Université Constantine 2
جامعة قسنطينة 2

Apprentissage machine 1

Chapitre 3 : Classification supervisée
Les arbres de décision

Ouadfel Salima
Faculté NTIC/IFA
salima.ouadfel@univ-constantine2.dz

Université Constantine 2
جامعة قسنطينة 2

Apprentissage machine 1

Chapitre 3 : Classification supervisée
Les arbres de décision

Faculté NTIC/IFA

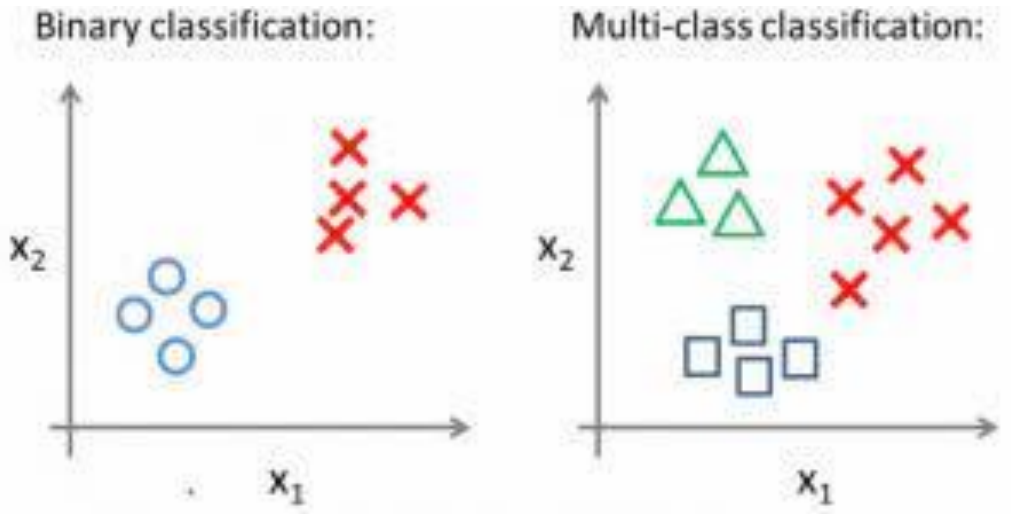
alima.ouadfel@univ-constantine2.dz

Etudiants concernés

Faculté/Institut	Département	Niveau	Spécialité
Nouvelles technologies	IFA	Master1	STIC

Classification supervisée

La classification supervisée est une méthode d'apprentissage supervisé qui cherche à **prédire une valeur qualitative (valeur discrete):** type d'une fleur, email spam ou non spam, la pluie tombera ou non demain,…….



Classification supervisée

Le jeu de données est constitué de n données. Chaque donnée x_i $i = 1..n$ est représentée par p variables explicatives quantitatives ou qualitatives. Chaque donnée x_i est associée avec une variable à expliquer y_i $i = 1..n$ qualitative. La variabley_i est l'étiquette de la classe de x_i . Les valeurs possible de y_i représentent les classes du jeu de données.

$x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$

4 Variables déclaratives

SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.0	3.6	1.4	0.2
...	...	...	...
6.7	3.0	5.2	2.3
6.3	2.5	5.0	1.9
6.5	3.0	5.2	2.0

Données d'apprentissage

Variable expliquée y_i

Species
Iris-setosa
Iris-setosa
Iris-setosa
Iris-setosa
Iris-setosa
...
Iris-virginica
Iris-virginica
Iris-virginica

Labels

Apprentissage

test

SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
5.9	3.0	5.1	1.8

Modèle

Prédiction

Species
Iris-setosa

y_i label

Les arbres de décision

Définition

Un arbre de décision est un modèle d'apprentissage supervisé utilisé pour **représenter visuellement et explicitement les décisions et la prise de décision** .
Il prend en entrée un ensemble d'apprentissage et le répartit en groupes homogènes en fonction d'un objectif connu.

But :

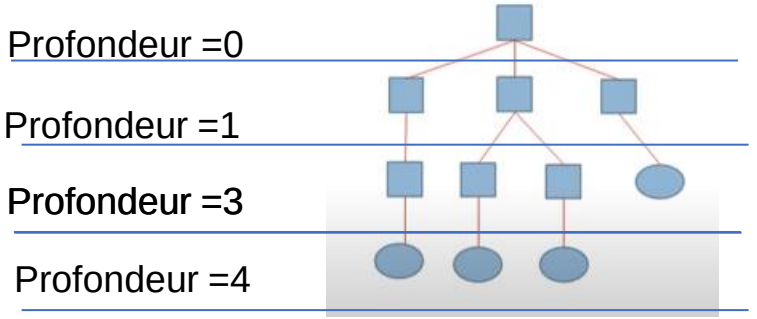
Prédire les valeurs prises par la variable cible

Si la variable cible est continue: arbre de décision de régression

Si la variable cible est discrète: arbre de décision de classification

Structure

Une structure arborescente qui se compose d'un nœud racine, de branches, de nœuds internes et de nœuds feuilles.
Chaque feuille correspond à la valeur à prédire



L'objectif est de créer un modèle qui prédit les valeurs de la variable cible, en se basant sur un ensemble de séquences de règles de décision déduites à partir des données d'apprentissage

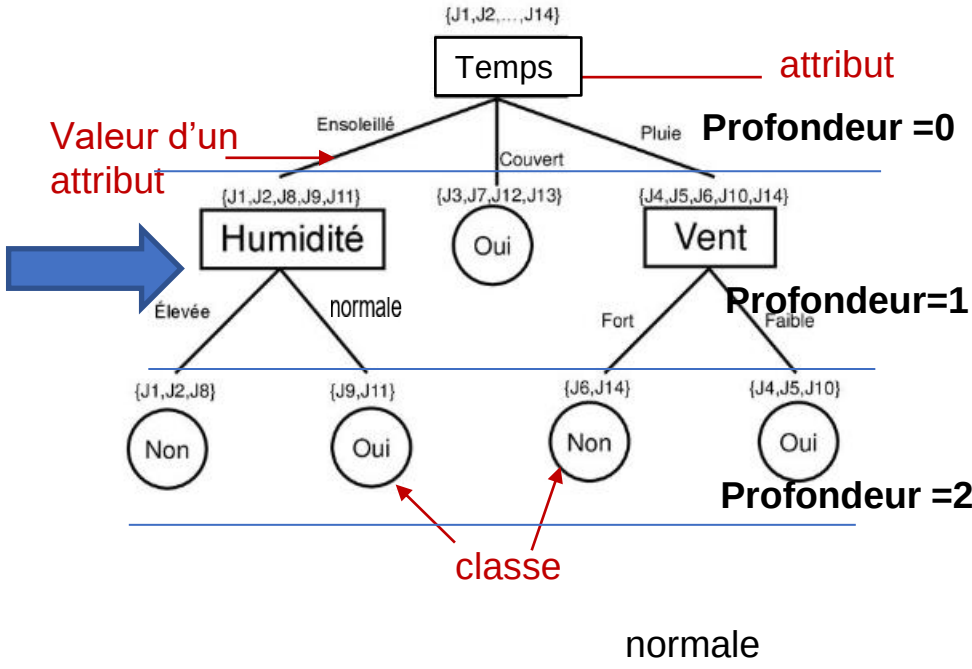
Les arbres de décision

Exemple décider s'il faut jouer au tennis ou non (deux classes)

attributs

classes

Jour	Temps	Température	humidité	vent	Jouer au tennis ?
1	Ensoleillé	Chaude	Élevée	Faible	Non
2	Ensoleillé	Chaude	Élevée	Fort	Non
3	Couvert	Chaude	Élevée	Faible	Oui
4	Pluie	Tiède	Élevée	Faible	Oui
5	Pluie	Fraîche	Normale	Faible	Oui
6	Pluie	Fraîche	Normale	Fort	Non
7	Couvert	Fraîche	Normale	Fort	Oui
8	Ensoleillé	Tiède	Élevée	Faible	Non
9	Ensoleillé	Fraîche	Normale	Faible	Oui
10	Pluie	Tiède	Normale	Faible	Oui
11	Ensoleillé	Tiède	Normale	Fort	Oui
12	Couvert	Tiède	Élevée	Fort	Oui
13	Couvert	Chaud	Normale	Faible	Oui
14	Pluie	Tiède	Élevée	Fort	Non



une **règle** est générée pour chaque **chemin** de l'arbre(de la racine à une feuille)

Les arbres de décision

Exemple Classification des fleurs d'iris

SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.0	3.6	1.4	0.2
...	...	...	...
6.7	3.0	5.2	2.3
6.3	2.5	5.0	1.9
6.5	3.0	5.2	2.0

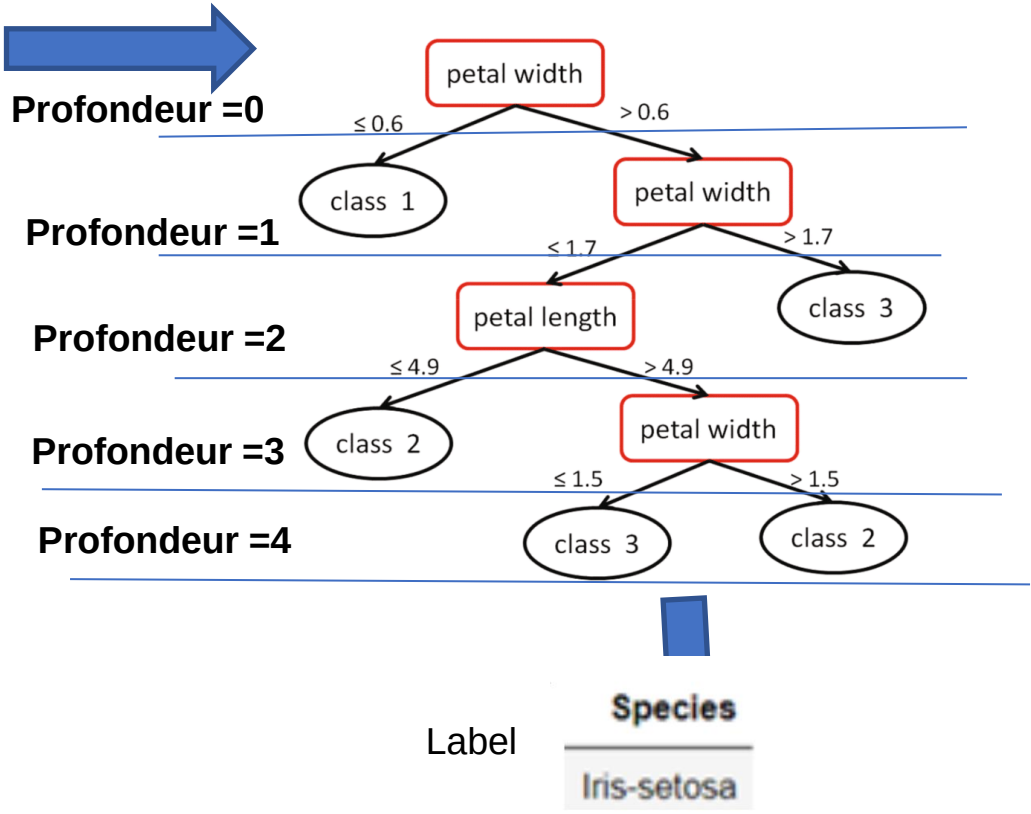
Données d'apprentissage

Species
Iris-setosa
Iris-setosa
Iris-setosa
Iris-setosa
Iris-setosa
...
Iris-virginica
Iris-virginica
Iris-virginica

Classes

test

SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
5.9	3.0	5.1	1.8



Label

Species
Iris-setosa

Les arbres de décision

Algorithme de base de construction

L'arbre est construit récursivement de haut en bas selon le principe "diviser pour régner"

1. Au début l'ensemble d'apprentissage S est dans le nœud racine

2. Déterminer le meilleur attribut dans S.

3. Diviser selon un critère sur l'attribut sélectionné, l'ensemble S en des sous-ensembles S_i $i = 1..m$.

4. Revenir vers l'étape 2 pour chaque sous-ensemble S_i .

5. On s'arrête quand tous les exemples (ou la majorité) d'un nœud appartiennent à la même classe ou le nombre minimum d'exemples dans le nœud est atteint ou le nombre de profondeur maximum est atteint ou on a plus d'attributs à tester.

Comment choisir le critère de sélection?
Comment choisir le critère de division?

Variantes de l'algorithme de base

- **ID3** (Iterative Dichotomiser 3): développé en 1986 par Ross Quinlan. Il est appliqué seulement sur les attributs nominales et pour la classification.
- **C4.5**: une extension de ID3 par Ross Quinlan. Il est appliqué pour tous types d'attributs et pour la classification.
- **CART** (Classification and Regression Trees): se base sur C4.5 mais avec d'autres métriques. L'algorithme CART s'applique pour la classification et pour la régression.

Différences entre les algorithmes

Prise en compte ou non les données continues

Critère de sélection du meilleur attribut

Critère de division des données

Attributs numériques versus attributs catégoriques

Jour	Temps	Température	humidité	vent	Jouer au tennis ?
1	Ensoleillé	Chaude	Élevée	Faible	Non
2	Ensoleillé	Chaude	Élevée	Fort	Non
3	Couvert	Chaude	Élevée	Faible	Oui
4	Pluie	Tiède	Élevée	Faible	Oui
5	Pluie	Fraîche	Normale	Faible	Oui
6	Pluie	Fraîche	Normale	Fort	Non
7	Couvert	Fraîche	Normale	Fort	Oui
8	Ensoleillé	Tiède	Élevée	Faible	Non
9	Ensoleillé	Fraîche	Normale	Faible	Oui
10	Pluie	Tiède	Normale	Faible	Oui
11	Ensoleillé	Tiède	Normale	Fort	Oui
12	Couvert	Tiède	Élevée	Fort	Oui
13	Couvert	Chaud	Normale	Faible	Oui
14	Pluie	Tiède	Élevée	Fort	Non

Valeurs catégoriques

Jour	Temps	Température	humidité	vent	Jouer au tennis ?
1	Ensoleillé	27,5	85	Faible	Non
2	Ensoleillé	25	90	Fort	Non
3	Couvert	26,5	86	Faible	Oui
4	Pluie	20	96	Faible	Oui
5	Pluie	19	80	Faible	Oui
6	Pluie	17,5	70	Fort	Non
7	Couvert	17	65	Fort	Oui
8	Ensoleillé	21	95	Faible	Non
9	Ensoleillé	19,5	70	Faible	Oui
10	Pluie	22,5	80	Faible	Oui
11	Ensoleillé	22,5	70	Fort	Oui
12	Couvert	21	90	Fort	Oui
13	Couvert	25,5	75	Faible	Oui
14	Pluie	20,5	91	Fort	Non

Valeurs numériques

Mesure de sélection du meilleur l'attribut

Comment choisir l'attribut qui apporte le plus d'information au résultat à prédire?

On veut faire de la classification, donc on choisit l'attribut qui maximise une mesure d'homogéinéité obtenue par le découpage guidé par cet attribut.

- Le gain d'information (ID3 et C4.5)
- L'index de Gini (CART)

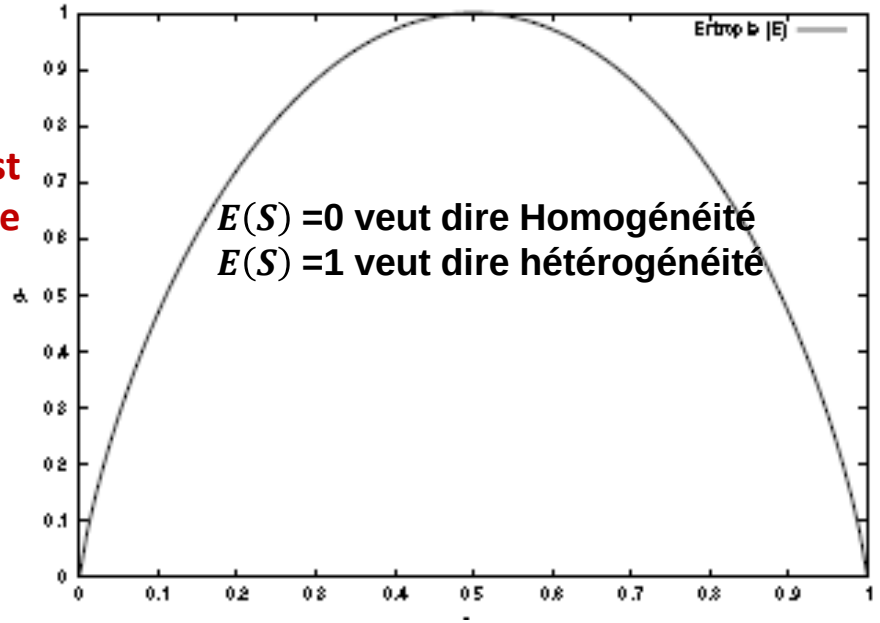
Sélection du meilleur l'attribut avec le gain d'information

1- Calcul de l'Entropie

L'entropie E de l'ensemble **S mesure son homogénéité et** est exprimée par:

$$\boldsymbol{E(S) = -\sum_{i=1}^{C} p_i log_2 p_i} \textbf{ avec } \boldsymbol{p_i = \frac{|C_i|}{|S|}} \textbf{ et C} = \text{nombre de classes}$$

Le nœud de l'arbre dont la valeur de l'entropie est nulle est considéré comme feuille de l'arbre de décision.



Sélection du meilleur l'attribut avec le gain d'information

2- Calcul du gain d'information

Soit S l'ensemble de données et soit a_j un attribut. Le gain d'information de S par rapport à a_j mesure la différence entre l'entropie originale de S et celle après sa division en des sous ensembles en se basant sur a_j.

$$IG(S, a_j) = E(S) - \sum_{v \in valeurs(a_j)} p(S_{a_j=v})\ E(S_{a_j=v})$$

Où

$valeurs(a_j)$ représentent toutes les valeurs possibles de l'attribut a_j et

$$p\left(S_{a_j=v}\right) = \frac{\left|S_{a_j=v}\right|}{|S|}$$

L'attribut a_j qui maximise le gain d'information est sélectionné.

Les arbres de décision

Exemple de construction de l'arbre de décision avec le gain d'information

Soit S l'ensemble de données du jeu de tennis,

1- calcul de L'entropie de l'ensemble S

On a 14 données , 9 oui et 5 non

$$E(S) = -\sum_{i=1}^{2} p_i log_2(p_i) = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14}$$

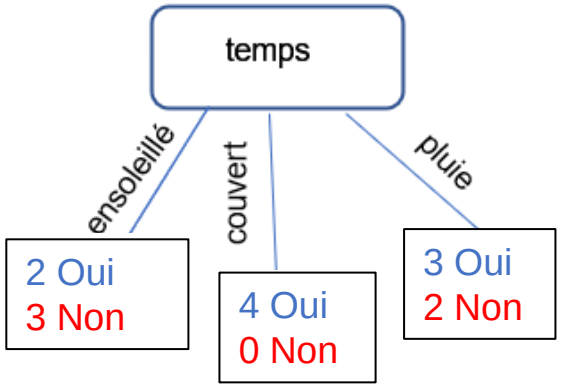
$$\boxed{E(S) = 0.940}$$

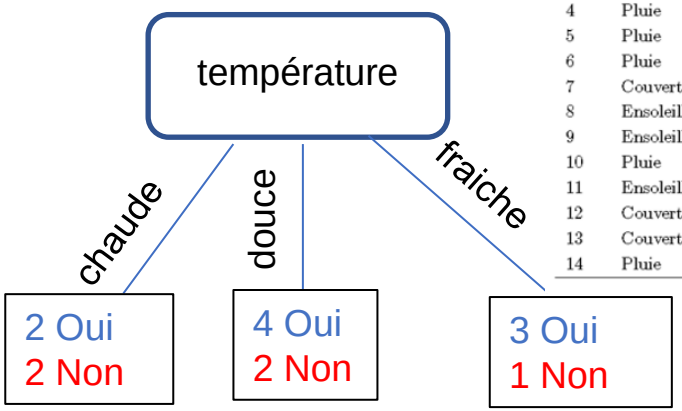
Jour	Temps	Température	humidité	vent	Jouer
1	Ensoleillé	Chaude	Élevée	Faible	Non
2	Ensoleillé	Chaude	Élevée	Fort	Non
3	Couvert	Chaude	Élevée	Faible	Oui
4	Pluie	Tiède	Élevée	Faible	Oui
5	Pluie	Fraîche	Normale	Faible	Oui
6	Pluie	Fraîche	Normale	Fort	Non
7	Couvert	Fraîche	Normale	Fort	Oui
8	Ensoleillé	Tiède	Élevée	Faible	Non
9	Ensoleillé	Fraîche	Normale	Faible	Oui
10	Pluie	Tiède	Normale	Faible	Oui
11	Ensoleillé	Tiède	Normale	Fort	Oui
12	Couvert	Tiède	Élevée	Fort	Oui
13	Couvert	Chaud	Normale	Faible	Oui
14	Pluie	Tiède	Élevée	Fort	Non

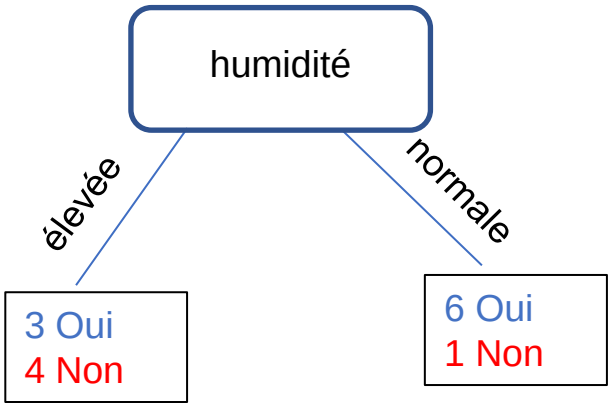
Exemple de construction de l'arbre de décision avec le gain d'information

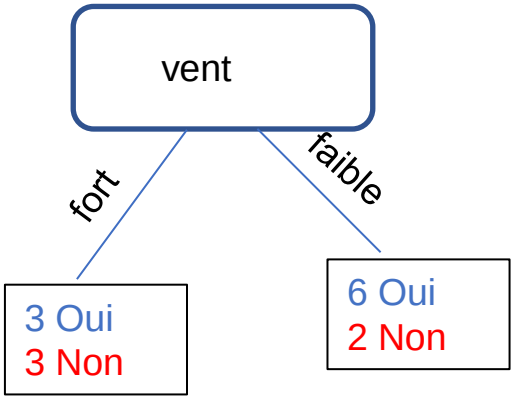
2- sélection du meilleur attribut de S

Jour	Temps	Température	humidité	vent	Jouer
1	Ensoleillé	Chaude	Élevée	Faible	Non
2	Ensoleillé	Chaude	Élevée	Fort	Non
3	Couvert	Chaude	Élevée	Faible	Oui
4	Pluie	Tiède	Élevée	Faible	Oui
5	Pluie	Fraîche	Normale	Faible	Oui
6	Pluie	Fraîche	Normale	Fort	Non
7	Couvert	Fraîche	Normale	Fort	Oui
8	Ensoleillé	Tiède	Élevée	Faible	Non
9	Ensoleillé	Fraîche	Normale	Faible	Oui
10	Pluie	Tiède	Normale	Faible	Oui
11	Ensoleillé	Tiède	Normale	Fort	Oui
12	Couvert	Tiède	Élevée	Fort	Oui
13	Couvert	Chaud	Normale	Faible	Oui
14	Pluie	Tiède	Élevée	Fort	Non









Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut de S

Si on choisi l'attribut temps:

temps	Oui	Non	Total
Ensoleillé	2	3	5
Couvert	4	0	4
Pluie	3	2	5

$$IG(S, temps) = E(S) - \sum_{v \in valeurs(temps)} p(S_{temps=v})\ E(S_{temps=v})$$

$$IG(S, temps) = 0.94 - [\, p(S_{temps=ensoleillé}) * E(S_{temps=ensoleillé}) + p(S_{temps=couvert}) * E(S_{temps=couvert}) + p(S_{temps=pluie}) * E(S_{temps=pluie}) \,]$$

$$p(S_{temps=ensoleillé}) * E(S_{temps=ensoleillé}) = \frac{5}{14} * \left(-\frac{2}{5} log_2 \frac{2}{5} - \frac{3}{5} log_2 \frac{3}{5} \right)$$

$$p(S_{temps=couvert}) * E(S_{temps=couvert}) = \frac{4}{14} * \left(-\frac{4}{4} log_2 \frac{4}{4} - \frac{0}{4} log_2 \frac{0}{4} \right)$$

$$p(S_{temps=pluie}) * E(S_{temps=pluie}) = \frac{5}{14} * \left(-\frac{3}{5} log_2 \frac{3}{5} - \frac{2}{5} log_2 \frac{2}{5} \right)$$

$$IG(S, temps) = 0.246$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut de S

Si on choisi l'attribut température:

température	Oui	Non	Total
Chaude	2	2	4
Douce	4	2	6
Fraiche	3	1	4

$$IG(S, temperature) = E(S) - \sum_{v \in valeurs(temperature)} p(S_{temperature=v})\ E(S_{temperature=v})$$

$$IG(S, temperature) = 0.94 - [\, p(S_{temperature=chaude}) * E(S_{temperature=chaude}) + p(S_{temperature=douce}) * E(S_{temperature=douce}) + p(S_{temperature=fraiche}) * E(S_{temperature=fraiche}) \,]$$

$$p(S_{temperature=chaude}) * E(S_{temperature=chaude}) = \frac{4}{14} * \left(-\frac{2}{4} log_2 \frac{2}{4} - \frac{2}{4} log_2 \frac{2}{4}\right) =$$

$$p(S_{temperature=douce}) * E(S_{temperature=douce}) = \frac{6}{14} * \left(-\frac{4}{6} log_2 \frac{4}{6} - \frac{2}{6} log_2 \frac{2}{6}\right)$$

$$p(S_{temperature=fraiche}) * E(S_{temperature=fraiche}) = \frac{4}{14} * \left(-\frac{3}{4} log_2 \frac{3}{4} - \frac{1}{4} log_2 \frac{1}{4}\right)$$

$$IG(S, temperature) = 0.030$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut de S

Si on choisi l'attribut humidité:

humidité	Oui	Non	Total
Élevée	3	4	7
normale	6	1	7

$$IG(S, humidité) = E(S) - \sum_{v \in valeurs(humidité)} p(S_{humidité=v})\ E(S_{humidité=v})$$

$$IG(S, humidité) = 0.94 - [\ p(S_{humidité=élevée}) * E(S_{humidité=élevée}) + p(S_{humidité=normale}) * E(S_{humidité=normale})\ ]$$

$$p(S_{humidité=élevée}) * E(S_{humidité=élevée}) = \frac{7}{14} * \left(-\frac{3}{7} log_2 \frac{3}{7} - \frac{4}{7} log_2 \frac{4}{7}\right)$$

$$p(S_{humidité=normale}) * E(S_{humidité=normale}) = \frac{7}{14} * (-\frac{6}{7} log_2 \frac{6}{7} - \frac{1}{7} log_2 \frac{1}{7})$$

$$IG(S, humidité) = 0.152$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut de S

Si on choisi l'attribut vent:

vent	Oui	Non	Total
fort	3	3	6
faible	6	2	8

$$IG(S, vent) = E(S) - \sum_{v \in valeurs(vent)} p(S_{vent=v})\ E(S_{vent=v})$$

$$IG(S, vent) = 0.94 - [\, p(S_{vent=fort}) * E(S_{vent=fort}) + p(S_{vent=faible}) * E(S_{vent=faible})]$$

$$p(S_{vent=fort}) * E(S_{vent=fort}) = \frac{6}{14} * \left(-\frac{3}{6} log_2 \frac{3}{6} - \frac{3}{6} log_2 \frac{3}{6}\right)$$

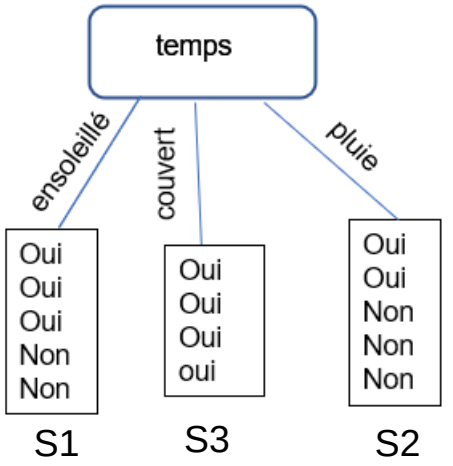
$$p(S_{vent=faible}) * E(S_{vent=faible}) = \frac{8}{14} * (-\frac{6}{8} log_2 \frac{6}{8} - \frac{2}{8} log_2 \frac{2}{8})$$

$$IG(S, vent\ ) = 0.048$$

Exemple de construction de l'arbre de décision avec le gain d'information

3- division de S selon l'attribut temps

Attribut	Gain d'information
Temps	**0.246**
Température	0.030
Humidité	0.152
Vent	0.048



S1

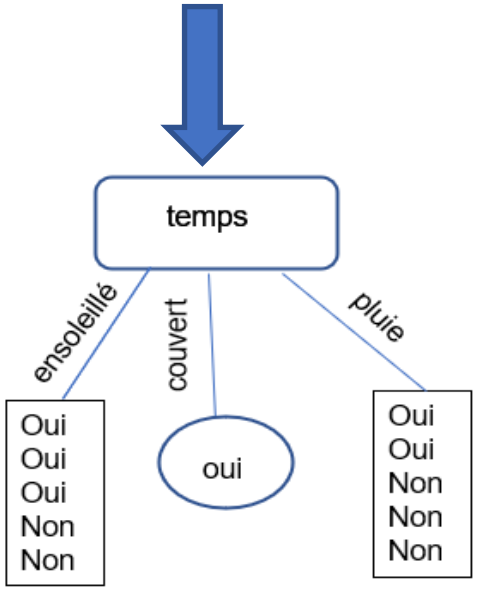
temps	température	humidité	vent	Jouer
Ensoleillé	chaude	Élevée	faible	non
	chaude	Élevée	fort	non
	douce	Élevée	faible	non
	fraiche	normale	faible	oui
	douce	normale	fort	oui

S2

temps	température	humidité	vent	Jouer
pluie	douce	Élevée	faible	oui
	fraiche	normale	faible	oui
	fraiche	normale	fort	non
	douce	normale	faible	oui
	douce	Élevée	fort	non

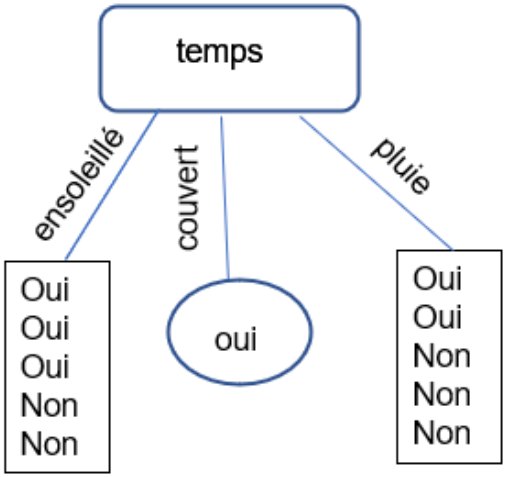
S3

temps	température	humidité	vent	Jouer
Couvert	chaude	Élevée	faible	oui
	fraiche	normale	fort	oui
	tiède	Élevée	fort	oui
	chaude	normale	faible	oui



Exemple de construction de l'arbre de décision avec le gain d'information

3- division de S selon l'attribut temps



On considère maintenant les deux sous ensembles S1 et S2:

temps	température	humidité	vent	Jouer
Ensoleillé	chaude	Élevée	faible	non
	chaude	Élevée	fort	non
	douce	Élevée	faible	non
	fraiche	normale	faible	oui
	douce	normale	fort	oui

S1

temps	température	humidité	vent	Jouer
pluie	douce	Élevée	faible	oui
	fraiche	normale	faible	oui
	fraiche	normale	fort	non
	douce	normale	faible	oui
	douce	Élevée	fort	non

S2

Exemple de construction de l'arbre de décision avec le gain d'information

temps	température	humidité	vent	Jouer
Ensoleillé	chaude	Élevée	faible	non
	chaude	Élevée	fort	non
	douce	Élevée	faible	non
	fraiche	normale	faible	oui
	douce	normale	fort	oui

1- Calcule de l'entropie de l'ensemble S1

$$E(S_1) = -\sum_{i=1}^{2} p_i log_2(p_i)$$

$$= -\frac{2}{5} log_2 \frac{2}{5} - \frac{3}{5} log_2 \frac{3}{5}$$

$$E(S_1) = 0.971$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut de S1

Si on choisi l'attribut température:

temps	Température	Oui	Non	Total
Ensoleillé	Chaude	0	2	2
	Douce	1	1	2
	Fraiche	1	0	1

$$IG(S_1, température) = E(S_1) - \sum_{v \in valeurs(température)} p(S_{1\ température=v})\ E(S_{1\ température=v})$$

$$IG(S_1, température) = 0.97 - [\, p(S_{1\ température=chaude}) * E(S_{1\ température=chaude}) + p(S_{1\ température=douce}) * E(S_{1\ température=douce}) + p(S_{1\ température=fraiche}) * E(S_{1\ température=fraiche}) \,]$$

$$p(S_{1\ température=chaude}) * E(S_{1\ température=chaude}) = \frac{2}{5} * \left(-\frac{0}{2} log_2 \frac{0}{2} - \frac{2}{2} log_2 \frac{2}{2} \right)$$

$$p(S_{1\ température=douce}) * E(S_{1\ température=douce}) = \frac{2}{5} * \left(-\frac{1}{2} log_2 \frac{1}{2} - \frac{1}{2} log_2 \frac{1}{2} \right)$$

$$p(S_{1\ température=fraiche}) * E(S_{1\ température=fraiche}) = \frac{1}{5} * \left(-\frac{1}{1} log_2 \frac{1}{1} - \frac{0}{1} log_2 \frac{0}{1} \right)$$

$$\boxed{IG(S_1, température) = 0.571}$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut de S1

Si on choisi l'attribut humidité:

temps	humidité	Oui	Non	Total
Ensoleillé	Elevée	0	3	3
	Normale	2	0	2

$$IG(S_1, humidité) = E(S_1) - \sum_{v \in valeurs(humidité)} p(S_{1humidité=v})\ E(S_{1humidité=v})$$

$$IG(S_1, humidité) = 0.97 - [\, p(S_{1humidité=élevée}) * E(S_{1humidité=élevée}) + p(S_{1humidité=normale}) * E(S_{1humidité=normale})]$$

$$p(S_{1humidité=élevée}) * E(S_{1humidité=élevée}) = \frac{3}{5} * \left(-\frac{0}{3} log_2 \frac{0}{3} - \frac{3}{3} log_2 \frac{3}{3}\right)$$

$$p(S_{1humidité=normale}) * E(S_{1humidité=normale}) = \frac{2}{5} * \left(-\frac{2}{2} log_2 \frac{2}{2} - \frac{0}{2} log_2 \frac{0}{2}\right)$$

$$\boxed{IG(S_1, humidité) = 0.971}$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut de S1

Si on choisi l'attribut vent:

temps	vent	Oui	Non	Total
Ensoleillé	Fort	1	1	2
	Faible	1	2	3

$$IG(S_1, vent) = E(S_1) - \sum_{v \in valeurs(vent)} p(S_{1\ vent=v})\ E(S_{1\ vent=v})$$

$$IG(S_1, vent) = 0.97 - [\, p(S_{1vent=fort}) * E(S_{1vent=fort}) + p(S_{1vent=faible}) * E(S_{1fort=faible})]$$

$$p(S_{1vent=fort}) * E(S_{1vent=fort}) = \frac{2}{5} * \left(-\frac{1}{2} log_2 \frac{1}{2} - \frac{1}{2} log_2 \frac{1}{2}\right)$$

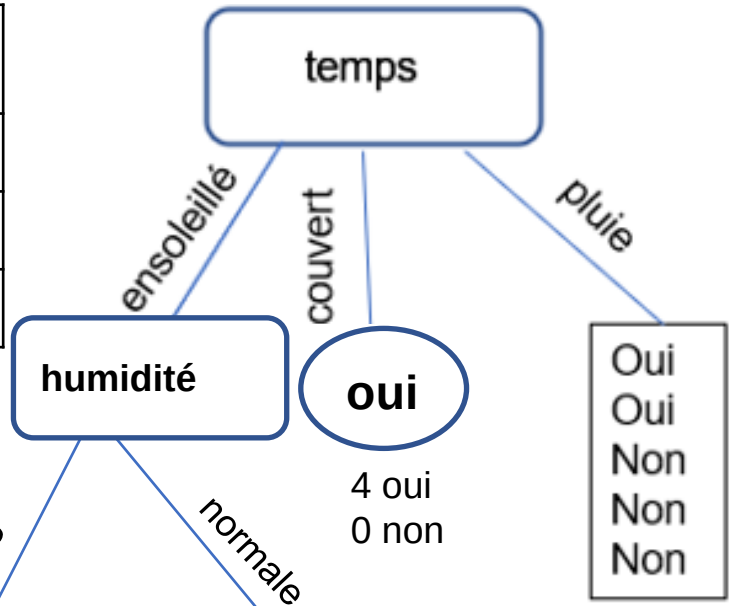
$$p(S_{1vent=faible}) * E(S_{1vent=faible}) = \frac{3}{5} * \left(-\frac{1}{3} log_2 \frac{1}{3} - \frac{2}{3} log_2 \frac{2}{3}\right)$$

$$IG(S_1, vent) = 0.020$$

Exemple de construction de l'arbre de décision avec le gain d'information

3- Division de S1 selon le meilleur attribut humidité

Attribut	Gain d'information
température	0.571
humidité	**0.971**
vent	0.020



temps	température	humidité	vent	Jouer
Ensoleillé	chaude	Élevée	faible	non
	chaude	Élevée	fort	non
	douce	Élevée	faible	non

temps	température	humidité	vent	Jouer
Ensoleillé	fraiche	normale	faible	oui
	douce	normale	fort	oui

Exemple de construction de l'arbre de décision avec le gain d'information

1- Calcule de l'entropie de l'ensemble S2

$$E(S_2) = -\sum_{i=1}^{2} p_i log_2(p_i)$$

$$= -\frac{3}{5} log_2 \frac{3}{5} - \frac{2}{5} log_2 \frac{2}{5}$$

$$E(S_2) = 0.971$$

temps	température	humidité	vent	Jouer
pluie	douce	Élevée	faible	oui
	fraiche	normale	faible	oui
	fraiche	normale	fort	non
	douce	normale	faible	oui
	douce	Élevée	fort	non

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut de S2

Si on choisi l'attribut température:

temps	Température	Oui	Non	Total
Pluie	Douce	2	1	3
	Fraiche	1	1	2

$$IG(S_2, température) = E(S_2) - \sum_{v \in valeurs(température)} p(S_{2\ température=v})\ E(S_{2\ température=v})$$

$$IG(S_2, température) = 0.97 - [p(S_{2\ température=douce}) * E(S_{2\ température=douce}) + p(S_{2\ température=fraiche}) * E(S_{2\ température=fraiche})]$$

$$p(S_{2\ température=douce}) * E(S_{2\ température=douce}) = \frac{3}{5} * \left(-\frac{2}{3} log_2 \frac{2}{3} - \frac{1}{3} log_2 \frac{1}{3}\right)$$

$$p(S_{2\ température=fraiche}) * E(S_{2\ température=fraiche}) = \frac{2}{5} * \left(-\frac{1}{2} log_2 \frac{1}{2} - \frac{1}{2} log_2 \frac{1}{2}\right)$$

$$IG(S_2, température) = 0.020$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut de S2

Si on choisi l'attribut humidité:

temps	humidité	Oui	Non	Total
Pluie	Elevée	1	1	2
	Normale	2	1	3

$$IG(S_2, humidité) = E(S_2) - \sum_{v \in valeurs(humidité)} p(S_{2\ humidité=v})\ E(S_{2\ humidité=v})$$

$$IG(S_2, humidité) = 0.97 - [\, p(S_{2\ humidité=élevée}) * E(S_{2\ humidité=élevée}) + p(S_{2\ humidité=normale}) * E(S_{2\ humidité=normale})]$$

$$p(S_{2\ humidité=élevée}) * E(S_{2\ humidité=élevée}) = \frac{2}{5} * \left(-\frac{1}{2} log_2 \frac{1}{2} - \frac{1}{2} log_2 \frac{1}{2}\right)$$

$$p(S_{2\ humidité=normale}) * E(S_{2\ humidité=normale}) = \frac{3}{5} * \left(-\frac{2}{3} log_2 \frac{2}{3} - \frac{1}{3} log_2 \frac{1}{3}\right)$$

$$\boxed{IG(S_2, humidité) = 0.020}$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut de S2

Si on choisi l'attribut vent:

temps	vent	Oui	Non	Total
Pluie	Fort	0	2	2
	Faible	3	0	3

$$IG(S_2, vent) = E(S_2) - \sum_{v \in valeurs(vent)} p(S_{2\ vent=v})\ E(S_{2\ vent=v})$$

$$IG(S_2, vent) = 0.97 - [\, p(S_{2\ vent=fort}) * E(S_{2\ vent=fort}) + p(S_{2\ vent=faible}) * E(S_{2\ fort=faible})]$$

$$p(S_{2\ vent=fort}) * E(S_{2\ vent=fort}) = \frac{2}{5} * \left(-\frac{0}{2} log_2 \frac{0}{2} - \frac{2}{2} log_2 \frac{2}{2}\right)$$

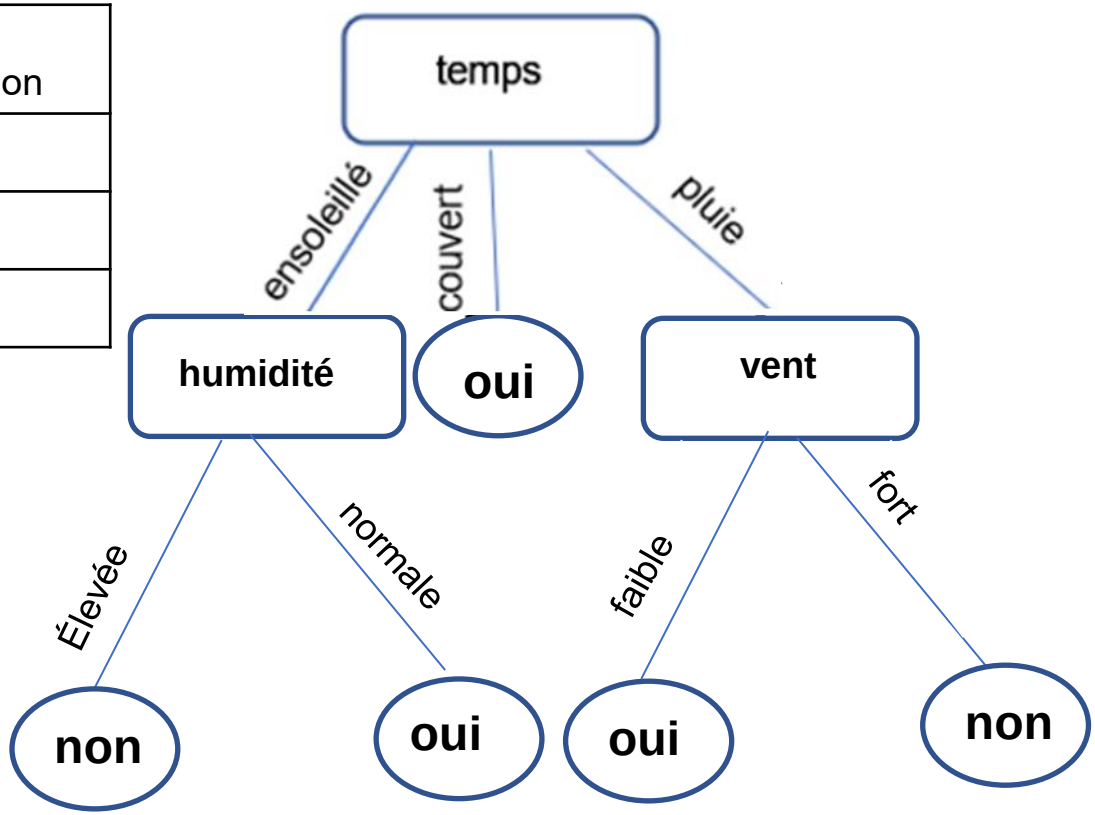
$$p(S_{2\ vent=faible}) * E(S_{2\ vent=faible}) = \frac{3}{5} * \left(-\frac{3}{3} log_2 \frac{3}{3} - \frac{0}{3} log_2 \frac{0}{3}\right)$$

$$\boxed{IG(S_2, vent) = 0.971}$$

Exemple de construction de l'arbre de décision avec le gain d'information

3- division de S2 selon le meilleur attribut vent

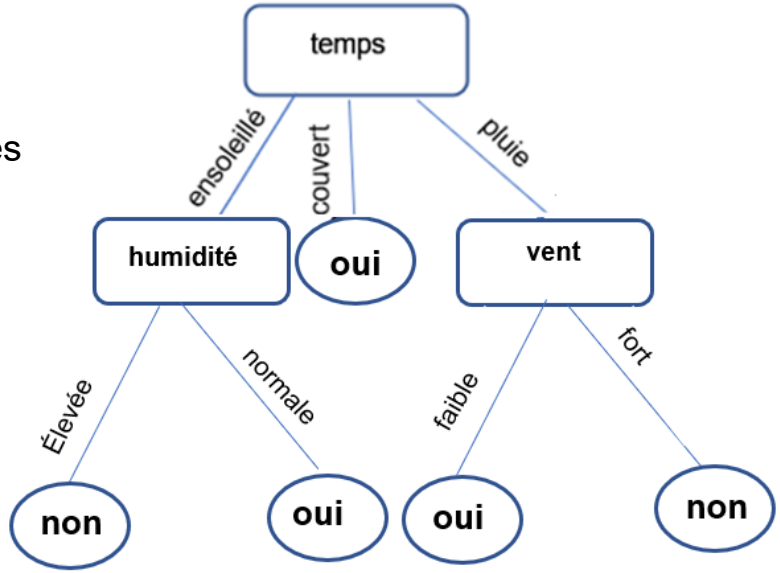
Attribut	Gain d'information
température	0.020
humidité	0.020
vent	**0.971**



Exemple de construction de l'arbre de décision avec le gain d'information

Extraction des règles de décision

Les règles de décision sont de la forme: si-alors
Chaque règle est construite de la racine vers l'une des feuilles de l'arbres
Les feuilles représentent les classes



Prédiction

- (Ensoleillé, Fraîche, Élevée, Fort) est classée comme « non » ;
- (Ensoleillé, Fraîche, Normale, Fort) est classée comme « oui » ;
- (Pluie, Chaude, Normale, Faible) est classée comme « oui » ;
- (Pluie, Fraîche, Élevée, Fort) est classée comme « non ».