

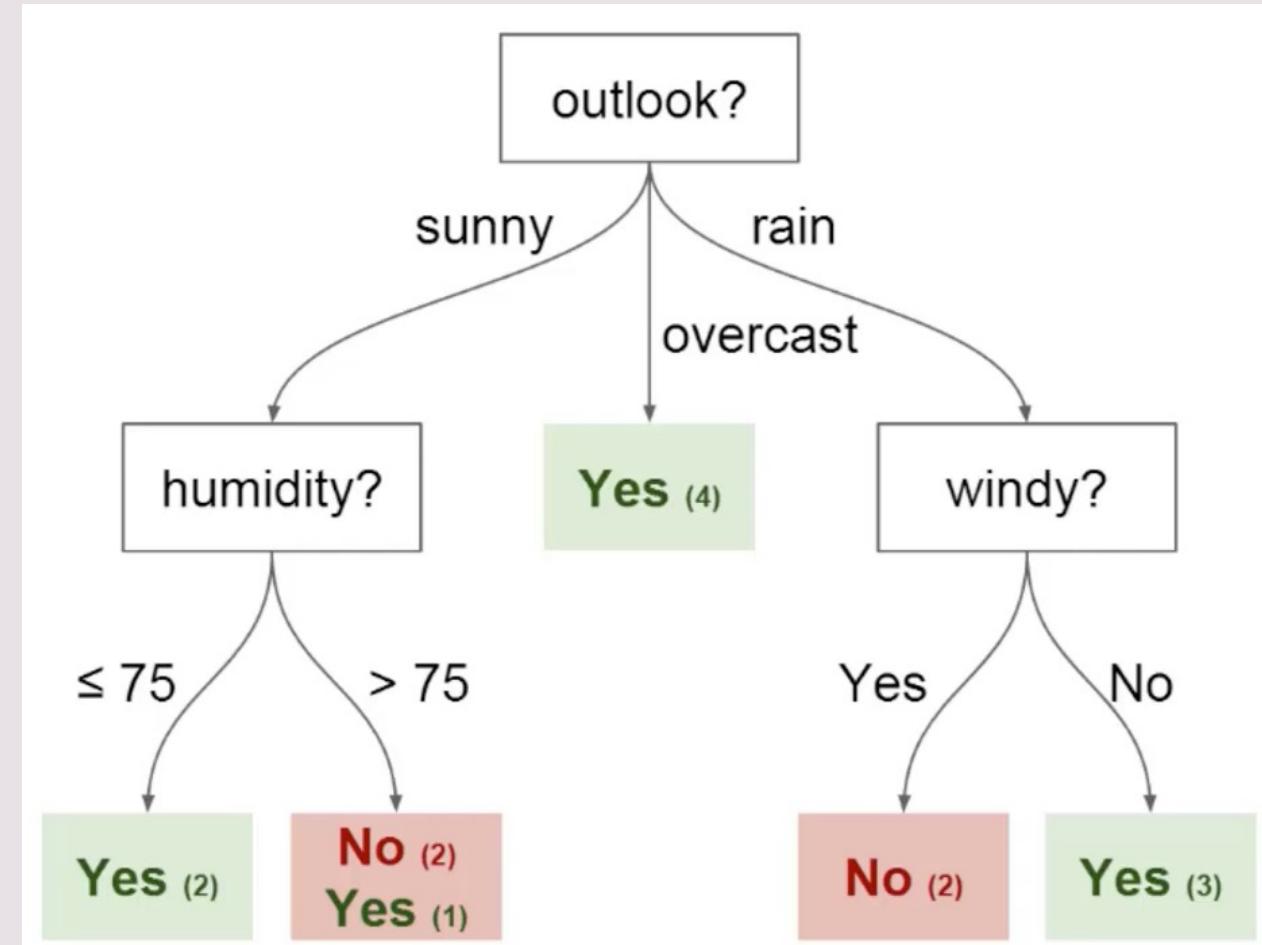
Từ trồng cây gây rừng đến XGBoost

Mì AI

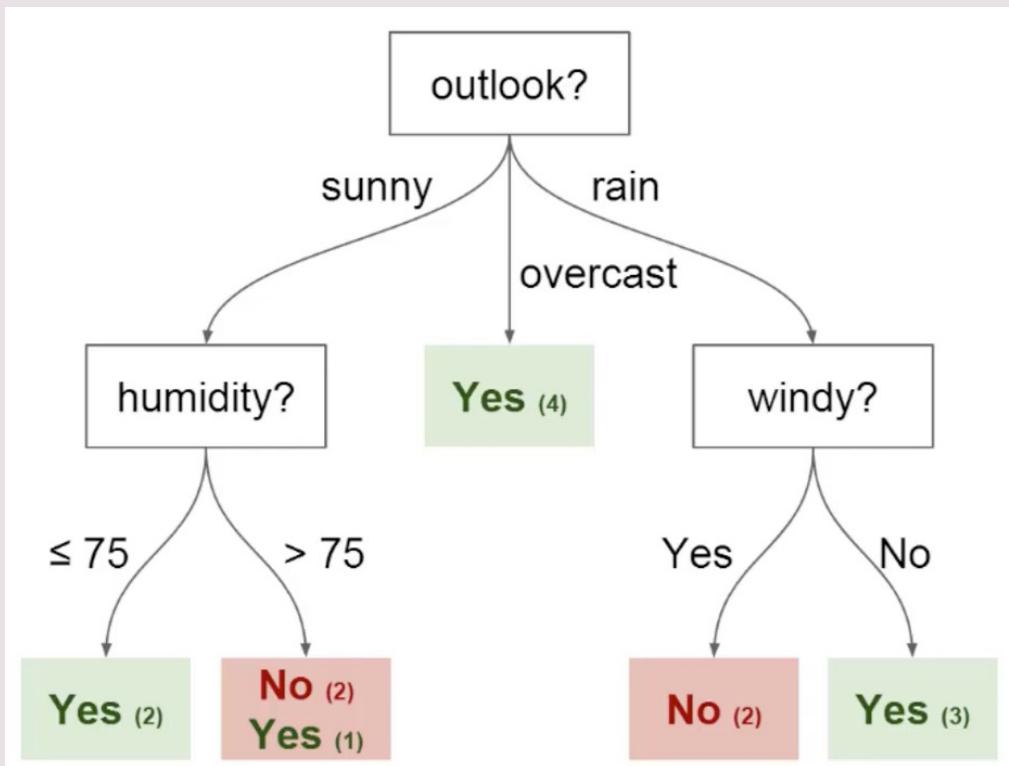
Decision Tree

- *A Decision Tree is a supervised Machine learning algorithm. It is used in both classification and regression algorithms.*
- *Decision Trees usually implement exactly the human thinking ability while making a decision, so it is easy to understand.*

Decision Tree

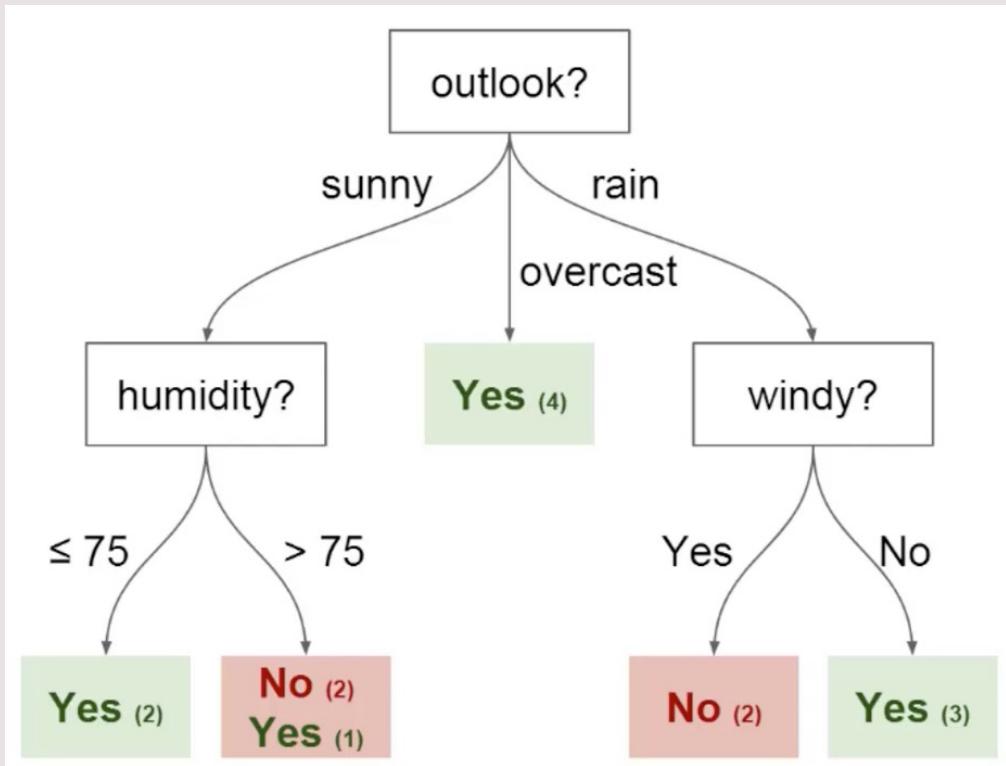


Decision Tree



- **Nodes:** It is The point where the tree splits according to the value of some attribute/feature of the dataset
- **Edges:** It directs the outcome of a split to the next node.

Decision Tree



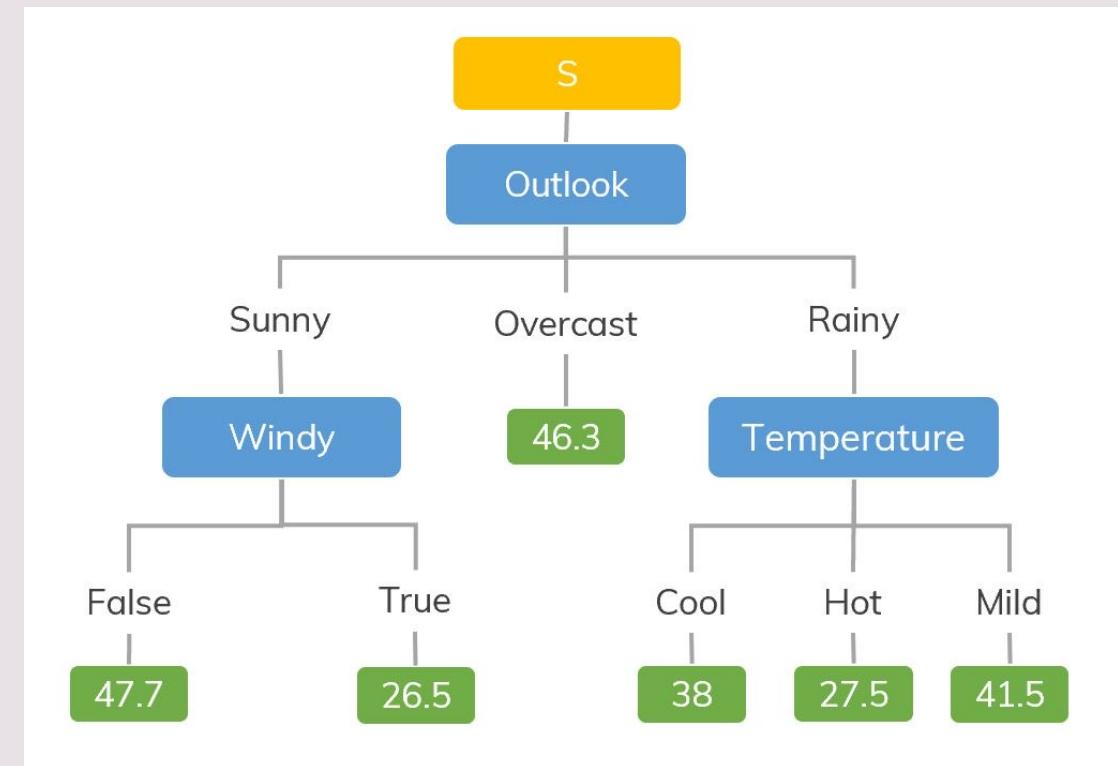
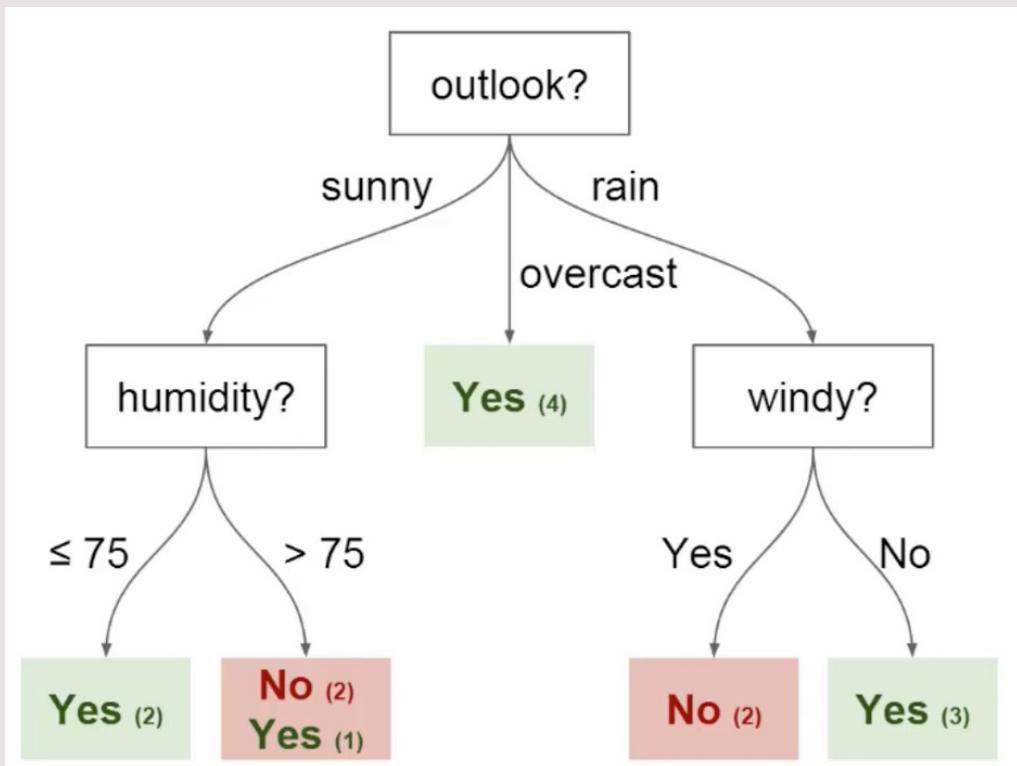
- **Root:** This is the node where the first split takes place
- **Leaves:** These are the terminal nodes that predict the outcome of the decision tree.

Decision Tree



Rainy	Cool	Normal	False	???
-------	------	--------	-------	-----

Decision Tree Type



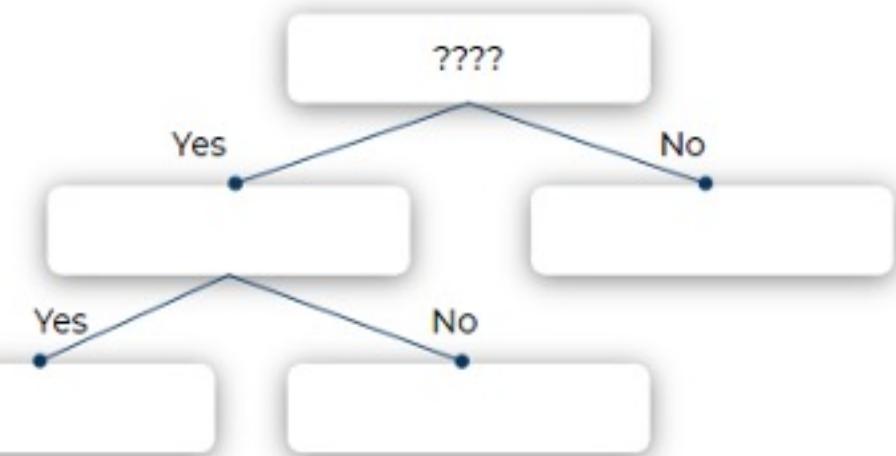
???

Classification Tree

- While building a Decision tree, the main thing is to select the best attribute from the total features list of the dataset for the root node as well as for sub-nodes. The selection of best attributes is being achieved with the help of a technique known as the Attribute selection measure (ASM).
- Methods:
 - Information Gain (ID3)
 - GINI

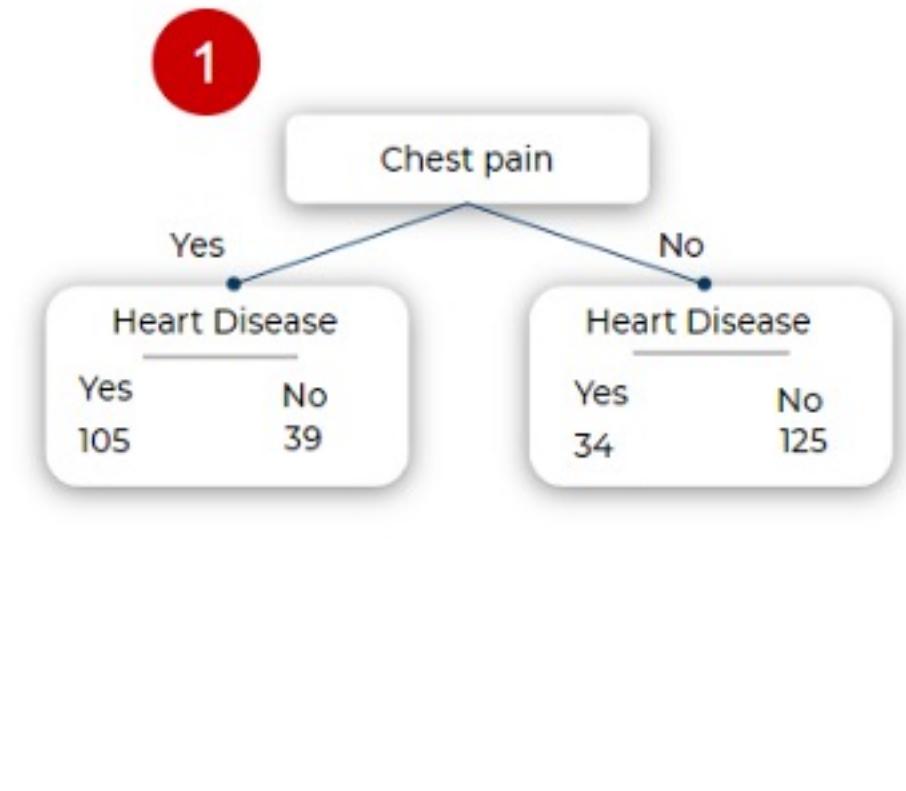
GINI

Chest pain	Good blood circulation	Blocked arteries	Heart disease
yes	no	yes	no
yes	yes	yes	yes
no	no	yes	no
yes	yes	no	yes
...



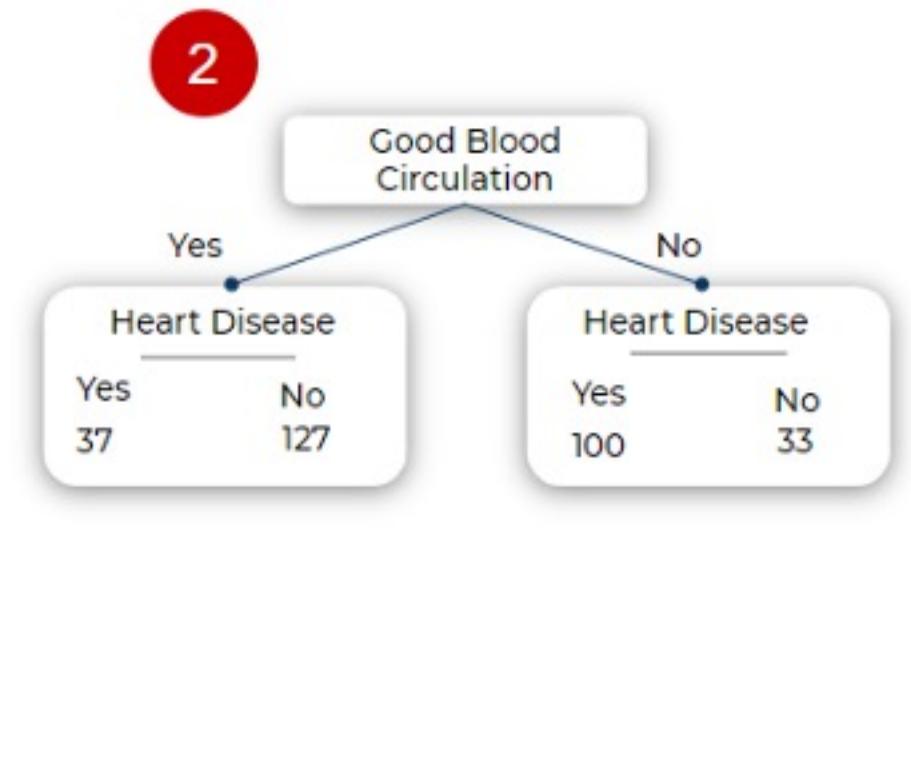
GINI

Chest pain	Good blood circulation	Blocked arteries	Heart disease
yes	no	yes	no
yes	yes	yes	yes
no	no	yes	no
yes	yes	no	yes
...



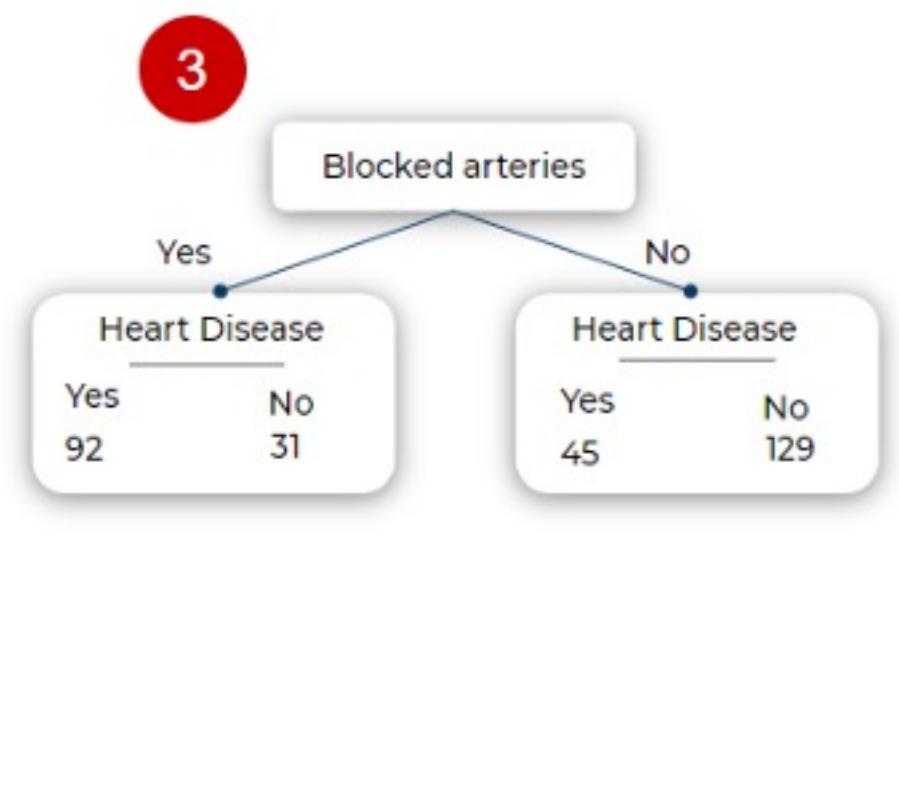
GINI

Chest pain	Good blood circulation	Blocked arteries	Heart disease
yes	no	yes	no
yes	yes	yes	yes
no	no	yes	no
yes	yes	no	yes
...

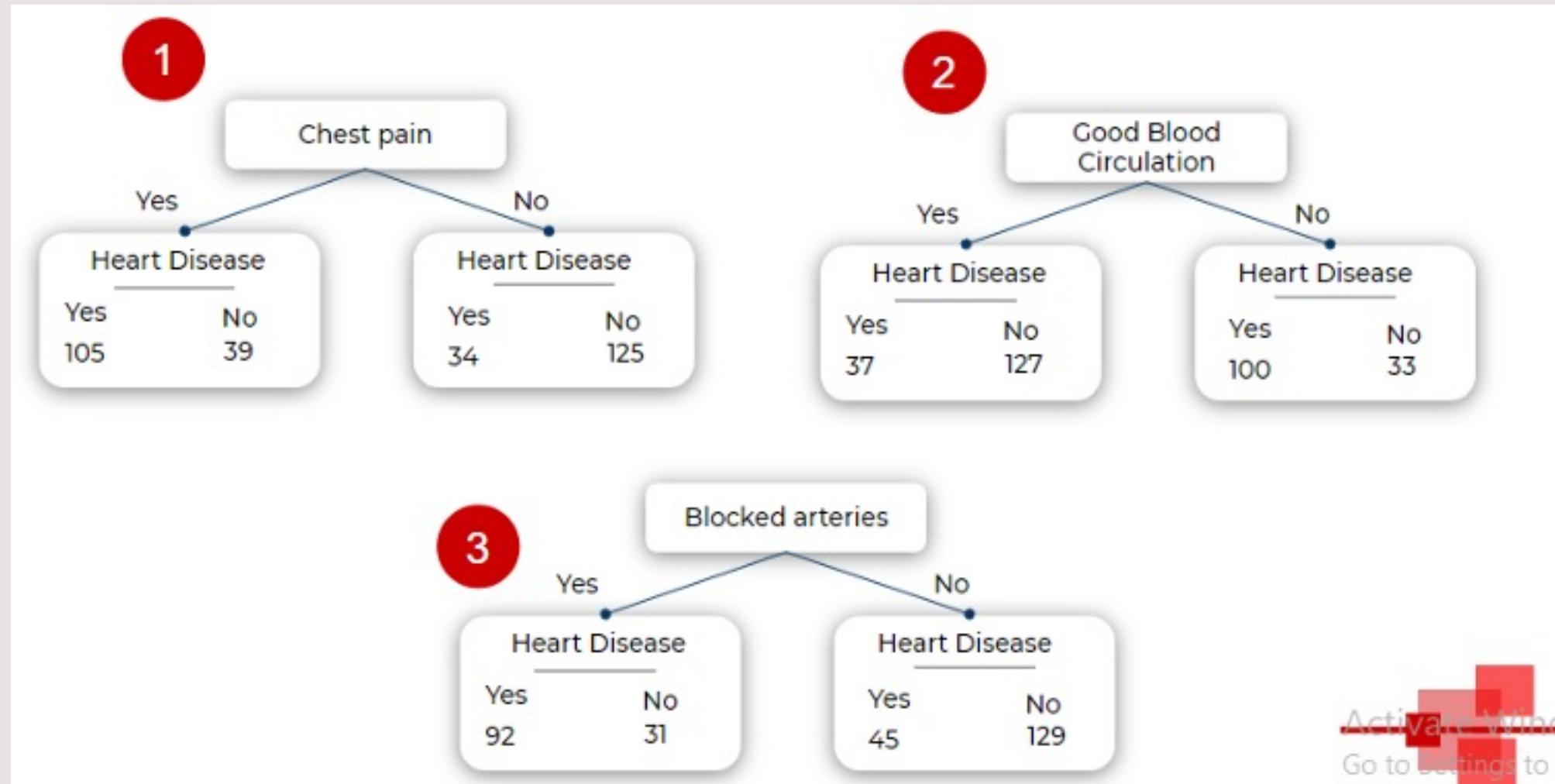


GINI

Chest pain	Good blood circulation	Blocked arteries	Heart disease
yes	no	yes	no
yes	yes	yes	yes
no	no	NA	no
yes	yes	no	yes
...



GINI



GINI

$$Gini \text{ impurity} = 1 - [P(\text{yes})]^2 - [P(\text{no})]^2$$

$$Gini \text{ impurity} = 1 - \left[\frac{105}{105+39} \right]^2 - \left[\frac{39}{105+39} \right]^2$$

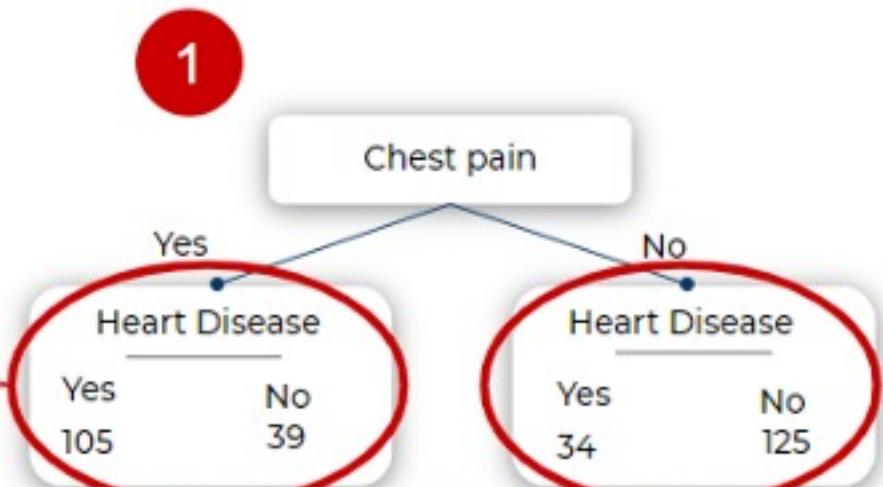
$$Gini \text{ impurity} = 0.395$$

$$Gini \text{ impurity} = 0.336$$

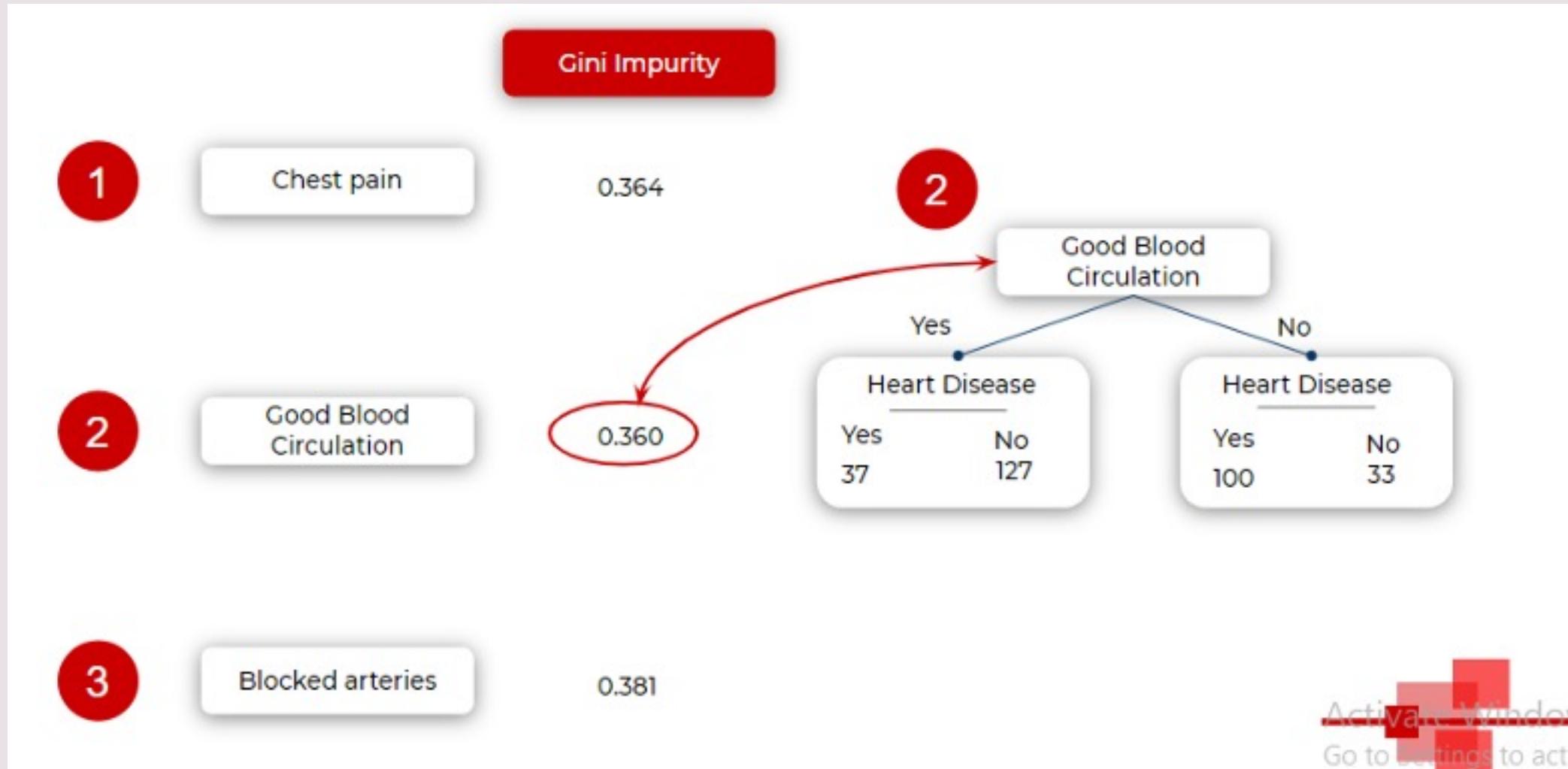
Gini impurity for chest pain = weighted avg. of leaf node

$$Gini \text{ impurity for chest pain} = \left(\frac{144}{144+159} \right) \cdot 0.395 + \left(\frac{159}{144+159} \right) \cdot 0.336$$

$$Gini \text{ impurity for chest pain} = 0.364$$

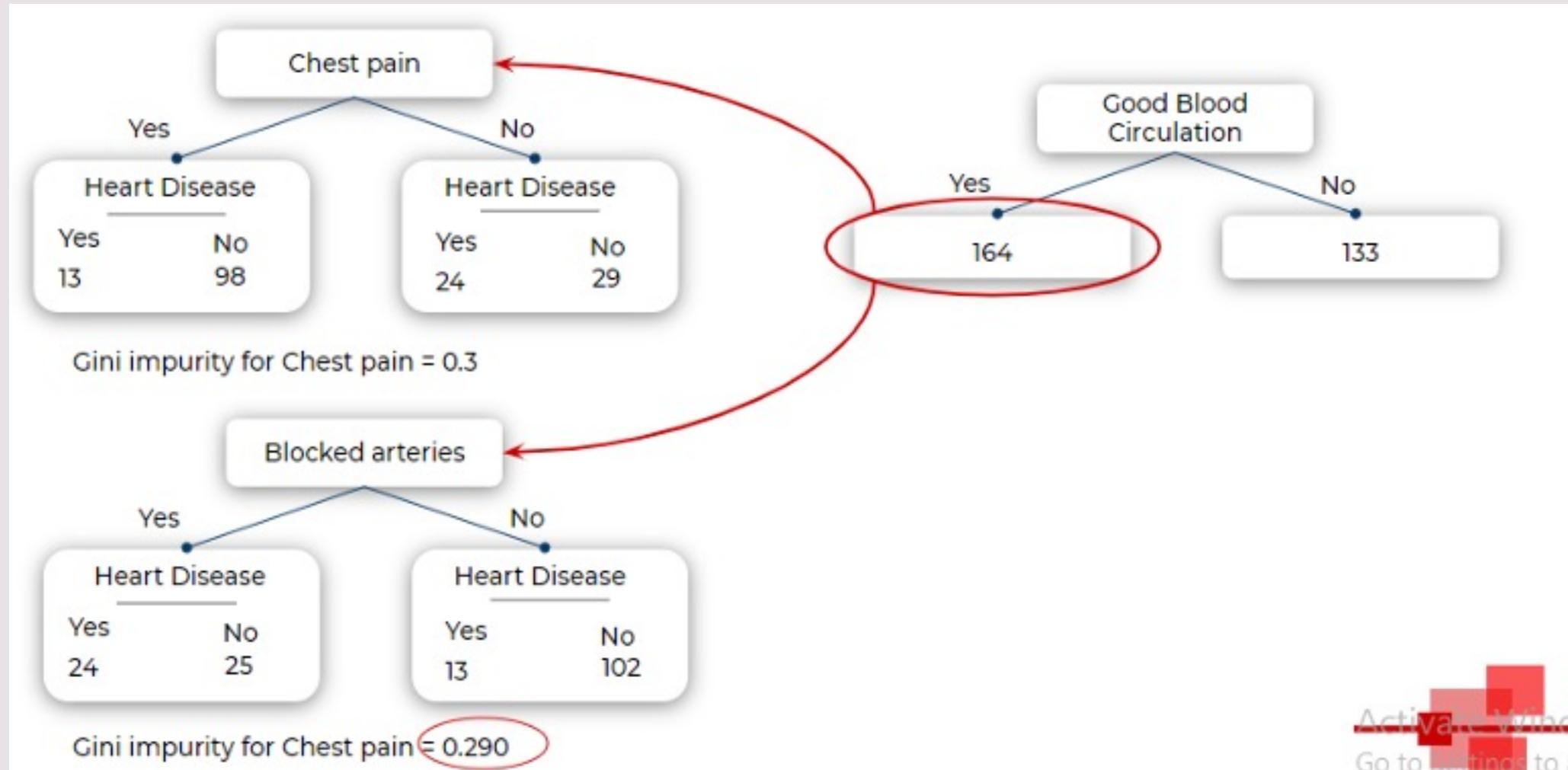


GINI

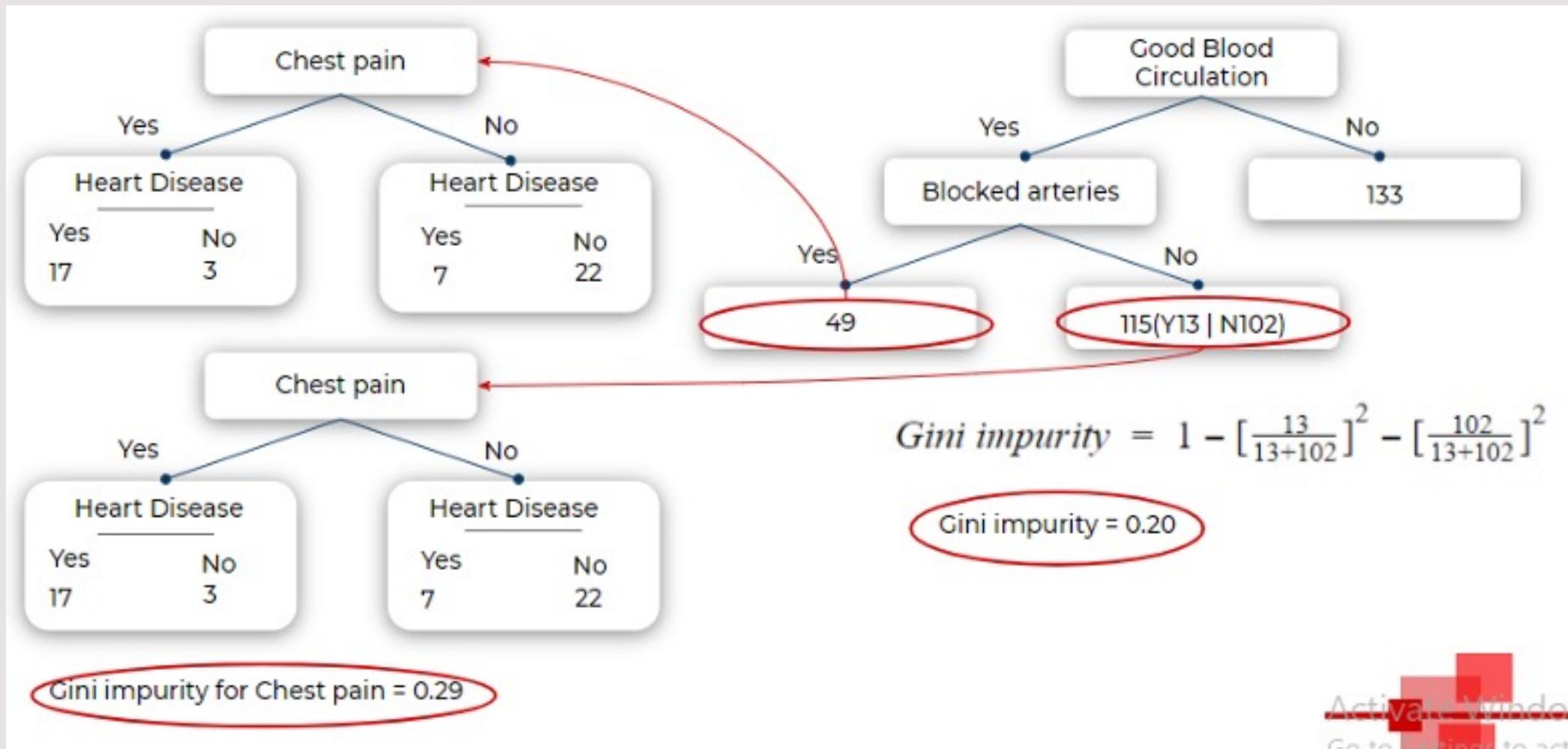


Activate Windows
Go to Settings to activate

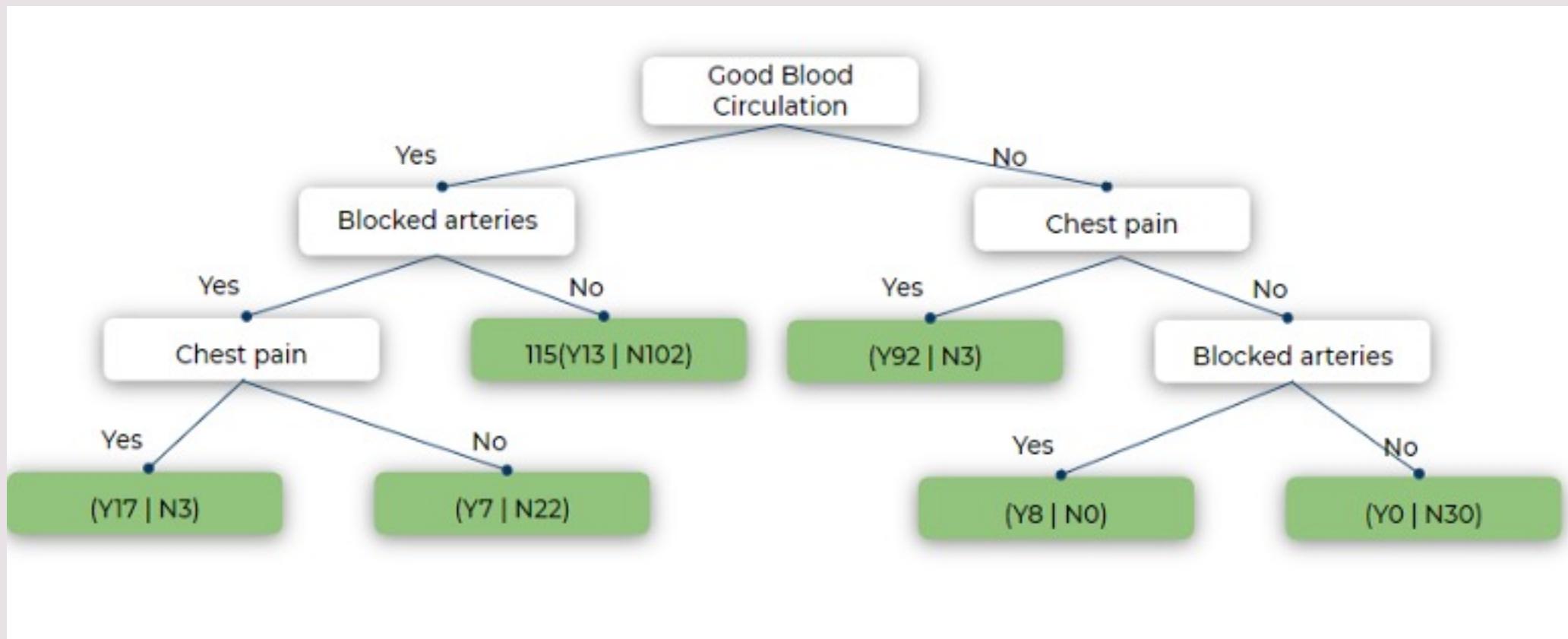
GINI



GINI



GINI



GINI

Weight	Heart disease	
155	no	
167.5		Gini impurity = 0.3
180	yes	Wt < 167.5
185		Gini impurity = 0.47
190	no	Wt < 185
200		Gini impurity = 0.27
220	yes	Wt < 200
225		Gini impurity = 0.4
222.5	yes	Wt < 222.5
...	...	

GINI

Rating	Heart disease
1	no
1	yes
3	no
2	yes
1	yes
...	...

Rating from 1 to 5

Gini impurity for 2 = 0.57

Rating < 2

Gini impurity for 3 = 0.17

Rating < 3

Gini impurity for 4 = 0.3

Rating < 4

Note: We don't need to calculate
Gini impurity for 1 & 5

GINI

Skin Color	Heart disease
White	no
Brown	yes
Brown	no
Black	yes
White	yes
...	...

Skin color: White, Brown, Black

Gini impurity for all following

White

White or Brown

Brown

White or Black

Black

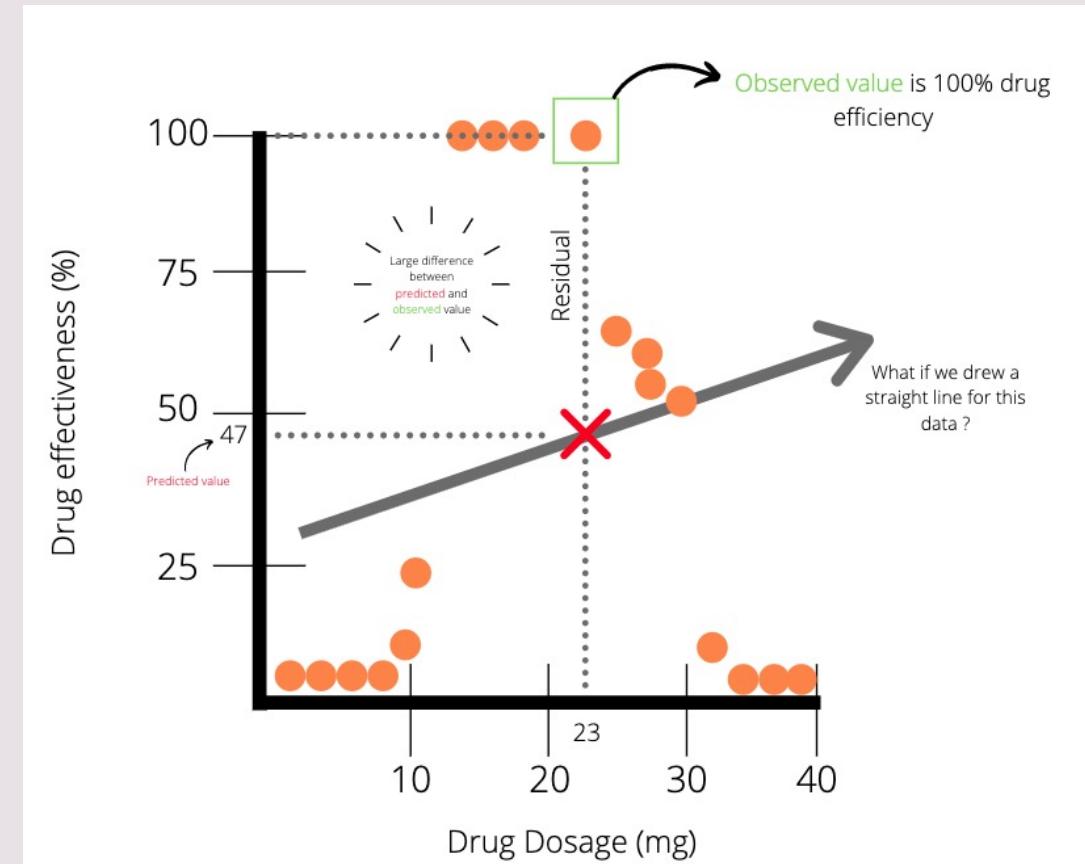
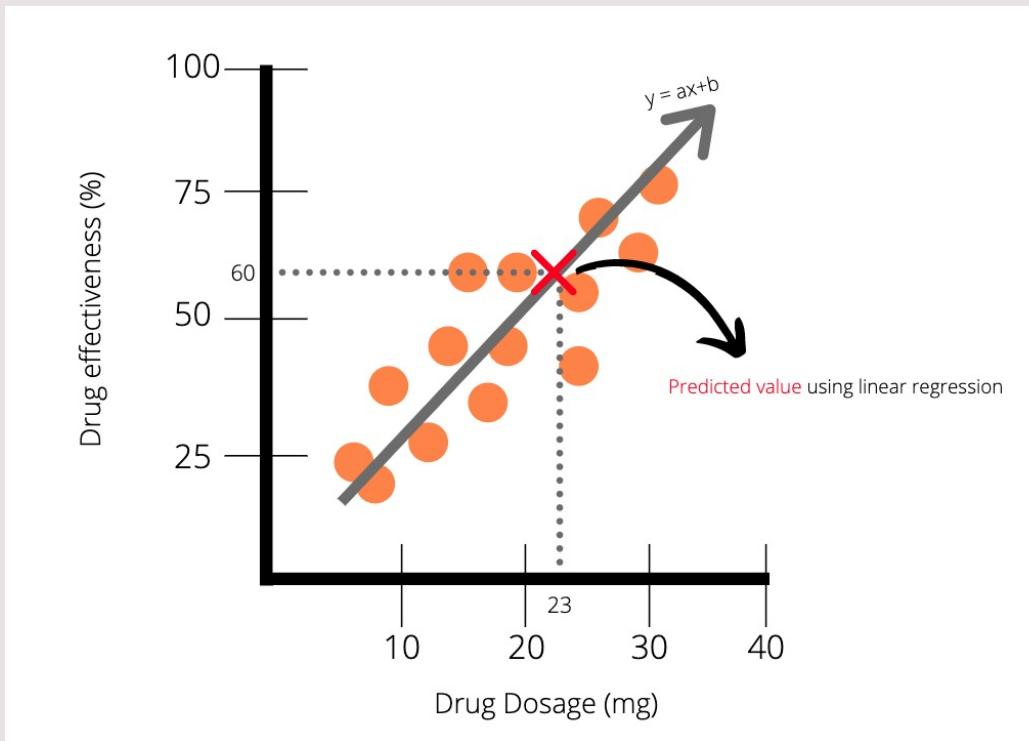
Black or Brown

Note: We don't need to calculate Gini impurity for White or Brown or Black

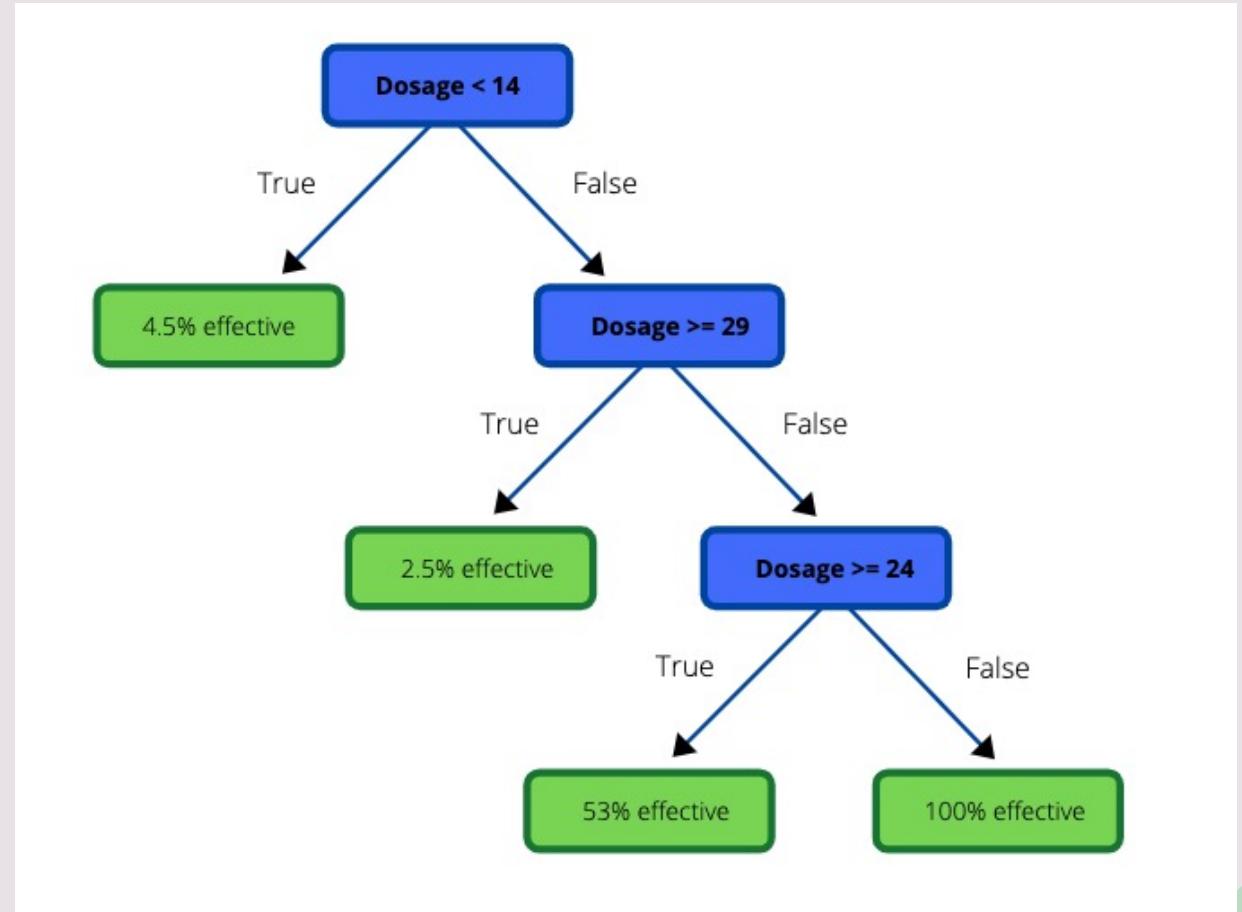


Go

Regression Tree

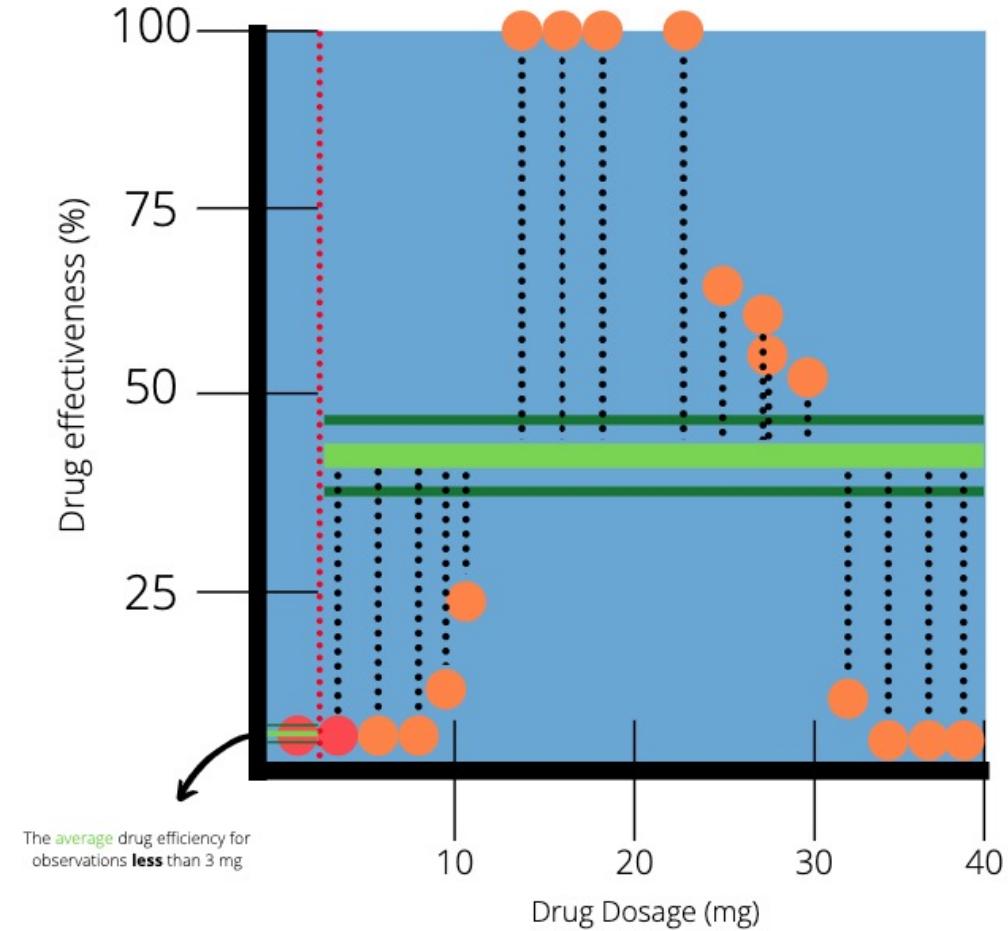
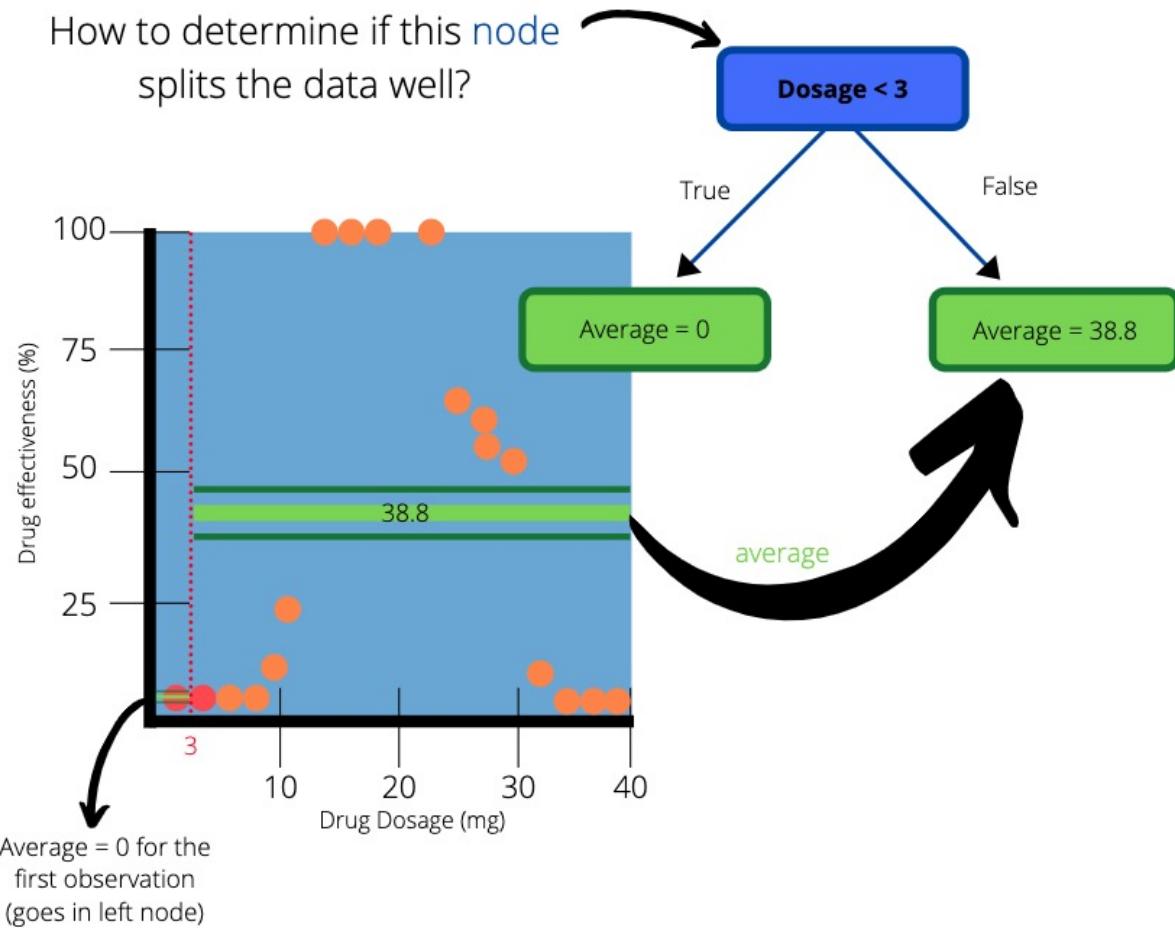


Regression Tree

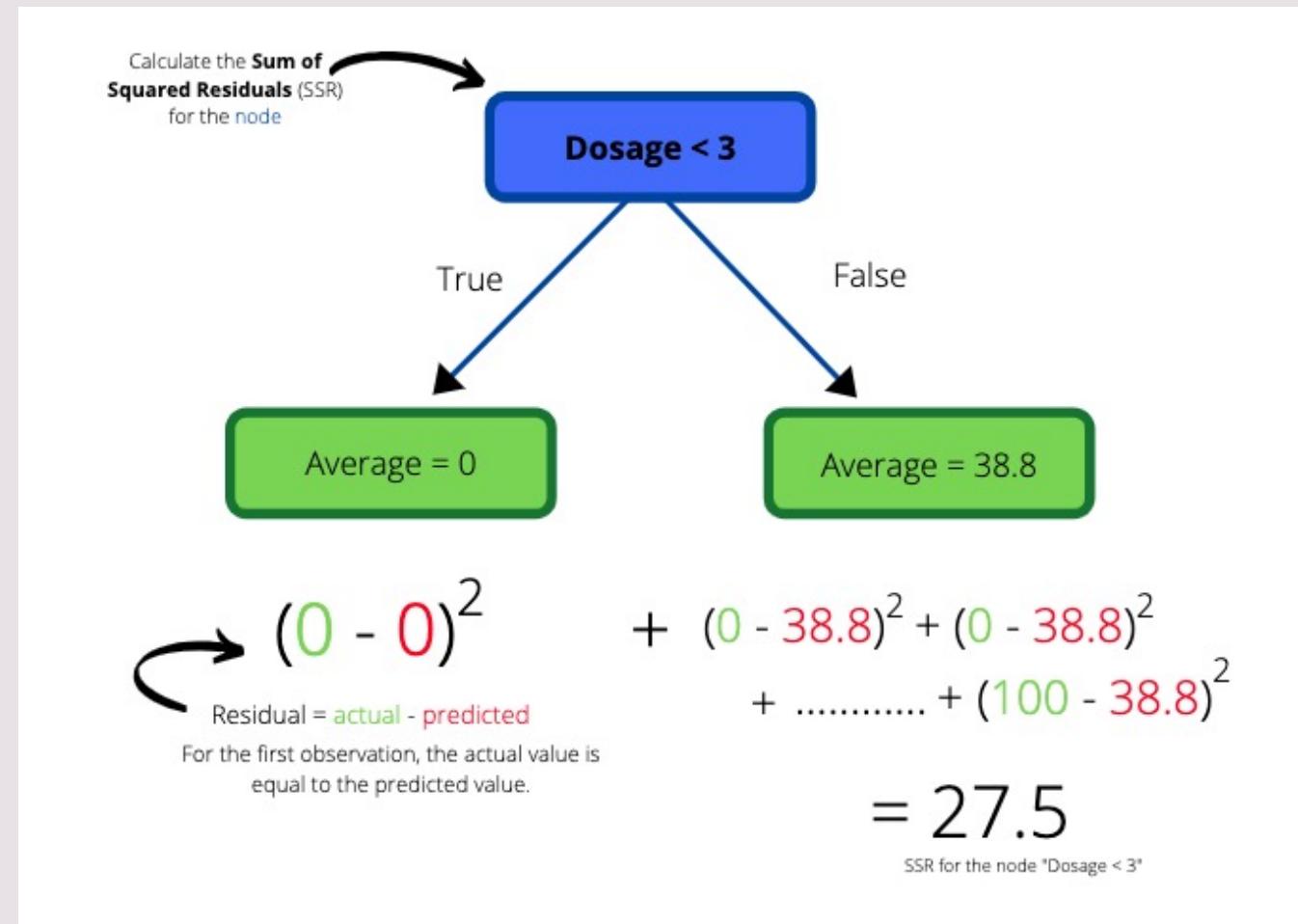


Regresion Tree

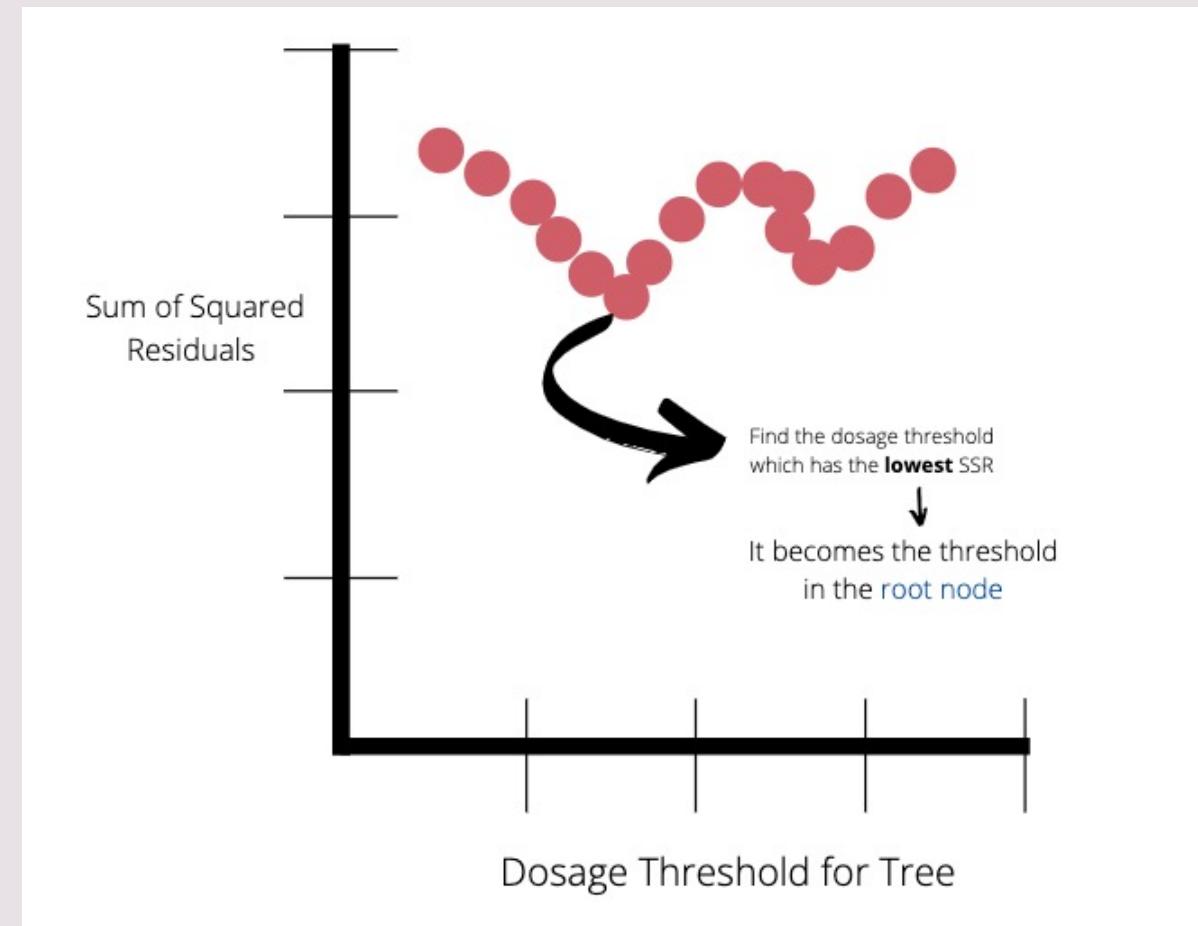
How to determine if this node splits the data well?



Regression Tree



Regression Tree



Pro and Cons

Advantages of a decision tree

- **Easy to visualize and interpret:** Its graphical representation is very intuitive to understand and it does not require any knowledge of statistics to interpret it.
- **Useful in data exploration:** We can easily identify the most significant variable and the relation between variables with a decision tree. It can help us create new variables or put some features in one bucket.
- **Less data cleaning required:** It is fairly immune to outliers and missing data, hence less data cleaning is needed.
- **The data type is not a constraint:** It can handle both categorical and numerical data.

Pro and Cons

Disadvantages of decision tree

- **Overfitting:** single decision tree tends to overfit the data which is solved by setting constraints on model parameters i.e. height of the tree and pruning. Sensitive to noisy data. It can overfit noisy data.
- **Not exact fit for continuous data:** It losses some of the information associated with numerical variables when it classifies them into different categories.
- The small variation(or variance) in data can result in the different decision tree. This can be reduced by bagging and boosting algorithms.
- Decision trees are biased with imbalance dataset, so it is recommended that balance out the dataset before creating the decision tree.

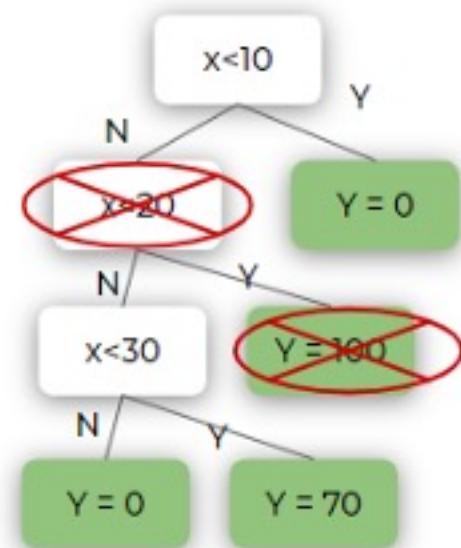
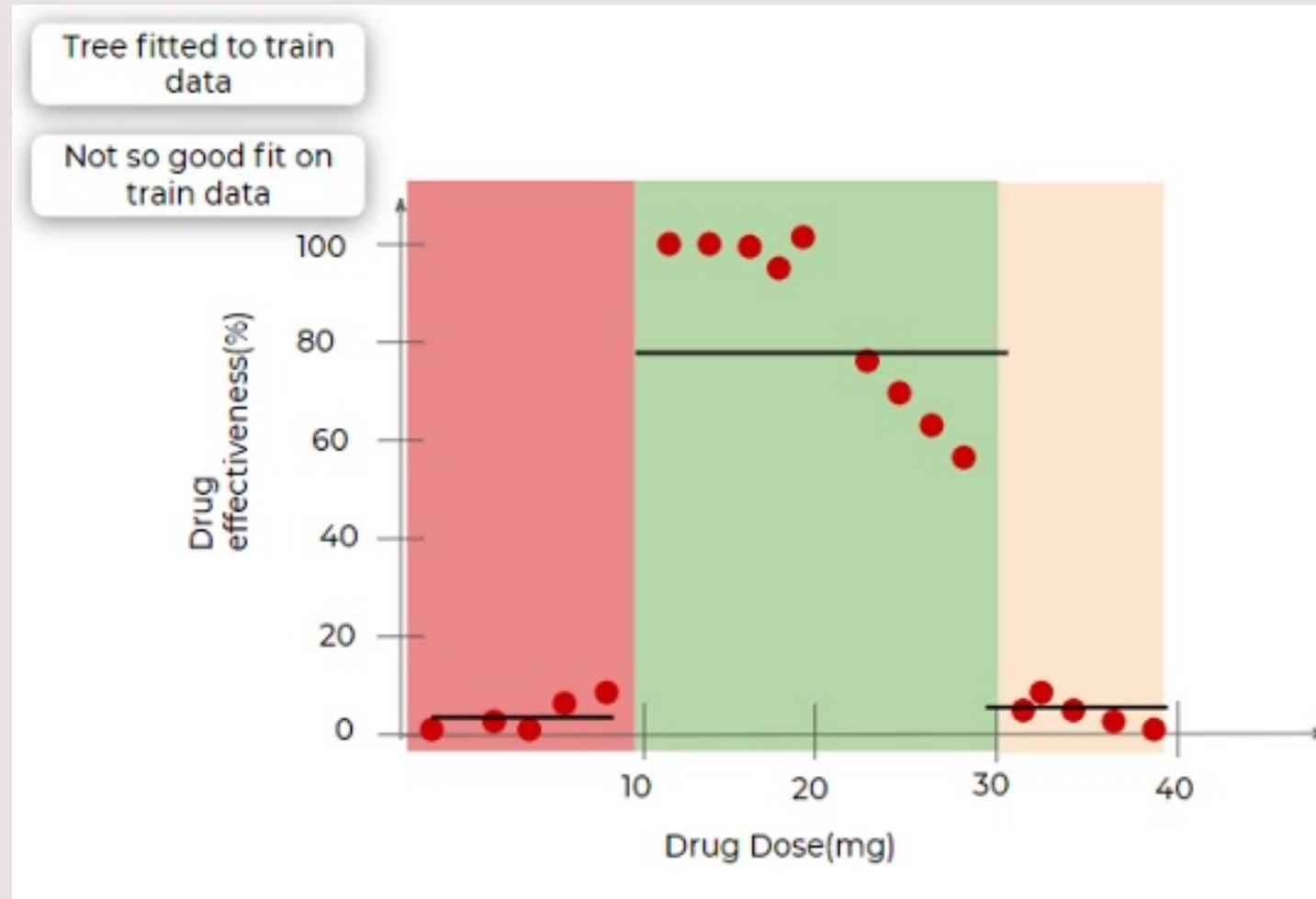
Overfit

Setting Constraints on tree size

- **Minimum samples for a node split**
- **Minimum samples for a leaf node**
- **Maximum depth of the tree (vertical depth)**
- **Maximum number of leaf nodes**
- **Maximum features to consider for a split**
 - The number of features to consider while searching for the best split. These will be randomly selected.
 - As a thumb-rule, the square root of the total number of features works great but we should check up to 30–40% of the total number of features.
 - Higher values can lead to over-fitting but depend on case to case.

Overfit

Pruning



Nấu Mì