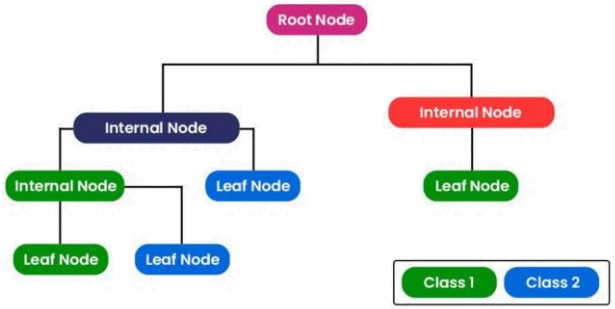
Ch6. Apprentissage Supervisé – Algorithme arbre de décision

Partie 1: Introduction et concepts de base

- ✓ Définition et principes des arbres de décision
- ✓ Applications en classification et régression
- ✓ Algorithme CART et construction d'arbres
- ✓ Critères de division: Gini et Entropie



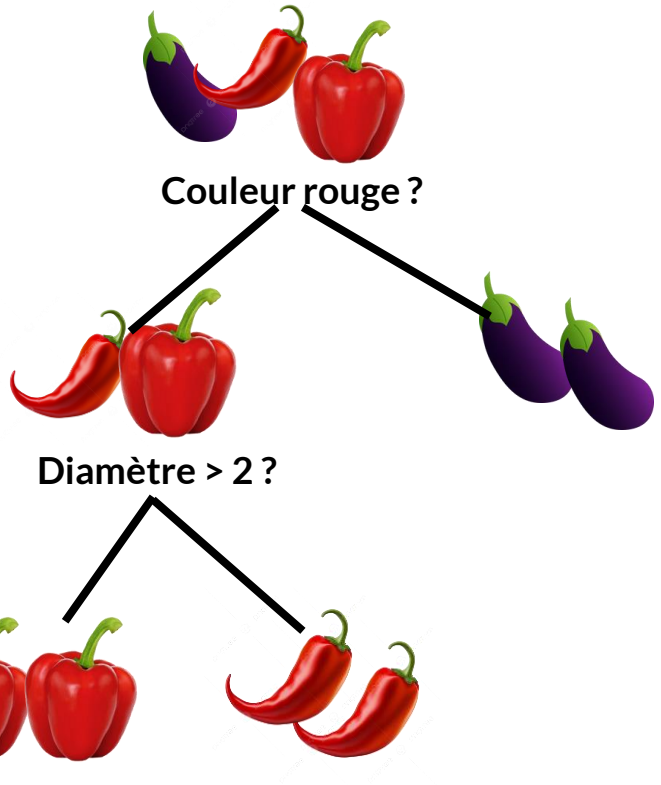
Partie 2: Implémentation pratique

- ✓ Implémentation en Python avec scikit-learn
- ✓ Visualisation et interprétation des arbres
- ✓ Évaluation des performances
- ✓ Hyperparamètres essentiels

Introduction et Définition

Un arbre de décision est un modèle hiérarchique de décisions séquentielles qui divise récursivement les données selon des règles simples de type « si-alors ».

Ce modèle non-paramétrique fonctionne comme un système de questions successives pour aboutir à une prédiction finale.



Fonctionnement Intuitif

Similaire à un organigramme de décision humain. Chaque nœud interne teste un attribut, chaque branche représente le résultat du test, et chaque feuille contient une prédiction (classe ou valeur).

La structure hiérarchique des arbres de décision reflète naturellement le raisonnement humain, ce qui les rend particulièrement interprétables comparés à d'autres modèles de machine learning.

Importance de cet algorithme

✓ Interprétabilité

Les arbres de décision sont visuellement compréhensibles, offrant une représentation graphique intuitive des décisions du modèle, ce qui les rend accessibles aux non-experts.

✓ Prise de décision transparenteté

Permettent de suivre logiquement le cheminement d'une décision depuis la racine jusqu'aux feuilles, ce qui facilite la compréhension des facteurs qui influencent une prédiction particulière.

✓ Adaptabilité à différents types de problèmes

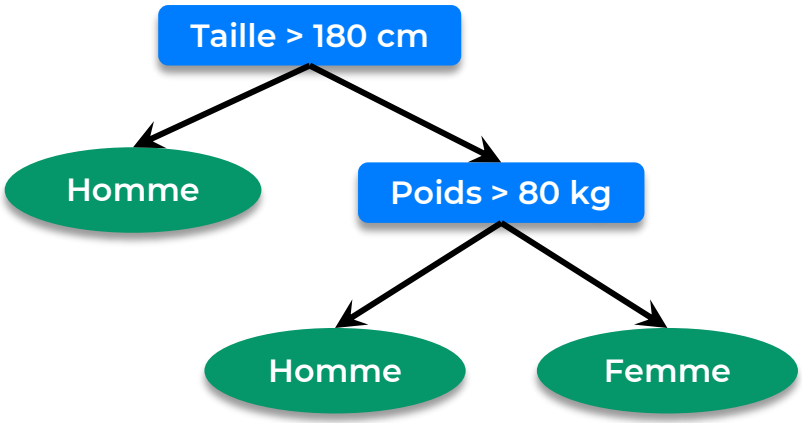
Les arbres de décision peuvent être utilisés pour des problèmes de classification et de régression, ce qui les rend adaptés à un large éventail de tâches d'apprentissage automatique.

✓ Gestion des données mixtes

Peuvent gérer à la fois des caractéristiques numériques et catégorielles sans nécessiter une transformation préalable des données, ce qui simplifie le processus d'ingénierie des caractéristiques.

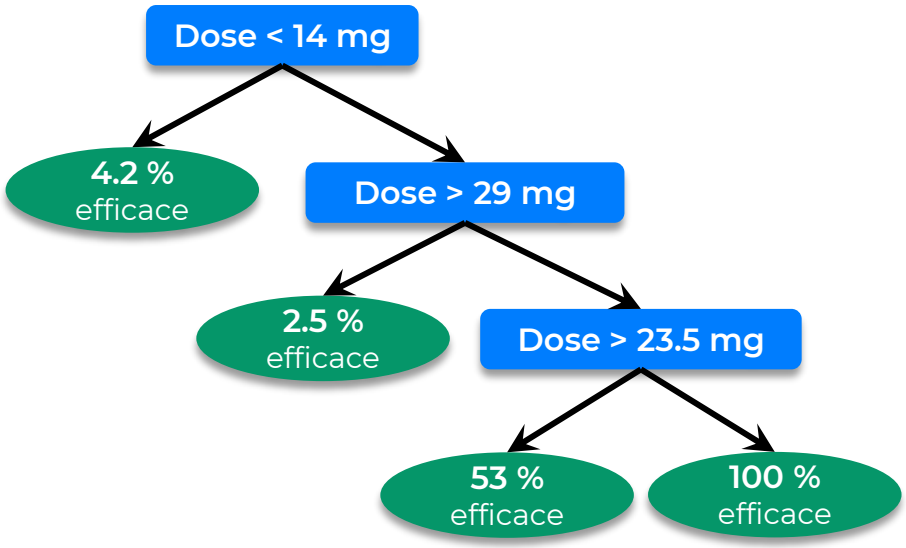
Classification vs Régression

- Lorsqu'un arbre de décision classe les éléments en catégories...



... cela s'appelle un arbre de classification.

- Et lorsqu'un arbre de décision prédit des valeurs numériques...

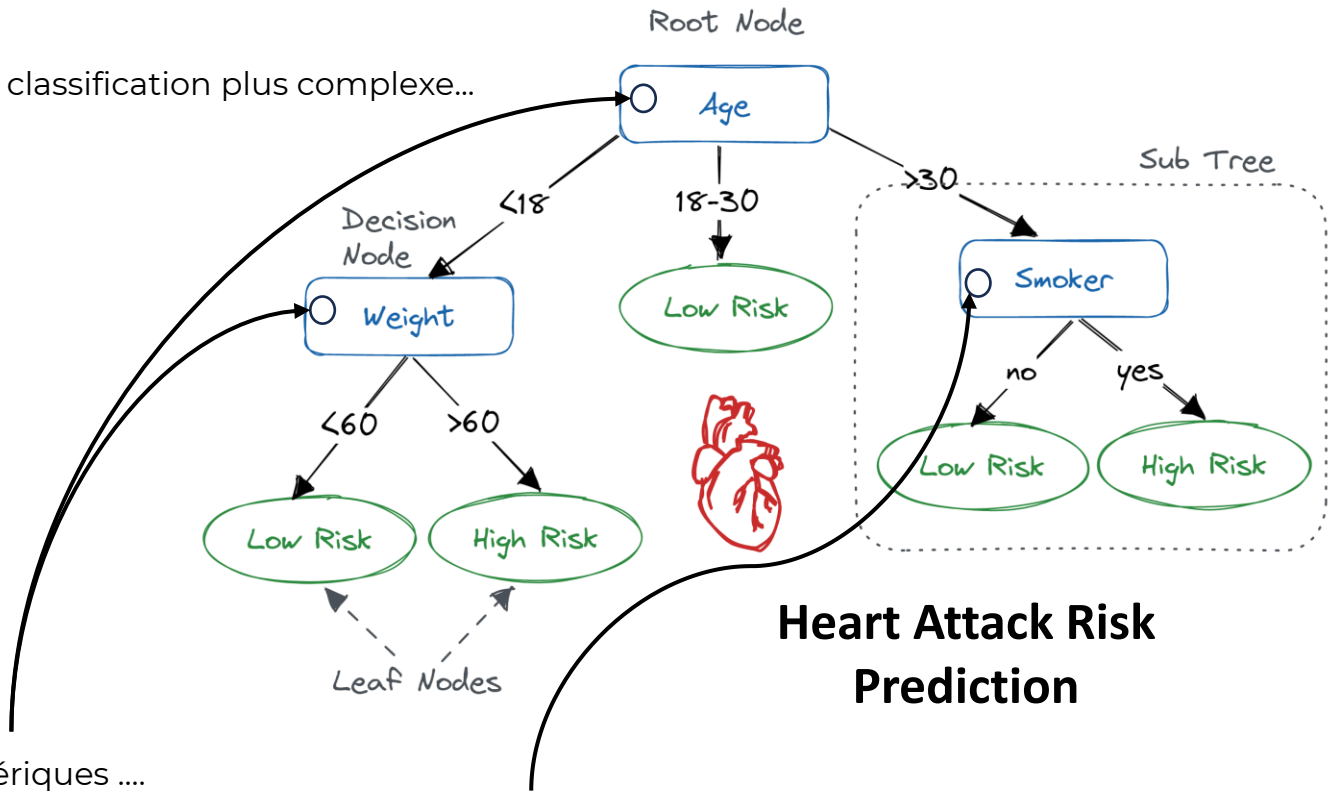


...on parle alors d'arbre de régression.

NOTE : Pour le reste de ce chapitre, nous allons nous concentrer sur les arbres de classification...

Arbre de Classification

Voici maintenant un arbre de classification plus complexe...



Il combine des données numériques ....

...avec des données oui/non.

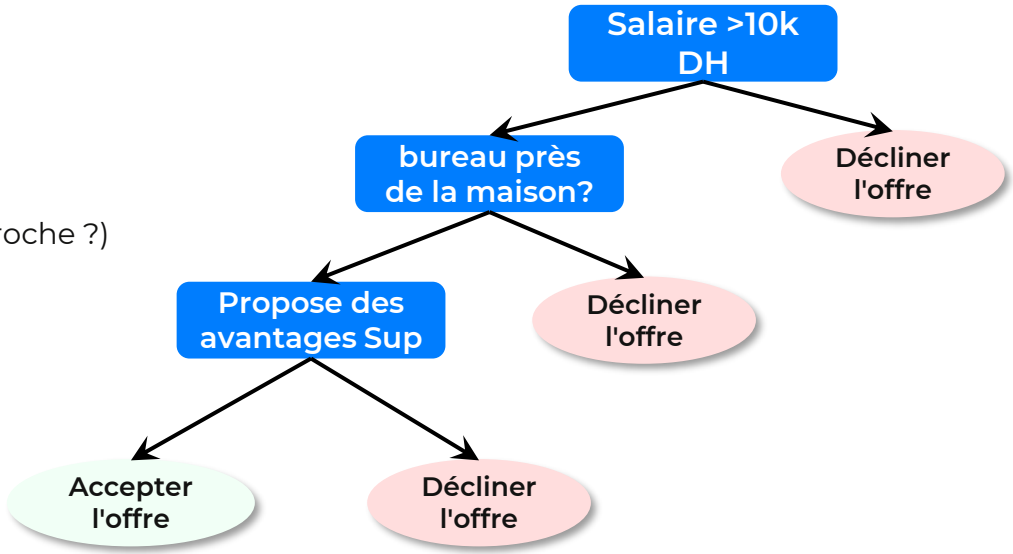
- Il est donc possible de mélanger différents types de données dans la même arborescence
- les classements finaux peuvent être répétés.

Fonctionnement des arbres de décision

SCÉNARIO : Accepter Une Offre D'emploi ?

Un candidat doit décider s'il accepte une offre d'emploi en fonction de trois critères principaux : le salaire, la distance bureau-domicile et les avantages proposés

1. **Racine**
   Vérification du salaire (≥ 10 000 DH ?)
2. **Si Oui**
   Vérification de la distance bureau-domicile (proche ?)
3. **Si Non**
   Décliner l'offre
4. **Feuilles**
   Décision finale (Accepter/Décliner l'offre)



En général, un arbre de décision fait une affirmation...
... puis prend une décision en fonction du fait que l'affirmation est « vraie » ou « fausse ».

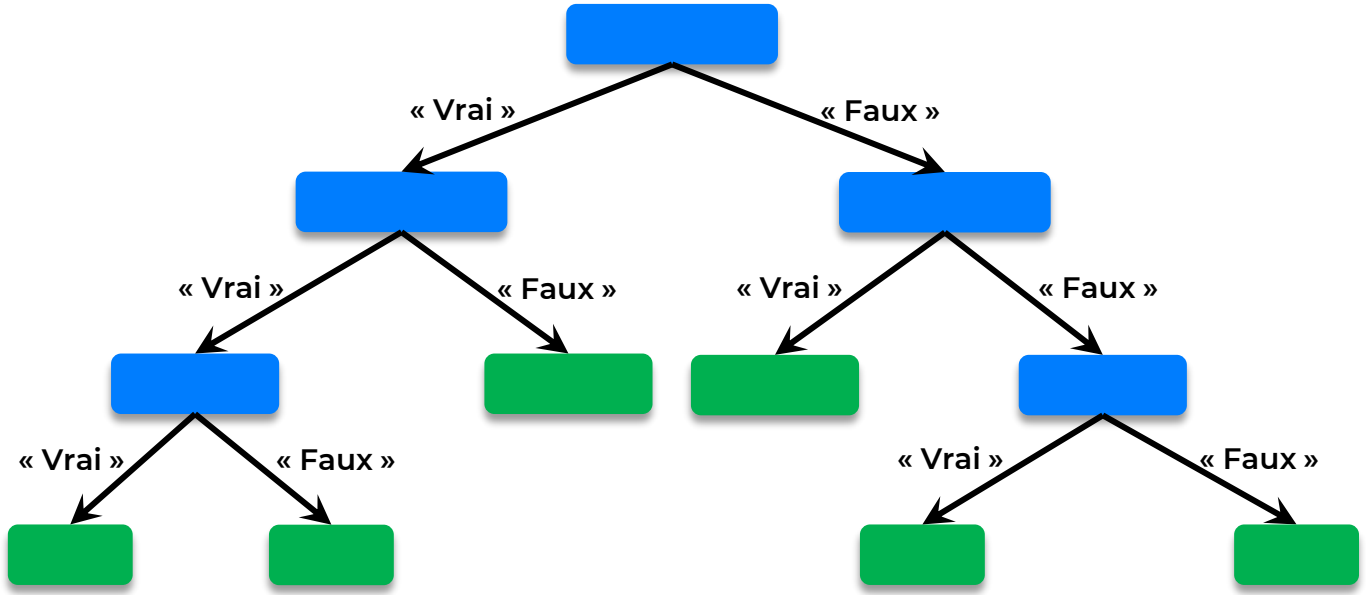
Fonctionnement des arbres de décision

Dans l'ensemble, les arbres de classification sont assez faciles à utiliser.

Vous commencez par le haut

... et vous descendez ...

jusqu'à ce que vous arriviez à un point où vous ne pouvez plus aller plus loin, et c'est ainsi que vous classerez quelque chose....



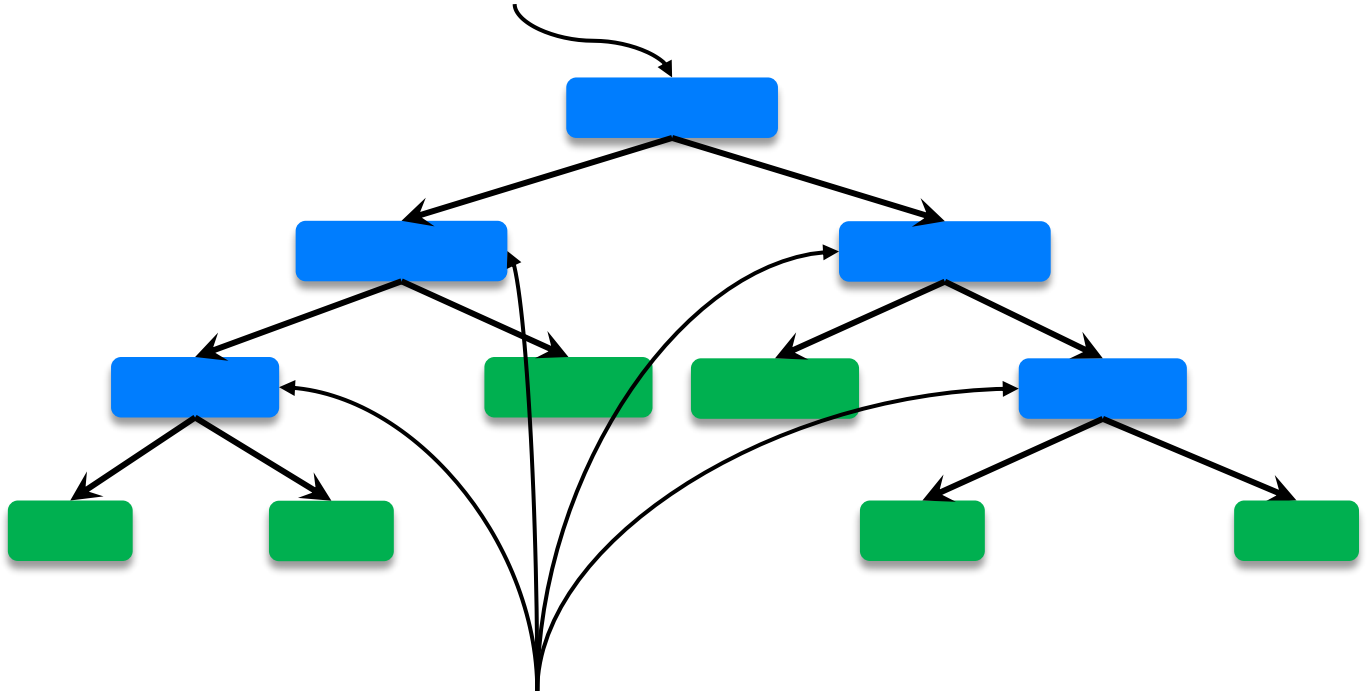
NOTE : Jusqu'à présent, les flèches ont été étiquetées « Vrai » ou « Faux »...

... mais généralement, on part du principe que si une affirmation est vraie, on va à gauche...

...et à droite, si une affirmation est fausse

Fonctionnement des arbres de décision

Le nœud le plus haut de l'arbre est appelé **nœud racine** (Root Node) ou « **racine** » (The Root).



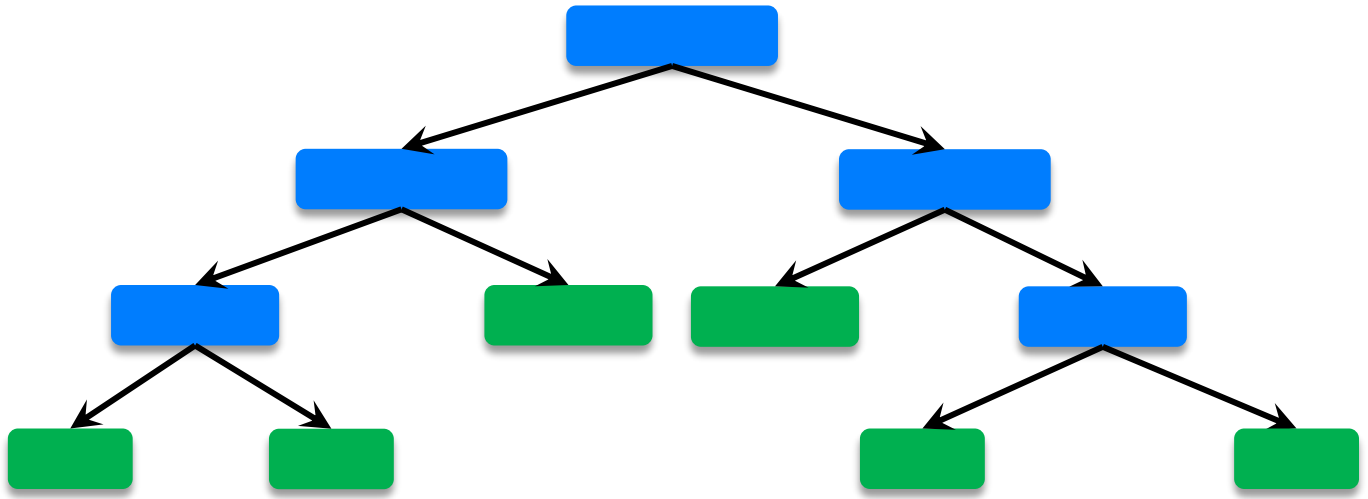
Ces nœuds sont appelés **nœuds internes** « Internal Nodes » ou **branches** (Branches).

Les branches ont des flèches qui pointent vers elles...

... et ils ont des flèches qui pointent vers l'extérieur.

Fonctionnement des arbres de décision

Enfin, ceux-ci sont appelés **nœuds feuilles** (Leaf Nodes) ou simplement **feuilles** (Leaves).



Les feuilles ont des flèches qui pointent vers elles...

...mais aucune flèche ne pointe en dehors d'elles.

Fonctionnement des arbres de décision

Maintenant que nous savons comment utiliser et interpréter les arbres de classification...

...apprenons à construire un arbre à partir de données brutes.

Tanjia	Humour	Age	Marrakchi
Yes	Yes	20	Yes
Yes	No	25	No
No	Yes	28	Yes
Yes	Yes	32	Yes
No	Yes	35	Yes
Yes	No	45	No
No	No	55	No
Yes	Yes	58	No

Fonctionnement des arbres de décision

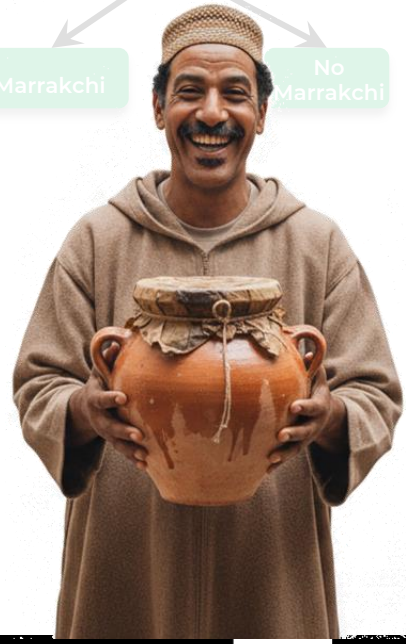
Ces données nous indiquent si quelqu'un aime préparer ou non le Tanjia, un plat traditionnel de Marrakech…

Tanjia	Humour	Age	Marrakchi
Yes	Yes	20	Yes
Yes	No	25	No
No	Yes	28	Yes
Yes	Yes	32	Yes
No	Yes	35	Yes
Yes	No	45	No
No	No	55	No
Yes	Yes	58	No

… si la personne a un sens de l'humour, ce qui est important dans la culture marrakchie…

… leur âge …

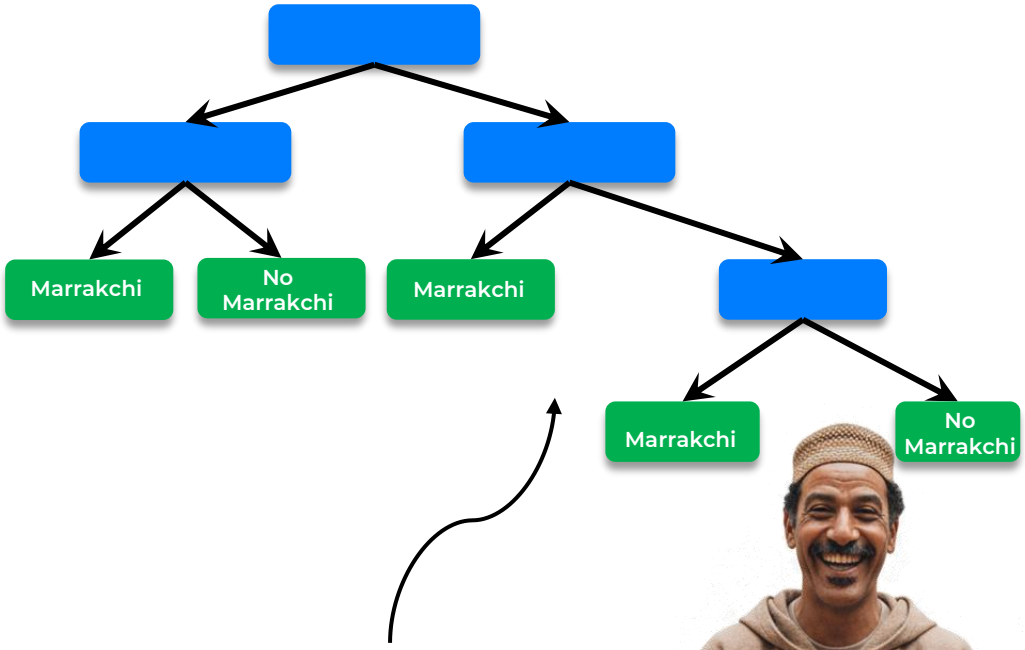
…et s'il est ou non un vrai Marrakchi.



Fonctionnement des arbres de décision

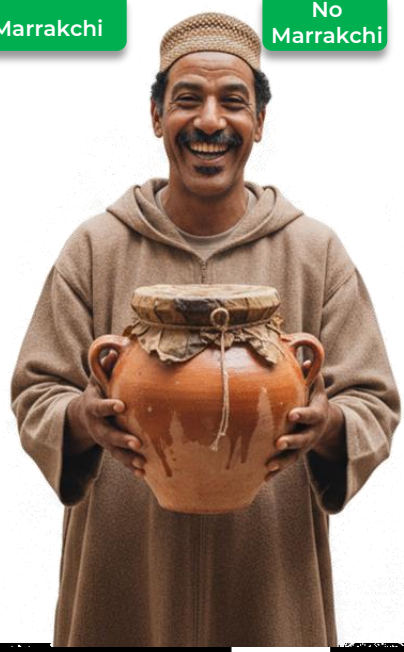
Tanjia	Humour	Age	Marrakchi
Yes	Yes	20	Yes
Yes	No	25	No
No	Yes	28	Yes
Yes	Yes	32	Yes
No	Yes	35	Yes
Yes	No	45	No
No	No	55	No
Yes	Yes	58	No

Nous allons donc utiliser ces données...



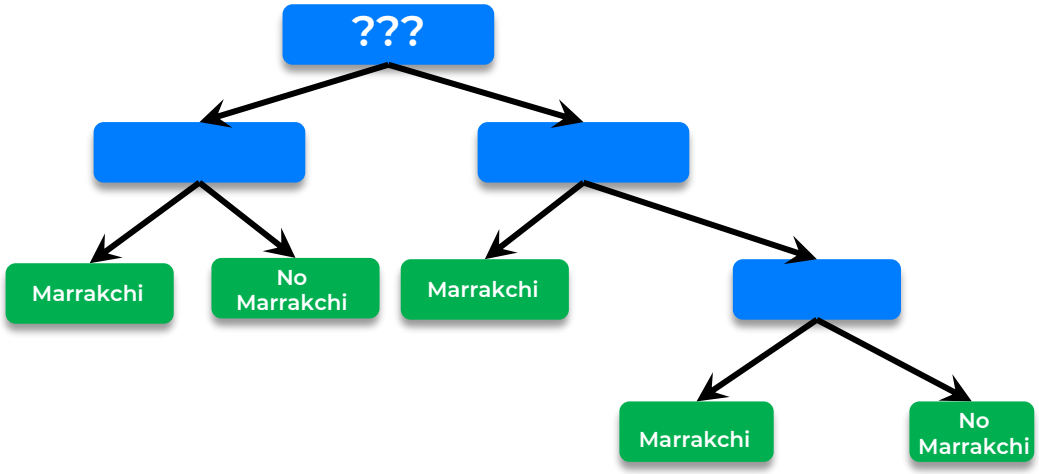
...pour construire cet arbre de classification...

... qui permet de déterminer si quelqu'un est un vrai Marrakchi ou non.



Fonctionnement des arbres de décision

Tanjia	Humour	Age	Marrakchi
Yes	Yes	20	Yes
Yes	No	25	No
No	Yes	28	Yes
Yes	Yes	32	Yes
No	Yes	35	Yes
Yes	No	45	No
No	No	55	No
Yes	Yes	58	No



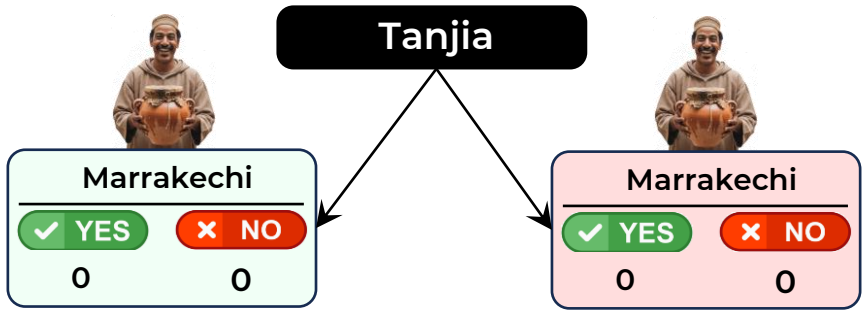
Maintenant, la question qui se pose ...

... est comment construire un arbre à partir de données seules ??

La 1ère chose à faire est de décider si le fait d'aimer préparer le **Tanjia**, d'avoir le sens de **l'humour** ou **l'âge** devrait être la question que nous posons en premier lieu.

Fonctionnement des arbres de décision

Tanjia	Humour	Age	Marrakchi
Yes	Yes	20	Yes
Yes	No	25	No
No	Yes	28	Yes
Yes	Yes	32	Yes
No	Yes	35	Yes
Yes	No	45	No
No	No	55	No
Yes	Yes	58	No



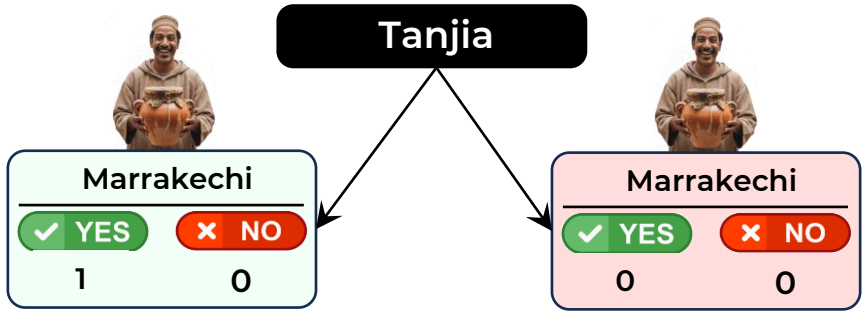
Pour prendre cette décision, nous commencerons par examiner dans quelle mesure le fait d'apprécier la préparation du Tanjia permet de prédire si une personne est un véritable Marrakechi.

Pour ce faire, nous allons créer un arbre très simple qui demande uniquement si quelqu'un aime préparer le Tanjia...

... puis nous faisons descendre les données dans l'arborescence.

Fonctionnement des arbres de décision

Tanjia	Humour	Age	Marrakchi
Yes	Yes	20	Yes
Yes	No	25	No
No	Yes	28	Yes
Yes	Yes	32	Yes
No	Yes	35	Yes
Yes	No	45	No
No	No	55	No
Yes	Yes	58	No



Par exemple, la première personne dans l'ensemble de données aime préparer le Tanjia...

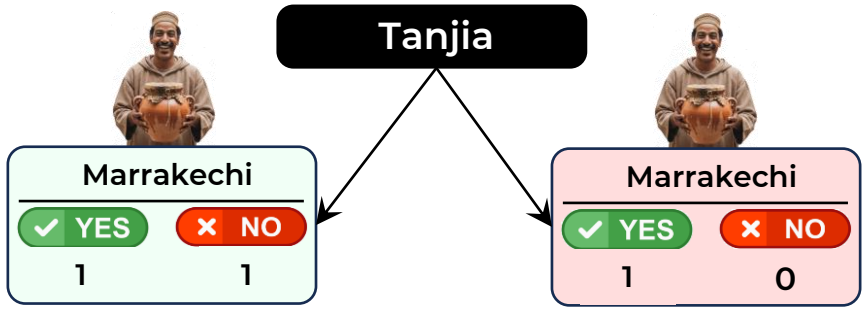
... alors ils se dirigent vers la feuille à gauche.

Et parce qu'il est un vrai Marrakechi...

...nous allons noter ça en mettant un 1 sous le mot Oui.

Fonctionnement des arbres de décision

Tanjia	Humour	Age	Marrakchi
Yes	Yes	20	Yes
Yes	No	25	No
No	Yes	28	Yes
Yes	Yes	32	Yes
No	Yes	35	Yes
Yes	No	45	No
No	No	55	No
Yes	Yes	58	No



La 2éme personne dans l'ensemble de données aime également préparer le Tanjia...

... donc ils vont aussi vers la feuille à gauche.

Et puisque ce n'est pas un vrai Marrakechi, nous mettons 1 sous « Non ».

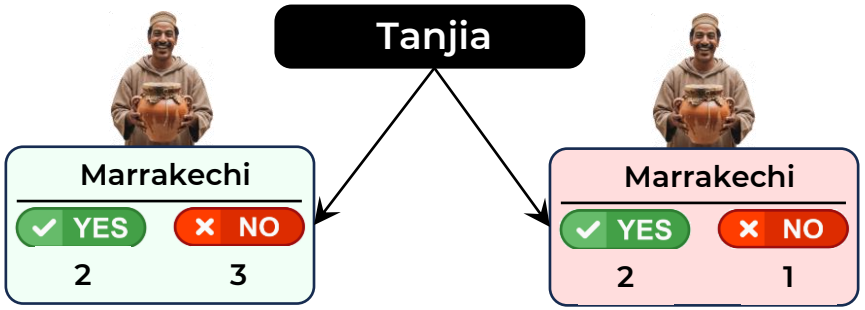
La troisième personne n'aime pas préparer le Tanjia...

... alors ils vont vers la feuille à droite.

Et parce qu'il est un vrai Marrakechi, nous allons mis un 1 sous le mot « Oui ».

Fonctionnement des arbres de décision

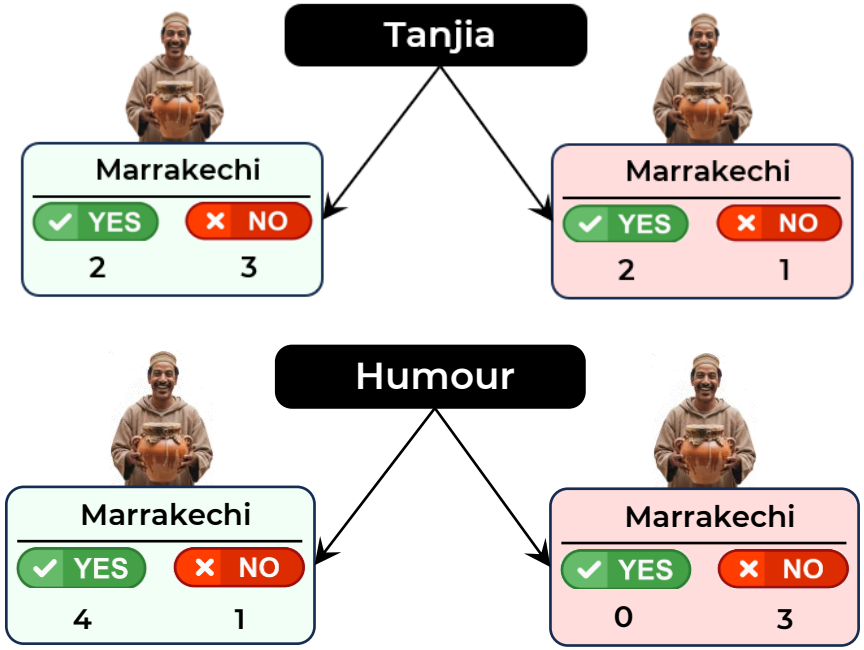
Tanjia	Humour	Age	Marrakchi
Yes	Yes	20	Yes
Yes	No	25	No
No	Yes	28	Yes
Yes	Yes	32	Yes
No	Yes	35	Yes
Yes	No	45	No
No	No	55	No
Yes	Yes	58	No



De même, nous parcourons les lignes restantes en suivant l'arborescence, en vérifiant si chacune d'entre elles correspond ou non à un vrai marrakechi.

Fonctionnement des arbres de décision

Tanjia	Humour	Age	Marrakchi
Yes	Yes	20	Yes
Yes	No	25	No
No	Yes	28	Yes
Yes	Yes	32	Yes
No	Yes	35	Yes
Yes	No	45	No
No	No	55	No
Yes	Yes	58	No

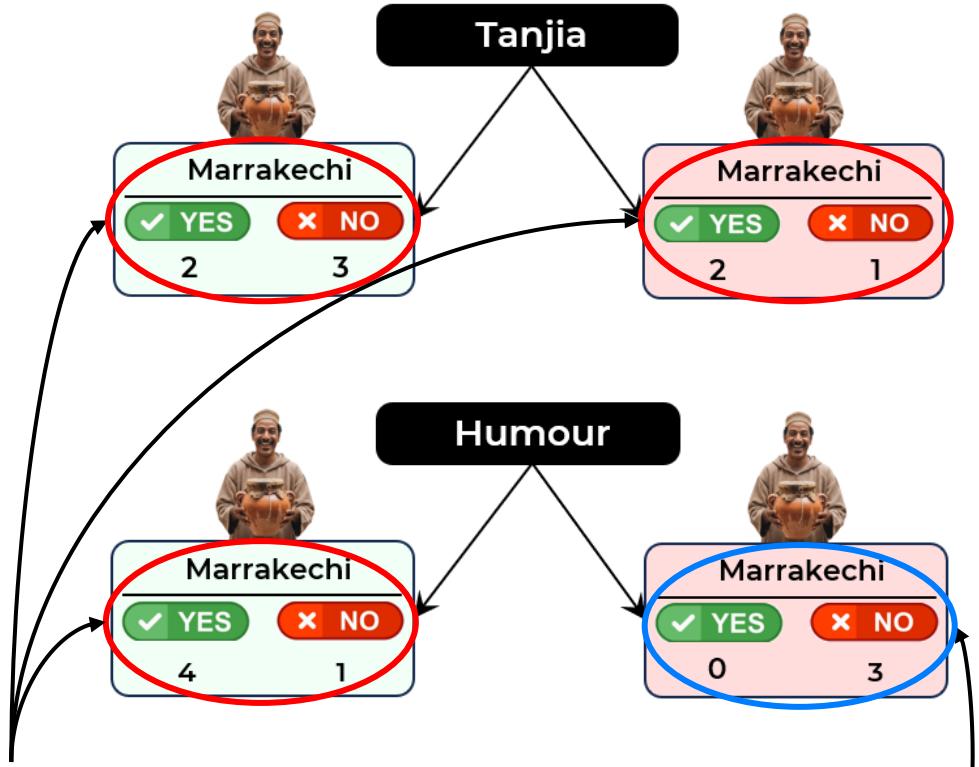


Maintenant, faisons exactement la même chose pour le sens de l'humour.

La première personne a le sens de l'humour et c'est un vrai Marrakchi.

Fonctionnement des arbres de décision

Tanjia	Humour	Age	Marrakchi
Yes	Yes	20	Yes
Yes	No	25	No
No	Yes	28	Yes
Yes	Yes	32	Yes
No	Yes	35	Yes
Yes	No	45	No
No	No	55	No
Yes	Yes	58	No



En analysant les deux petits arbres, nous voyons que les deux ne sont pas parfaitement efficaces pour prédire si quelqu'un est ou non un vrai Marrakechi.

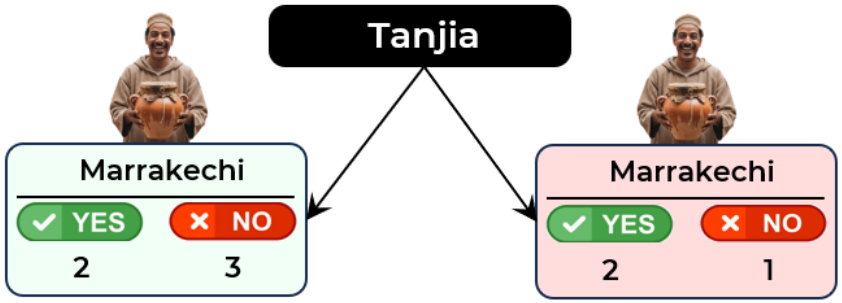
Plus précisément, ces trois Leaves regroupent des personnes qui sont et qui ne sont pas de vrais Marrakechis.

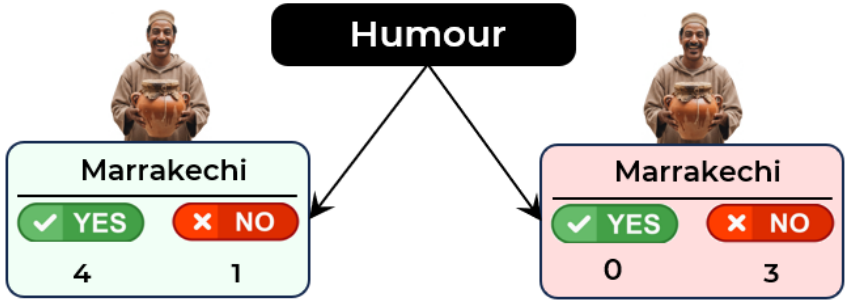
Par contre, cette feuille ne contient que des personnes qui ne sont pas de vrais Marrakchis.

Comme ces 3 feuilles regroupent à la fois des personnes qui sont et ne sont pas de vrais marrakchis...

...on les appelle les **Impures**.

Fonctionnement des arbres de décision





Comme les deux feuilles de l'arbre « **Tanjia** » sont **impures**...

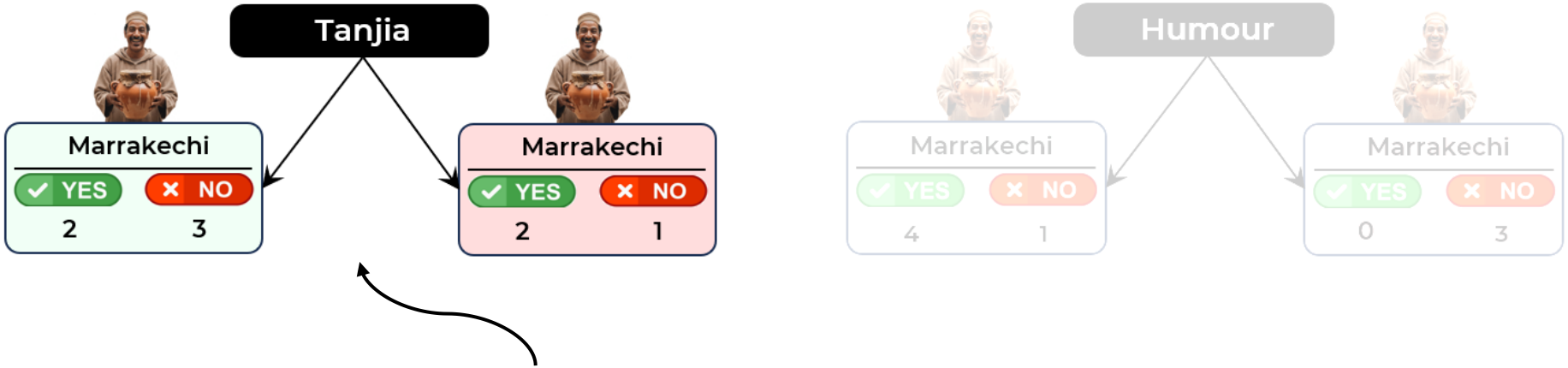
... et qu'une seule feuille de l'arbre « **humour** » **n'est pas impure**...

... il semble que le sens de l'humour soit un meilleur indicateur pour déterminer qui est ou n'est pas un vrai Marrakechi...

... mais il serait intéressant de pouvoir quantifier les différences entre « Tanjia » et « Humour ».

Il existe plusieurs façons de quantifier l'impureté des feuilles.
L'une des méthodes les plus populaires est appelée « **impureté de Gini** », bien qu'il existe également des méthodes aux telles que « **entropie** » et « **gain d'information** ».
D'un point de vue numérique, ces méthodes sont toutes assez similaires*, nous nous concentrerons donc sur l'impureté de Gini, car non seulement elle est très populaire, mais elle est la plus simple.

Fonctionnement des arbres de décision



Commençons donc par calculer **l'impureté de Gini** pour l'arbre « **Tanjia** ».

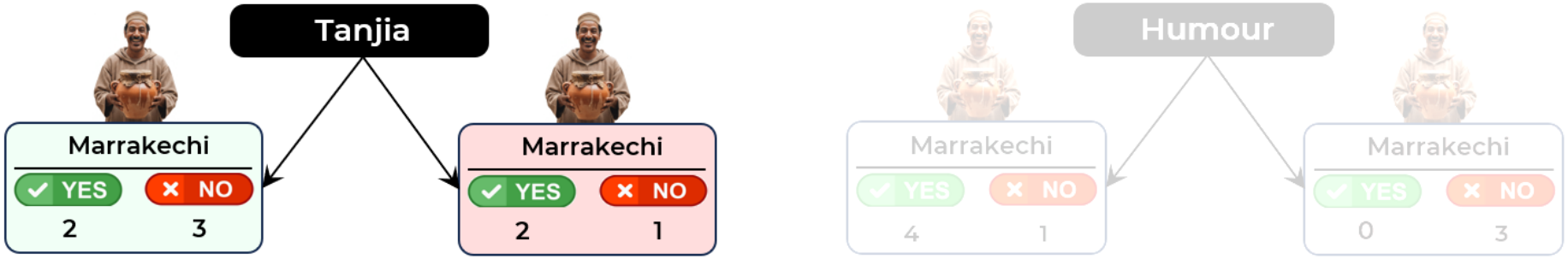
Pour calculer l'impureté de Gini pour « Tanjia », nous commençons par calculer l'impureté de Gini pour les feuilles individuelles

L'impureté de Gini (Imp_{GINI}) pour la feuille de gauche est...

$$Imp_{GINI} = 1 - \sum P_i^2$$ Où P_i est la probabilité d'occurrence de chaque classe

L'impureté de Gini pour une feuille = 1 - (probabilité de « Oui »)² - (probabilité de « Non »)²

Fonctionnement des arbres de décision



$$IG_g = 1 - \sum P_i^2$$

Nous commençons donc par 1...

$$= \boxed{1} - \boxed{\left(\frac{2}{2+3}\right)^2} - \boxed{\left(\frac{3}{2+3}\right)^2} = 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = 0.48$$

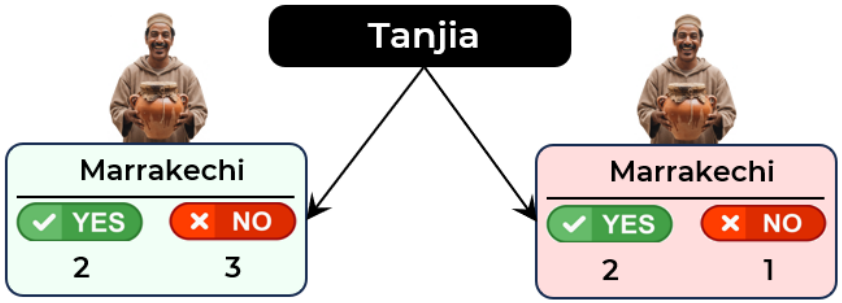
... puis nous retirons la probabilité au carré que quelqu'un dans cette feuille adore préparer le Tanjia...

... qui correspond à 2, divisé par le nombre total de personnes dans la feuille, soit 5, au carré.

Enfin, nous déduisons la probabilité au carré qu'une personne de cette feuille n'aime pas préparer le Tanjia...

...qui est égale à 3, divisé par le nombre total de personnes de la feuille (5) au carré.

Fonctionnement des arbres de décision



$IG_g = 0.48$

$IG_d \approx 0.444$

De même pour la feuille à droite

$$IG_d = 1 - \sum P_i^2$$

$$= 1 - \left(\frac{2}{2+1}\right)^2 - \left(\frac{1}{2+1}\right)^2 = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 \approx 0.444$$

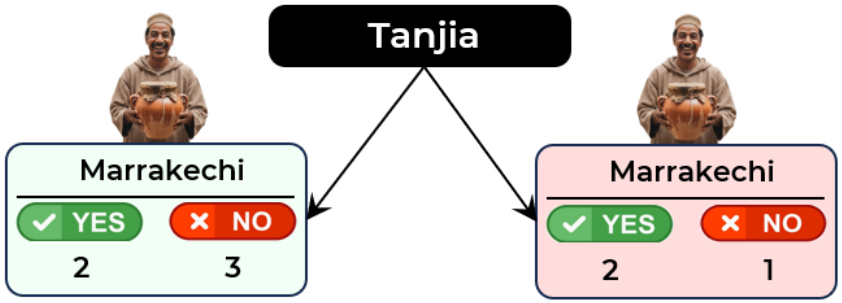
Maintenant, comme la feuille à gauche contient 5 personnes...

...et celle à droite seulement 3...

...les feuilles ne représentent pas le même nombre de personnes.

➔ Par conséquent, **l'impureté totale de Gini est la moyenne pondérée des impuretés des feuilles**.

Fonctionnement des arbres de décision



$IG_g = 0.48$

$IG_d \approx 0.444$

Impureté totale de Gini = moyenne pondérée des impuretés de Gini pour les feuilles

$$IG_{Total} = \left(\frac{5}{8}\right) \times 0.48 + \left(\frac{3}{8}\right) \times 0.444 = 0.4665$$

qui correspond au nombre total de personnes dans chaque feuille, la première 5 et la deuxième 3...

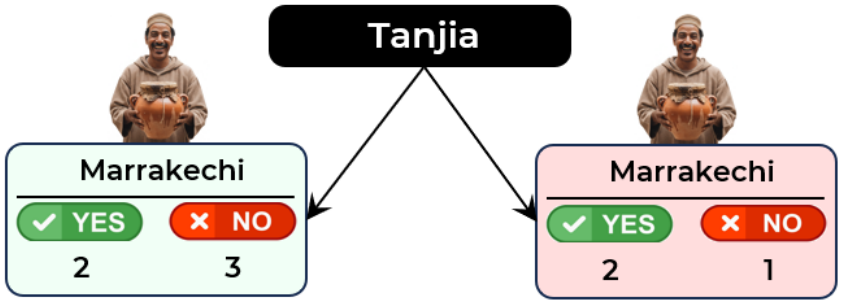
... divisé par le nombre total de personnes dans les deux feuilles, soit 8.

multiplié par l'impureté de Gini associée, respectivement 0,48 et 0,444.

Donc l'impureté de Gini pour « **Tanjia** » est de **0,4665**.

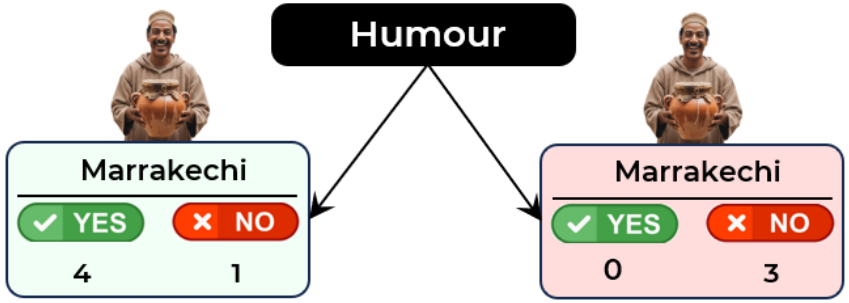
Fonctionnement des arbres de décision

$IG_{Total} = 0.4665$



$IG_g = 0.48$ $IG_d \approx 0.444$

$IG_{Total} = 0.2$



$IG_g = 0.32$ $IG_g = 0$

$$IG_g = 1 - \left(\frac{4}{4+1}\right)^2 - \left(\frac{1}{4+1}\right)^2 = 1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = 0.32$$

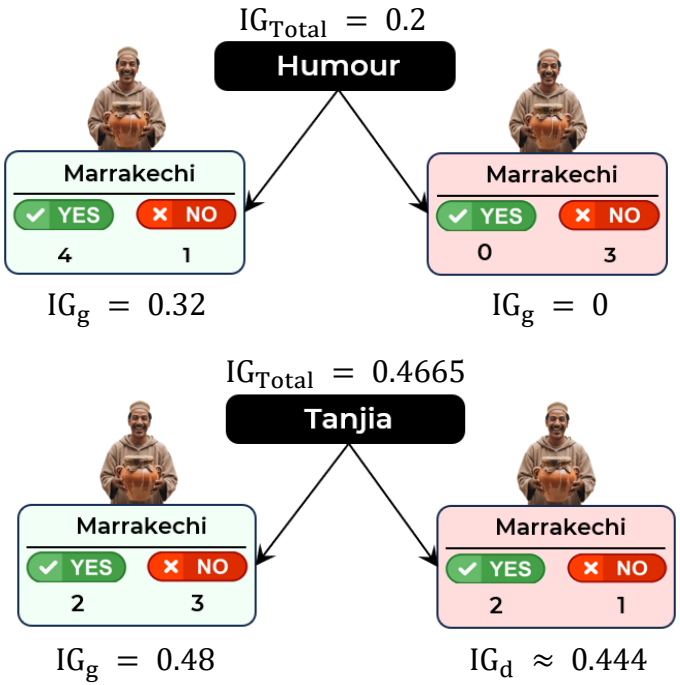
$$IG_d = 1 - \left(\frac{0}{0+3}\right)^2 - \left(\frac{3}{0+3}\right)^2 = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 0$$

$$IG_{Total} = \left(\frac{5}{8}\right) \times 0.32 + \left(\frac{3}{8}\right) \times 0 = 0.2$$

Fonctionnement des arbres de décision

Tanjia	Humour	Age	Marrakechi
Yes	Yes	20	Yes
Yes	No	25	No
No	Yes	28	Yes
Yes	Yes	32	Yes
No	Yes	35	Yes
Yes	No	45	No
No	No	55	No
Yes	Yes	58	No

22.5 | 26.5 | 30 | 33.5 | 40 | 50 | 56.5



Nous devons maintenant calculer l'impureté de Gini pour l'âge...

...Cependant, vu que l'âge contient des données numériques, et pas seulement des valeurs oui/non, le calcul de l'impureté de Gini est un peu plus compliqué

La première chose à faire est de trier les lignes par âge, de la valeur la plus basse à la valeur la plus élevée (dans ce cas, nos données sont déjà triées)

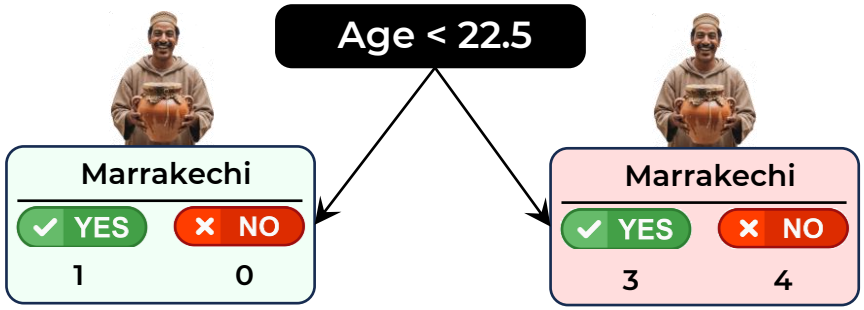
Ensuite, nous calculons l'âge moyen de toutes les personnes adjacentes

Enfin, nous calculons les valeurs de « Gini Impurity » pour chaque âge moyen.

Fonctionnement des arbres de décision

Tanjia	Humour	Age	Marrakechi
Yes	Yes	20	Yes
Yes	No	25	No
No	Yes	28	Yes
Yes	Yes	32	Yes
No	Yes	35	Yes
Yes	No	45	No
No	No	55	No
Yes	Yes	58	No

22.5 | 26.5 | 30 | 33.5 | 40 | 50 | 56.5



Par exemple, pour calculer l'impureté de Gini pour la première valeur...

...nous mettons Âge < 22,5 dans la racine...

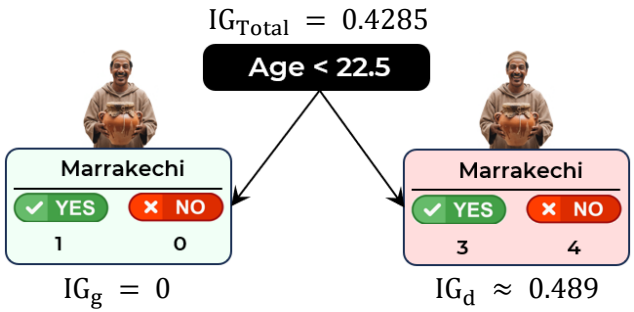
...et comme la seule personne dont l'âge est inférieur à 22,5 ans est un vrai Marrakechi...

...nous mettons un 0 sous « Non » et un 1 sous « Oui ».

Ensuite, toutes les personnes dont l'âge est supérieur ou égal à 22,5 ans vont dans la feuille de droite.

Fonctionnement des arbres de décision

Tanjia	Humour	Age	Marrakechi
Yes	Yes	20	Yes
Yes	No	22.5	No
No	Yes	25	Yes
Yes	Yes	26.5	Yes
No	Yes	28	Yes
Yes	Yes	30	Yes
No	Yes	32	Yes
Yes	No	33.5	No
No	No	35	No
Yes	Yes	40	No
		45	
		50	
		55	
		56.5	
		58	



Nous calculons maintenant l'impureté de « Age < 22.5 » ...

$$IG_g = 1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2 = 1 - 1 = 0$$

Et cela est logique, car chaque personne dans cette feuille est un véritable Marrakéchi...

...il n'y a donc aucune impureté.

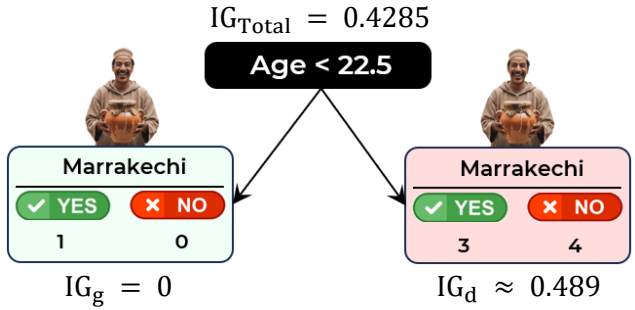
$$IG_d = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 \approx 0.4897$$

$$IG_{Total} = \left(\frac{1}{8}\right) \times 0 + \left(\frac{7}{8}\right) \times 0.4897 \approx 0.4285$$

Fonctionnement des arbres de décision

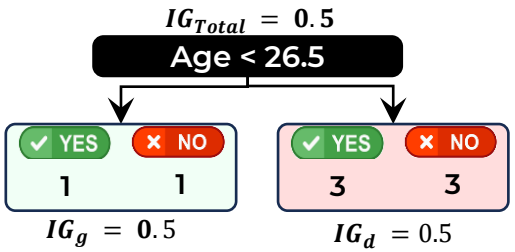
Tanjia	Humour	Age	Marrakechi
Yes	Yes	20	Yes
Yes	No	25	No
No	Yes	28	Yes
Yes	Yes	32	Yes
No	Yes	35	Yes
Yes	No	45	No
No	No	55	No
Yes	Yes	58	No

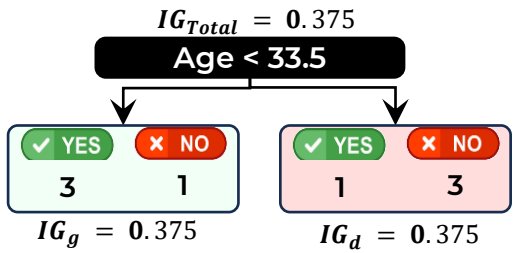
22.5 | 26.5 | 30 | 33.5 | 40 | 50 | 56.5

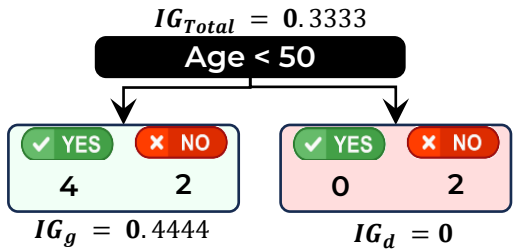


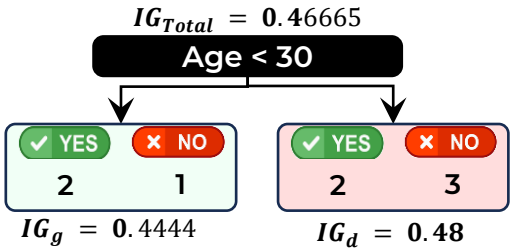
De même pour les autres moyennes d'âge

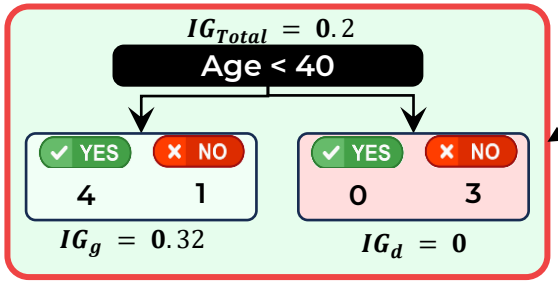
Le seuil de ce candidat, 40, est le plus bas en termes d'impureté, 0,16... ... donc, nous en choisirons 40.

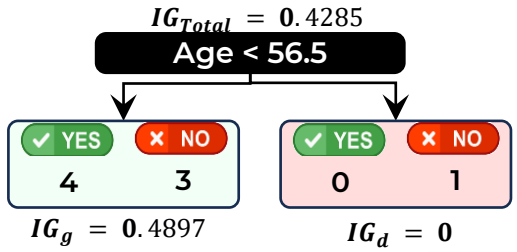




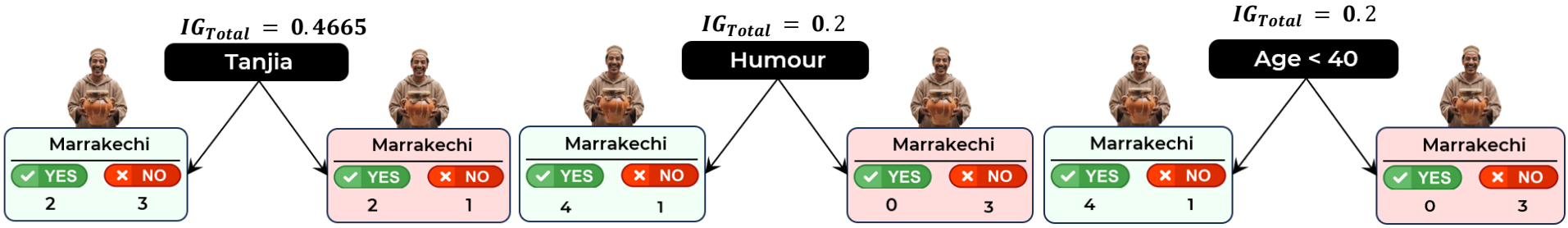








Fonctionnement des arbres de décision



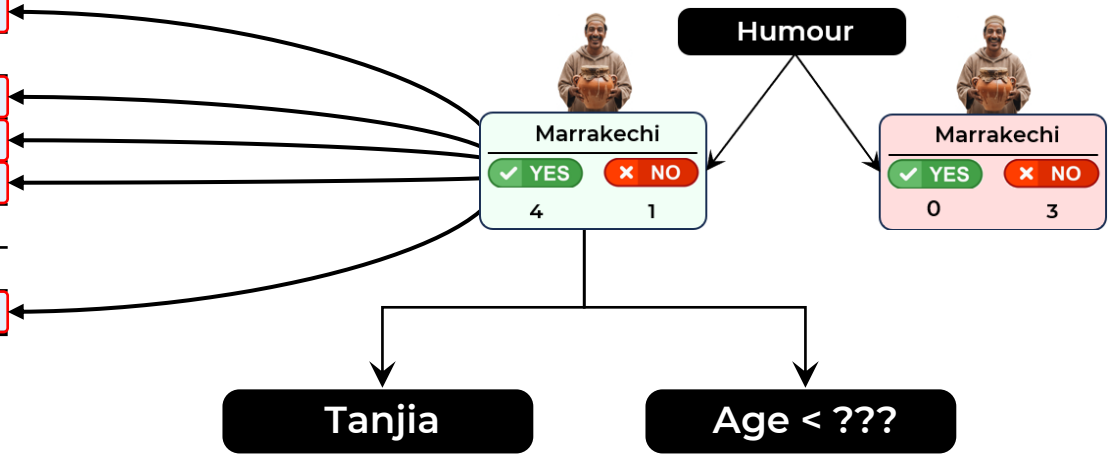
RAPPELEZ-VOUS, nous comparons les valeurs d'impureté de Gini pour « Tanjia », « Humour » et « Age ». Pour décider quelle caractéristique doit se trouver tout en haut de l'arbre

- Et par ce que « Humour » et « Age » ont le Gini d'impureté le plus bas (0.2)
- Leurs feuilles « Leaves » ont le moins d'impuretés

➔Alors nous avons placé le sens de l'humour ou l'âge au sommet de l'arbre (voir le slide suivant)

Fonctionnement des arbres de décision

Tanjia	Humour	Age	Marrakchi
Yes	Yes	20	Yes
Yes	No	25	No
No	Yes	28	Yes
Yes	Yes	32	Yes
No	Yes	35	Yes
Yes	No	45	No
No	No	55	No
Yes	Yes	58	No

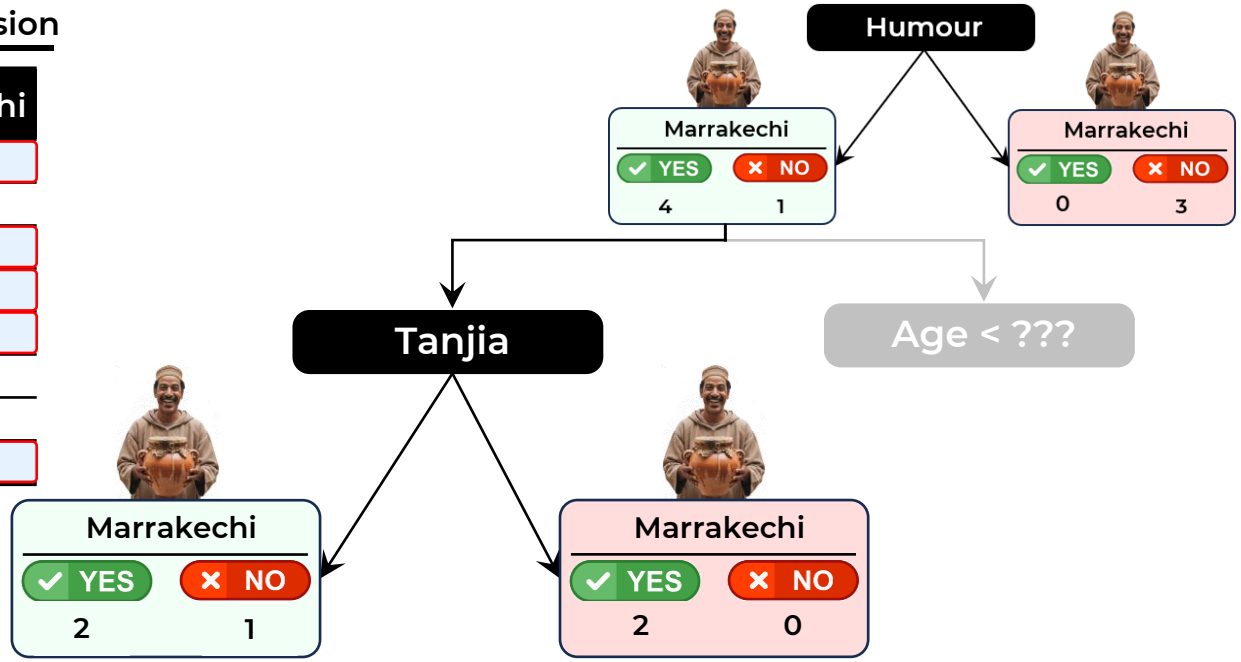


Maintenant, regardons le nœud à gauche (les 5 personnes qui ont le sens de l'humour sont dans ce nœud).
4 d'entre elles sont de vrais Marrakechis et 1 ne l'est pas ➔ Ce nœud est donc **impur**.

Voyons donc si nous pouvons réduire l'impureté en séparant les personnes qui ont le sens de l'humour en fonction de leur enthousiasme à préparer le Tanjia ou de leur âge <=40 ans.

Fonctionnement des arbres de décision

Tanjia	Humour	Age	Marrakchi
Yes	Yes	20	Yes
Yes	No	25	No
No	Yes	28	Yes
Yes	Yes	32	Yes
No	Yes	35	Yes
Yes	No	45	No
No	No	55	No
Yes	Yes	58	No



$$IG_{Total} = 0.2666$$

Nous commencerons par demander aux 5 personnes qui ont le sens de l'humour si elles aiment préparer le Tanjia.

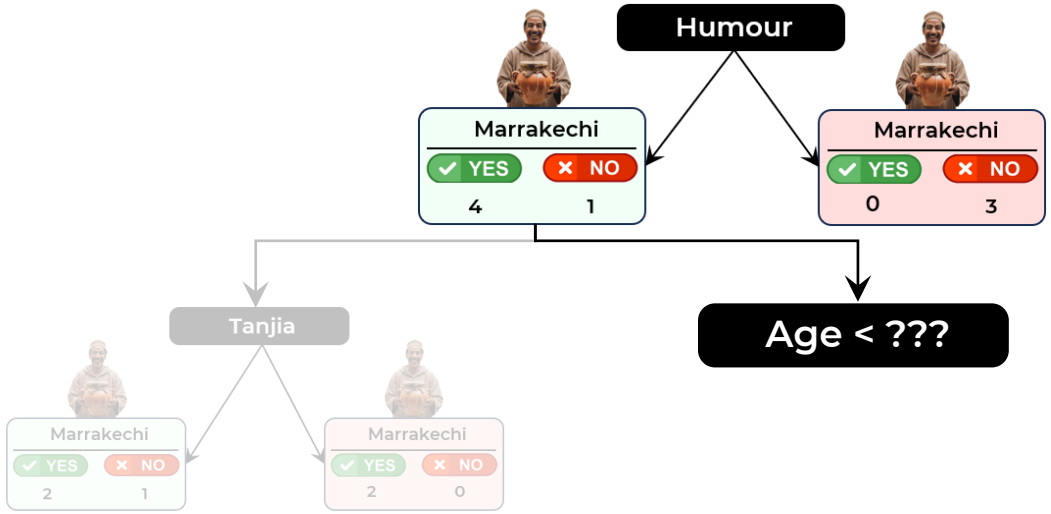
- 3 des 5 personnes qui ont le sens de l'humour aiment également préparer le Tanjia

  ➔ elles se retrouvent dans la feuille de gauche.
- Les 2 autres restants, n'aiment pas préparer le Tanjia, se retrouvent à droite.
- Et l'indice de Gini total pour cette division est de 0,266.

Fonctionnement des arbres de décision

Tanjia	Humour	Age	Marrakchi
Yes	Yes	20	Yes
Yes	No	25	No
No	Yes	28	Yes
Yes	Yes	32	Yes
No	Yes	35	Yes
Yes	No	45	No
No	No	55	No
Yes	Yes	58	No

24 | 30 | 33.5 | 46.5



Nous testons maintenant différents seuils d'âge, comme précédemment...

... mais cette fois-ci, nous ne prenons en compte que l'âge des personnes qui ont le sens de l'humour

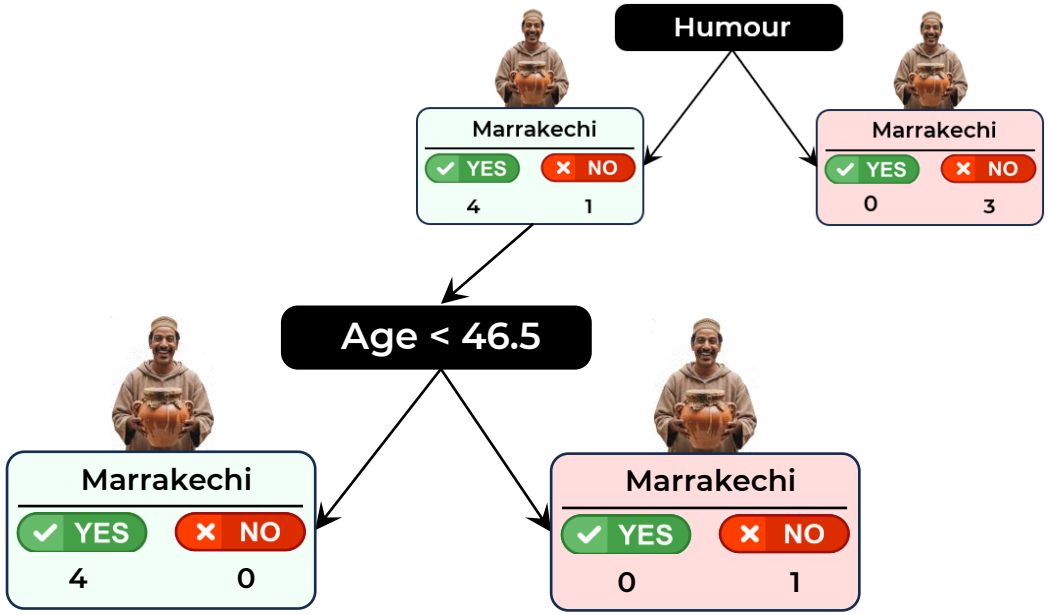
Age < 24				Age < 30				Age < 33.5				Age < 46.5			
✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
1	0	3	1	2	0	2	1	3	0	1	1	4	0	0	1

$IG_{Total} = 0.3$ $IG_{Total} = 0.2666$ $IG_{Total} = 0.2$ $IG_{Total} = 0$

et l'âge <= 46,5 nous donne l'impureté la plus faible, 0. Les deux feuilles n'ont aucune impureté.

Fonctionnement des arbres de décision

Tanjia	Humour		Age	Marrakchi
Yes	Yes		20	Yes
Yes	No	24	25	No
No	Yes	30	28	Yes
Yes	Yes	33.5	32	Yes
No	Yes		35	Yes
Yes	No	46.5	45	No
No	No		55	No
Yes	Yes		58	No



Maintenant, comme 0 est inférieur à 0,266, nous utiliserons Âge < 46.5 pour diviser ce nœud en feuilles.

Remarque :

il s'agit des feuilles, car il n'y a aucune raison de continuer à diviser ces personnes en groupes plus petits.

En général, les valeurs de sortie de chaque feuille correspondent à la catégorie qui a obtenu le plus grand nombre de votes.

L'algorithme CART

CART (Classification And Regression Trees) est un algorithme fondamental développé par **Breiman** et al. en **1984**, qui **permet de construire des arbres de décision** pour la classification et la régression

Principes clés

- **Division binaire :** Chaque nœud se divise en exactement deux branches (oui/non)
- **Division recursive :** Application répétée du même processus à chaque sous-ensemble
- **Optimisation locale :** Recherche de la meilleure division à chaque étape
- **Critères d'arrêt :** Profondeur maximale, nombre minimal d'échantillons, pureté

CART est devenu la base de la plupart des implémentations modernes d'arbres de décision, y compris celles de scikit-learn, grâce à son efficacité et sa simplicité conceptuelle

L'algorithme CART

CART (Classification And Regression Trees) est un algorithme fondamental développé par **Breiman** et al. en **1984**, qui **permet de construire des arbres de décision** pour la classification et la régression

> L'algorithme évalue toutes les divisions possibles (toutes les variables et tous les seuils) et sélectionne celle qui maximise l'homogénéité (ou minimise l'impureté), garantissant ainsi une amélioration optimale à chaque étape.

Notation mathématique

Pour une variable **X** et un seuil **s**, la division est définie par :

X ≤ s (branche gauche) ou **X > s** (branche droite)

La qualité de la division est évaluée par la réduction d'impureté :

$$\Delta I = I(parent) - [\mathrm{P_g} \times I(gauche) + \mathrm{P_d} \times I(droite)]$$

où $\mathrm{P_g}$ et $\mathrm{P_d}$ sont les proportions d'exemples dans les nœuds enfants.

Mesures d'impureté

Gini : Mesure la probabilité de mauvaise classification

Entropie : Mesure le désordre ou l'incertitude

L'algorithme CART

L'indice de Gini mesure la probabilité qu'un élément soit mal classé si on lui attribue aléatoirement une étiquette selon la distribution des classes dans le nœud.

$$\text{Gini}(i) = 1 - \sum(p_i^2)$$

où p_i est la proportion d'éléments de la classe i dans le nœud t.

Gini = 0 ➔ nœud pur (une seule classe) – Gini proche de 1 ➔ distribution uniforme des classes

L'entropie mesure le niveau de désordre ou d'incertitude dans un ensemble de données.

$$H(t) = -\sum(p_i \times \log_2 p_i)$$

où p_i est la proportion d'éléments de la classe i dans le nœud t.

Le gain d'information mesure la réduction d'entropie obtenue en divisant un nœud.

$$IG(t, \alpha) = H(t) - \sum\left(\left(\frac{t_v}{t}\right) \times H(t_v)\right)$$

où t est le noeud parent, t_v sont les noeuds enfants.

L'entropie et le gain d'information sont particulièrement utiles pour les problèmes multi-classes, car ils pénalisent davantage les distributions non-uniformes.

L'algorithme CART

EXEMPLE CHIFFRÉ:

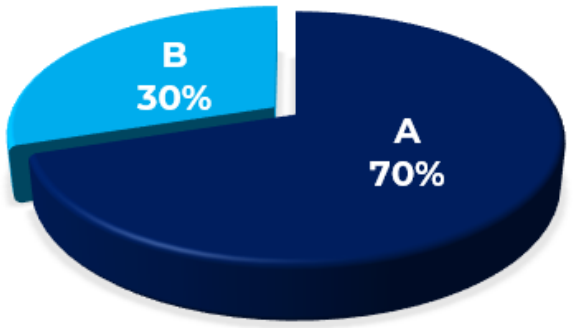
Calcul De L'indice De Gini Étape Par Étape :

Jeu de données d'exemple

Considérons un nœud contenant 10 exemples :

7 de classe A et 3 de classe B.

Classe	Nombre	Proportion
A	7	0.7
B	3	0.3



Calcul de l'indice de Gini

1. Calculer les proportions :

$$P_A = 7/10 = 0.7, \quad P_B = 3/10 = 0.3$$

2. Calculer les carrés :

$$P_A^2 = (0.7)^2 = 0.49, \quad P_B^2 = (0.3)^2 = 0.09$$

3. Somme des carrés :

$$0.49 + 0.9 = 0.58$$

4. Indice de Gini :

$$Gini = 1 - [(0.7)^2 + (0.3)^2] = 1 - 0.58 = 0.42$$

L'indice de Gini de 0.42 indique un niveau d'impureté modéré. Plus la valeur est proche de 0, plus le nœud est pur (homogène). Une valeur de 0.5 représenterait une impureté maximale pour un problème binaire.

L'algorithme CART

EXEMPLE CHIFFRÉ:

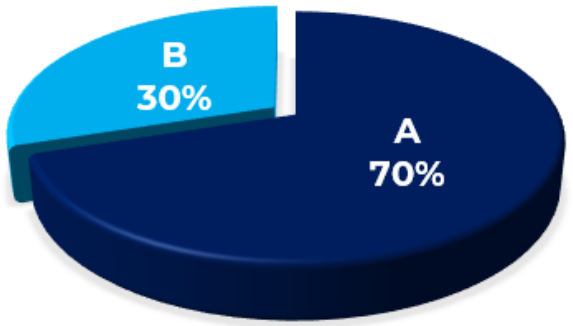
Calcule de l'entropie et du gain d'information

Jeu de données d'exemple

Considérons un nœud contenant 10 exemples :

7 de classe A et 3 de classe B.

Classe	Nombre	Proportion
A	7	0.7
B	3	0.3



Calcul de l'indice de Gini

1. $P_A \times \log_2 P_A = 0.7 \times (-0.51) = -0.36$
2. $P_B \times \log_2 P_B = 0.3 \times (-1.74) = -0.52$
3. $\text{Entropie} = -((-0.36) + (-0.52)) = 0.88$
4. Entropy

$$H(t) = -[0.7 \times \log_2(0.7) + 0.3 \times \log_2(0.3)] = 0.88$$

L'entropie de 0.88 indique un niveau d'incertitude modéré. Une entropie de 0 signifierait un nœud parfaitement pur, tandis qu'une entropie de 1 représenterait une incertitude maximale pour un problème binaire équilibré.

Implémentation Python

Les arbres de décision s'appliquent aussi aux problèmes de régression

Aspect	Classification	Régression
Classe	DecisionTreeClassifier	DecisionTreeRegressor
Critère	gini, entropy	squared_error
Prédiction	Classes discrètes	Valeurs continues
Évaluation	Accuracy, F1-score	MSE, MAE, R^2

Les arbres de régression divisent les données pour minimiser la variance dans chaque nœud. À chaque feuille, la prédiction est la moyenne des valeurs cibles des échantillons dans cette feuille.

Implémentation Python

HYPERPARAMÈTRES IMPORTANTS

Le réglage des hyperparamètres optimise les performances du modèle

- **Criterion** : 'gini' ou 'entropy'

Pour éviter le sur-apprentissage, plusieurs hyperparamètres permettent de limiter la complexité des arbres de décision :

max_depth : Profondeur maximale de l'arbre (nombre de niveaux)

min_samples_split : Nombre minimal d'échantillons requis pour diviser un nœud

min_samples_leaf : Nombre minimal d'échantillons requis dans chaque feuille

min_impurity_decrease : Réduction minimale d'impureté requise pour une division

max_features : Nombre maximal de caractéristiques à considérer pour chaque division

MARRAKECH
جامعة القاضي عياض
UNIVERSITÉ CADI AYYAD

كليـــــة العلـــــوم
السمـلاليـــة - مراكـــش
FACULTÉ DES SCIENCES
SEMLALIA – MARRAKECH

Démonstration Pratique