Apprentissage d'arbre de décision

Day	Temperature	Humidity	Wind	Weather	Play
1	Hot	High	Weak	Sunny	no
2	Hot	High	Strong	Sunny	no
3	Hot	High	Weak	Overcast	yes
4	Mild	High	Weak	Rain	yes
5	Cool	Normal	Weak	Rain	yes
6	Cool	Normal	Strong	Rain	no
7	Cool	Normal	Strong	Overcast	yes
8	Mild	High	Weak	Sunny	no
9	Cool	Normal	Weak	Sunny	yes
10	Mild	Normal	Weak	Rain	yes
11	Mild	Normal	Strong	Sunny	yes
12	Mild	High	Strong	Overcast	yes
13	Hot	Normal	Weak	Overcast	yes
14	Mild	High	Strong	Rain	no

Question 1 : Trouvez avec votre intuition un arbre de décision qui va indiquer si l'utilisateur va jouer ou non et qui soit cohérent avec ces données.

Question 2 : L'entropie caractérise le désordre d'une collection d'instances. Soit S un ensemble d'instances dont la classification est binaire (les instances sont classées comme positives ou négatives). Une interprétation de la théorie de l'information est que l'entropie indique le nombre de bits nécessaires pour encoder la classification d'un membre arbitraire de S. L'entropie de S pour cette classification booléenne est définie par :

$$\text{entropie}(S) = -p_{\oplus} log_2(p_{\oplus}) - p_{\ominus} log_2(p_{\ominus})$$

où

- $p_{\oplus}$ est la proportion d'examples positifs dans S
- $p_{\ominus}$ est la proportion d'examples négatifs dans S

On veut calculer le gain en information qui est la réduction d'entropie espérée en partitionnant les instances suivant la valeur d'un attribut. Pour un ensemble d'instance S et un attribut A, le gain est défini par :

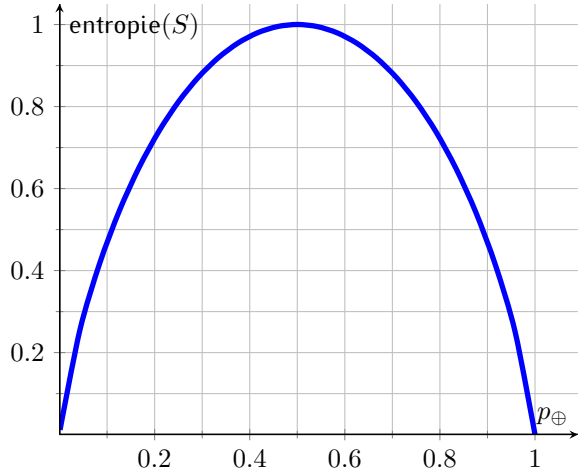
$$\text{gain}(S, A) = \text{entropie}(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \text{entropie}(S_v)$$

où S_v est l'ensemble des instances dont l'attribut A a pour valeur v. Appliquez la formule du gain pour les attributs `Wind` et `Temperature`.

Question 3 : L'application numérique donne le tableau suivant :

	Weather	Humidity	Wind	Temperature
Gain	0.246	0.151	0.048	0.029

— Vérifiez (à la maison) que votre expression est correcte
— Traitez maintenant le cas où le temps est ensoleillé : donnez l'expression du gain. Sans faire les calculs, pouvez vous avoir une bonne idée sur le résultat ?
— Pouvez vous deviner l'attribut pour le cas où il pleut ?
— Faites de même pour le cas où le temps est nuageux.
— Dessinez l'arbre de décision produit par l'algorithme.



entropie(S) pour un ensemble de données S dont $p_\oplus$ est la proportion d'instances positives ($1 - p_\oplus$ est la proportion d'instances négatives).