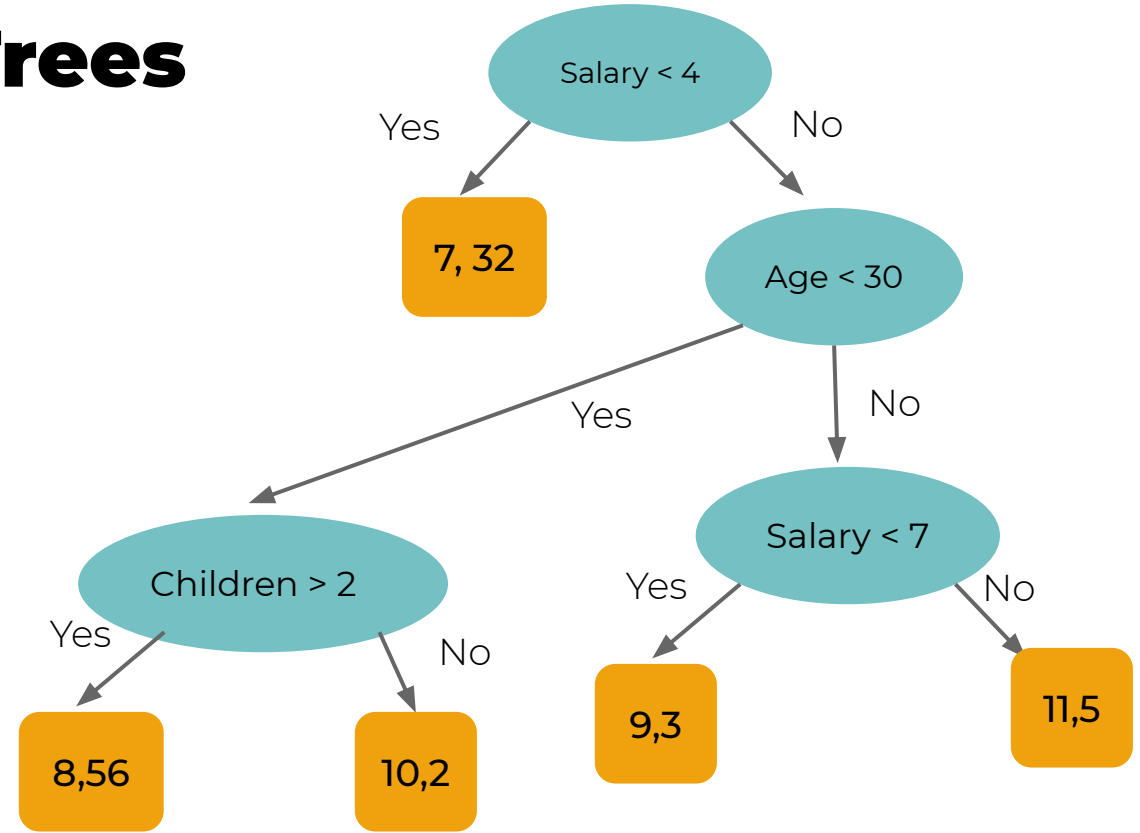
Model trees

Regression trees



Regression Trees

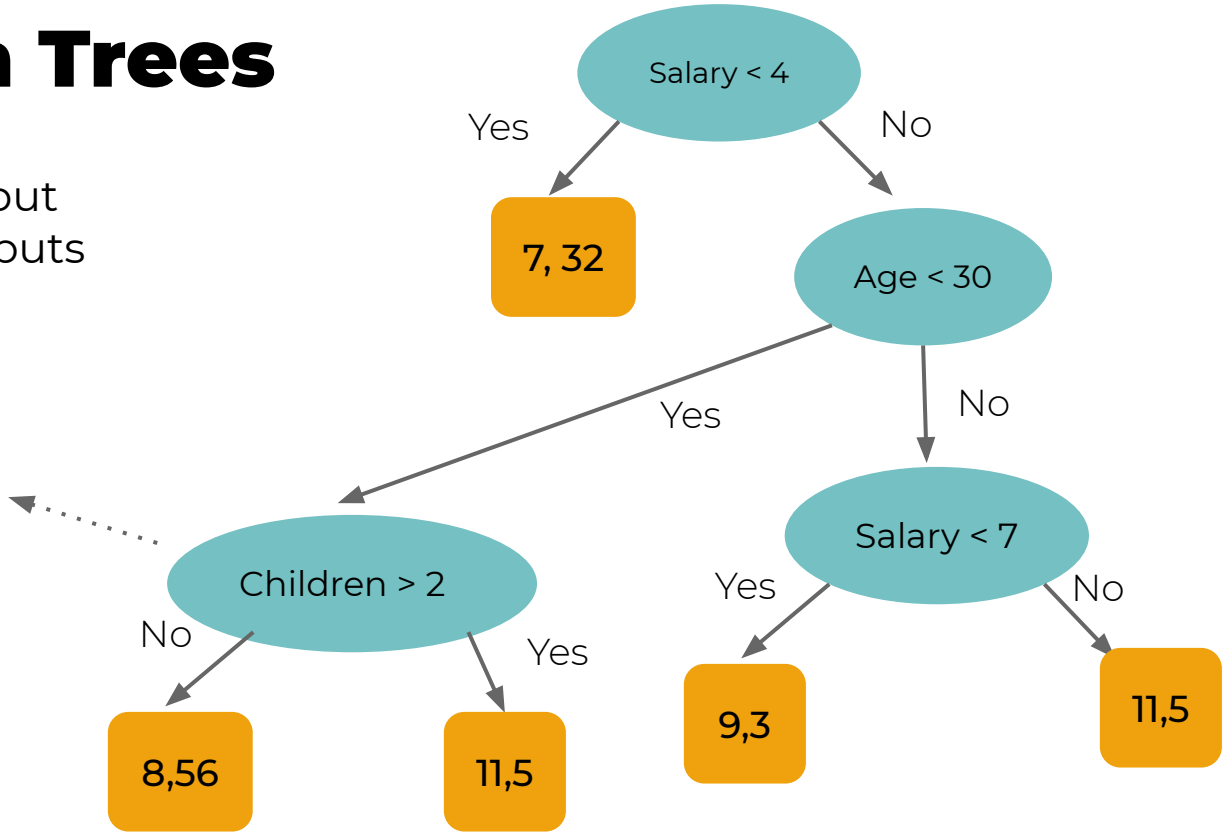
On the leaves, we will put the average of the outputs of the data that have arrived at each leaf.



Regression Trees

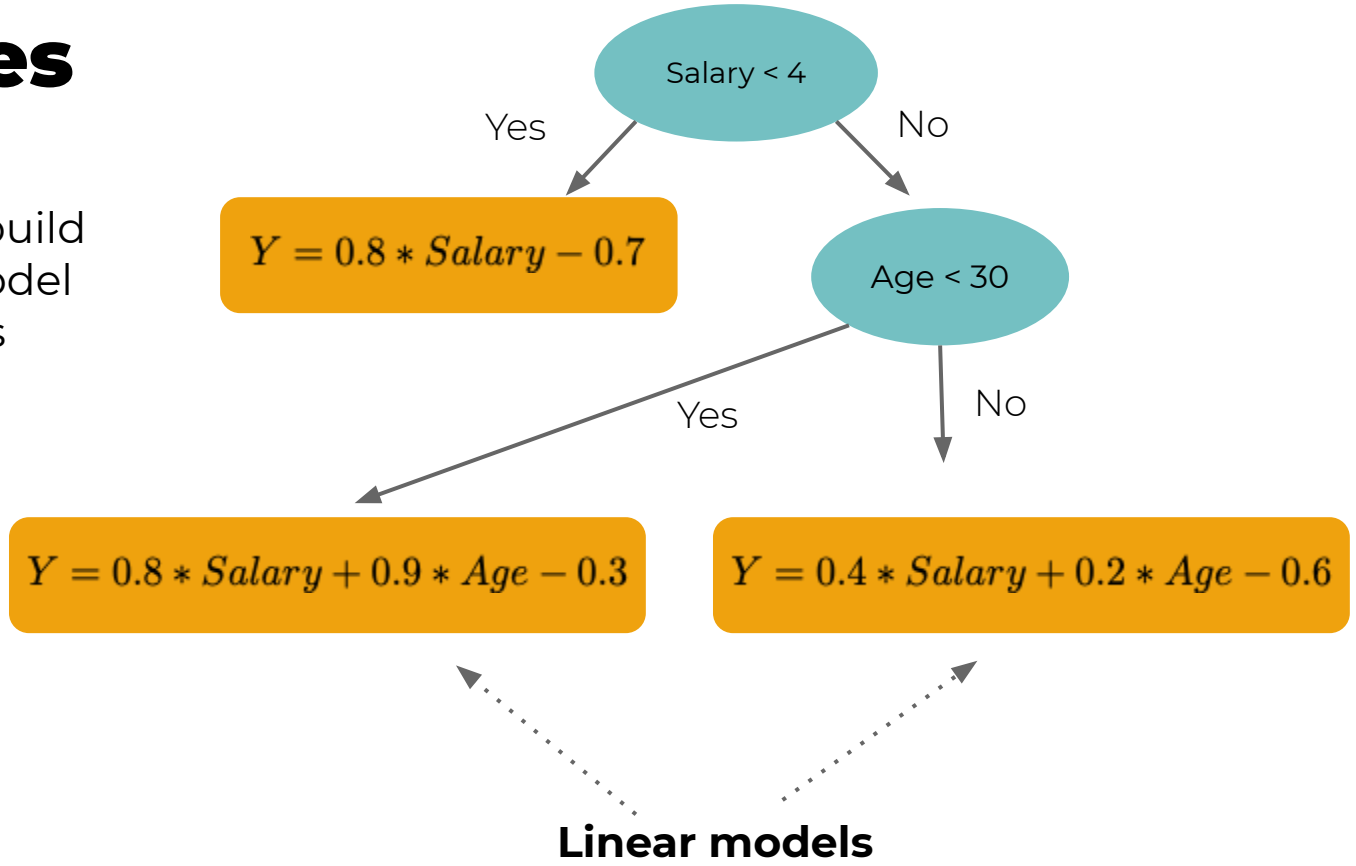
On the leaves, we will put the average of the outputs of the data that have arrived at each leaf.

Children	Salary
0	8,56
2	9.5
4	12
5	13



Model Trees

On each leaf, we will build a linear regression model with the data that has reached that leaf.



Advantages and Disadvantages of Trees

Pro

- Trees are very easy to explain to people (maybe easier to Linear regression)
- More closely mirror human decision-making
- Can be displayed graphically
- Easily interpreted even by a non-expert (small trees)
- Trees can easily handle qualitative predictors without the need to create dummy variables

Cons

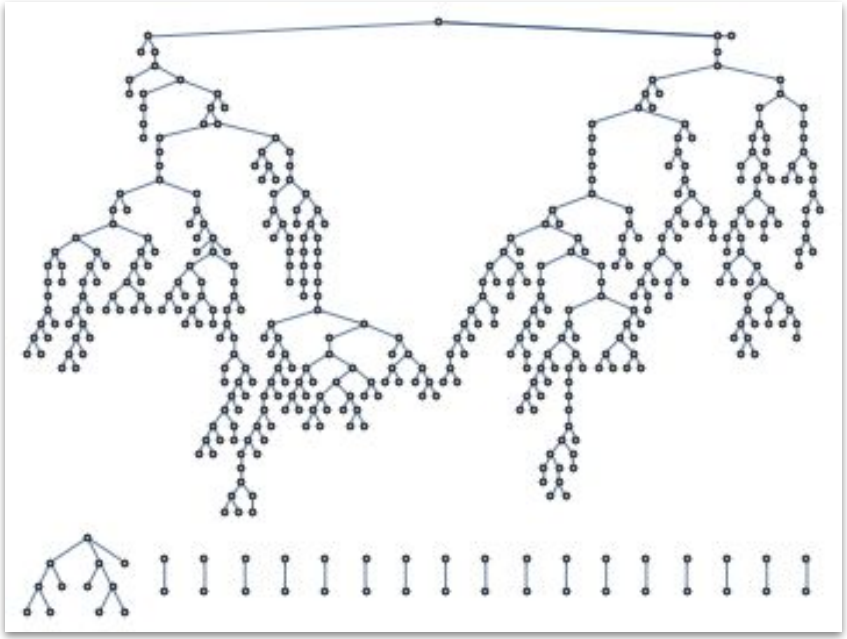
- Trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches

Tree Pruning

As the number of splits in DTs increase, their complexity rises.

In general, simpler DTs are preferred over super complex ones, since they are easier to understand and they are less likely to fall into overfitting.

Pruning: reduces the size of DTs by removing sections of the Tree that provide little predictive or classification power.



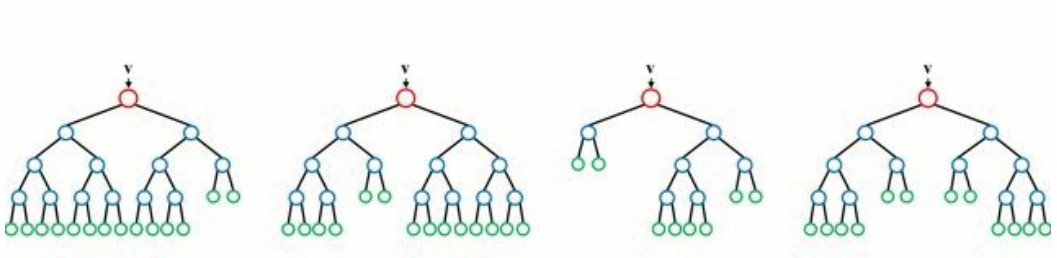
IMMUNE CODING INSTITUTE

Random Forest

To improve performance, we can use many trees with a random sample of features chosen as the split.

- A new random sample of feature is chosen for **every single tree at every single split.**

- For **classification,** *m* (sample of features) is typically chosen to be the square root of the total of features. .



Random Forest

What's the idea behind?

Suppose there we have a very **strong feature** in our data set. When using a bag of trees (**bagging)**, most of the trees will use that feature as the top split, having an ensemble of similar trees that are very similar or **highly correlated.**

By randomly leaving out candidate features from each split, **Random Forest "decorrelates" the trees**, such that the averaging process can reduce the variance of the resulting model

Bagging?

Boosting?

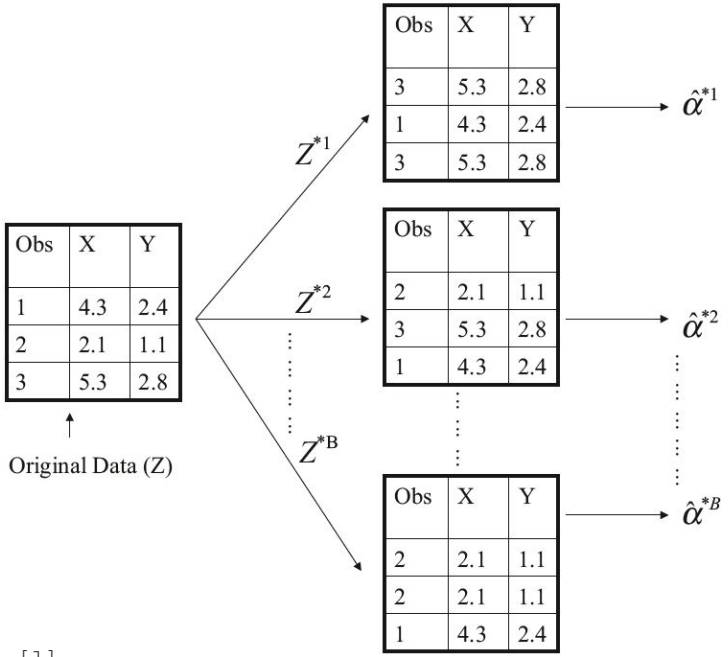
Ensemble algorithms?

IMMUNE CODING INSTITUTE

Bootstrap resampling

Quantify the uncertainty associated with statistics of a population or skills of a machine learning model (resampled → sample → population)

- from our data set (Z) to create bootstrap data sets (Z^{*1},Z^{*2},...,Z^{*B}) where B = 1000, 10 000,...
- > same observation can be sampled more than once
- Each bootstrap sample will have the as the original data set

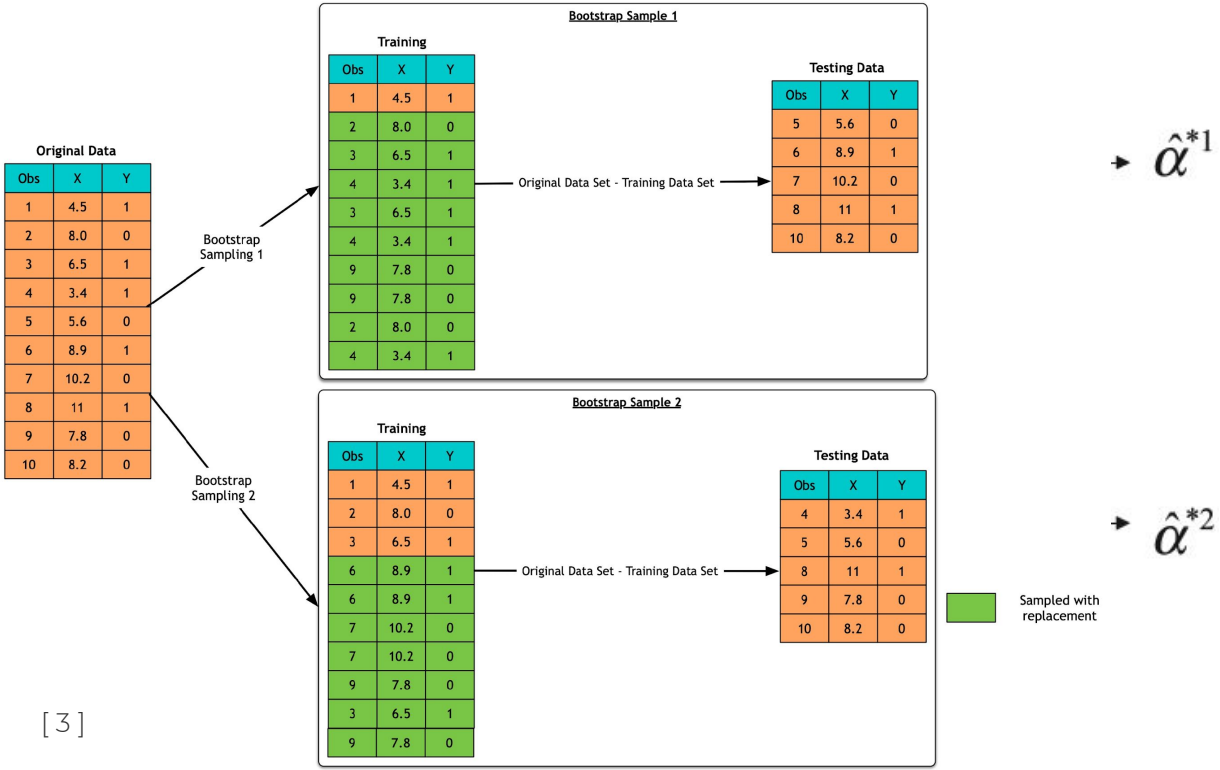


[1]

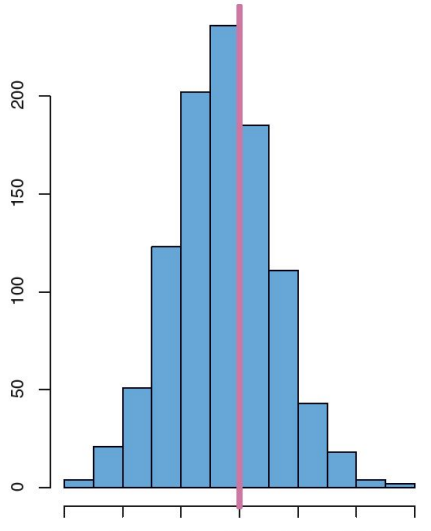
IMMUNE CODING INSTITUTE

Bootstrap

out-of-bag observations (OOB)



[3]

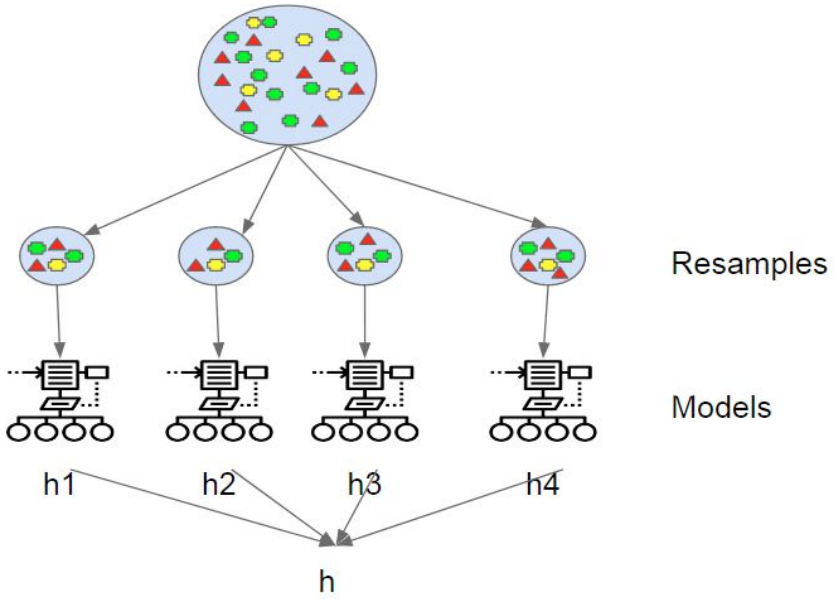


[1]

Bagging

It refers to (Bootstrap Aggregators).

“Bagging predictors is a method for generating multiples versions of a predictor and using these to get an aggregated predictor”.



Once each model has develop a prediction, the models use voting for classification or averaging for regression

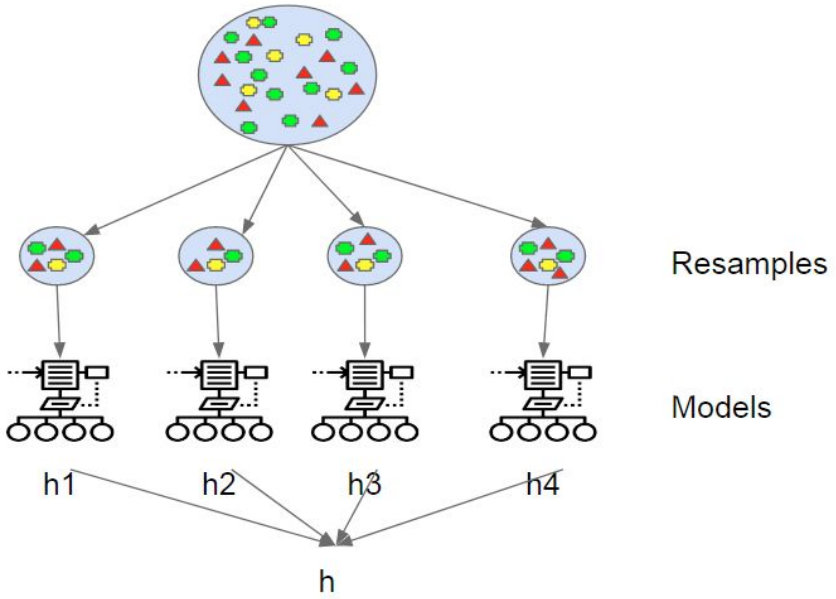
ref:
https://link.springer.com/content/pdf/10.1023/A:1018054314350.pdf

This helps to decrease **variance** i.e. reduce the **overfit**.

Bagging

It refers to (Bootstrap Aggregators).

“Bagging predictors is a method for generating multiples versions of a predictor and using these to get an aggregated predictor”.

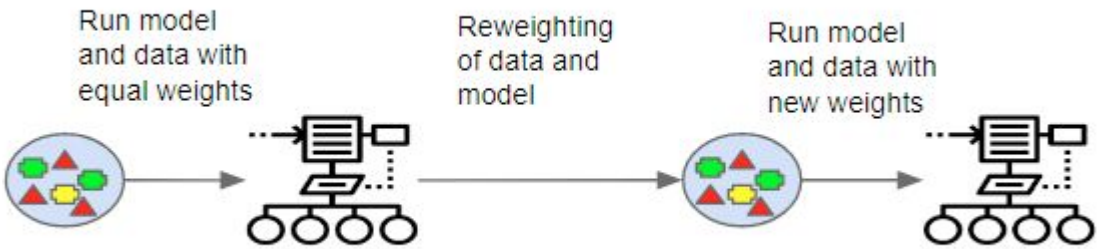


Once each model has develop a prediction, the models use voting for classification or averaging for regression

ref:
https://link.springer.com/content/pdf/10.1023/A:1018054314350.pdf

Boosting

It refers to a group of algorithms that utilize weighted averages to make weak learners into stronger learners.



The models with better outcomes have a stronger pull on the final output.


IMMUNE
CODING INSTITUTE