

Processus d'extraction de connaissances

1 Le golf

Les échantillons suivants correspondent à un ensemble de conditions météorologiques qui permettent (P pour positif) ou pas (N pour Négatif) la pratique du golf :

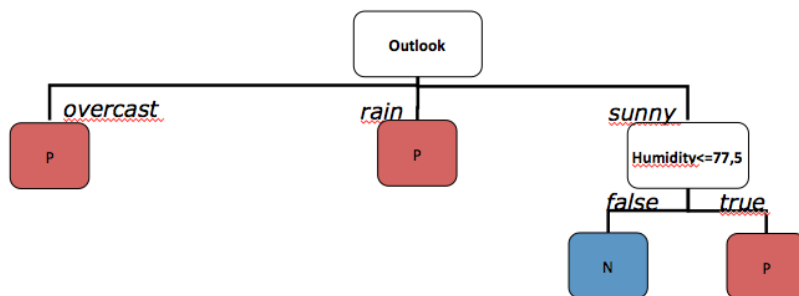


FIGURE 1 – Arbre de décision obtenu avec un critère d'arrêt.

num	Outlook	Temperature	Humidity	Windy	Play
1	sunny	hot	78	false	N
2	sunny	hot	90	true	N
3	overcast	hot	80	false	P
4	rainy	mild	96	false	P
5	rainy	cool	76	false	P
6	rainy	cool	75	false	P
7	overcast	cool	65	true	P
8	overcast	mild	83	false	N
9	sunny	cool	72	false	P
10	rainy	mild	74	false	P

TABLE 1 – Exemples de test

Q. 1 : Etablissez la matrice de confusion obtenue par l'arbre de décision de la figure 1 sur les exemples de test de la table 1.

Q. 2 : A quelle(s) valeur(s) doit-on comparer le taux d'erreur pour savoir si le modèle est a priori intéressant ?

Q. 3 : Calculez le rappel pour les classes P et N .

Q. 4 : Calculez la précision pour les classes P et N .

Q. 5 : Ce modèle est-il utile ? Justifiez.

2 Le Vin

On souhaite prédire la qualité d'un vin à partir de variables descriptives (non fournies ici).

Q. 1 : Calculez le taux d'erreur au regard de la table 2.

		<i>Classes prédites</i>		
		mauvais	moyen	bon
<i>Classes réelles</i>	mauvais	11	0	1
	moyen	6	1	4
	bon	0	0	11

TABLE 2 – Matrice de confusion obtenue sur le problème de prédiction de la qualité du vin.

Q. 2 : A quelle(s) valeur(s) doit-on comparer le taux d'erreur pour savoir si le modèle est bon ?

Q. 3 : Calculez le rappel pour la classe *mauvais*.

Q. 4 : Calculez la précision pour la classe *mauvais*.

Q. 5 : Ce modèle est-il utile ? Justifiez.

3 Titanic

A partir d'une base de données¹ dont les premiers enregistrements figurent dans la table 3, nous souhaitons obtenir un modèle pour étudier ce qui s'est passé lors du naufrage du Titanic en 1912. Nous avons donc utilisé un arbre de décision dont le résultat est présenté dans la figure 2.

1. Titanic in the Classroom, Department of Education & Children Services, Southern Australia

TABLE 3 – Extrait de la base du Titanic¹. La base complète contient 755 passagers. La variable "Class" peut prendre les valeurs "1st", "2nd" ou "3rd".

Name	Class	Age	Sex	Status
"Allen, Miss Elisabeth Walton"	1st	29.0	female	survived
"Allison, Miss Helen Loraine"	1st	2.0	child	died
"Allison, Mr Hudson Joshua Creighton"	1st	30.0	male	died
"Allison, Mrs Hudson JC (Bessie Waldo Daniels)"	1st	25.0	female	died
"Allison, Master Hudson Trevor"	1st	0.92	child	survived
"Anderson, Mr Harry"	1st	47.0	male	survived
"Andrews, Miss Kornelia Theodosia"	1st	63.0	female	survived
"Andrews, Mr Thomas, jr"	1st	39.0	male	died
.....	...	...	...	...

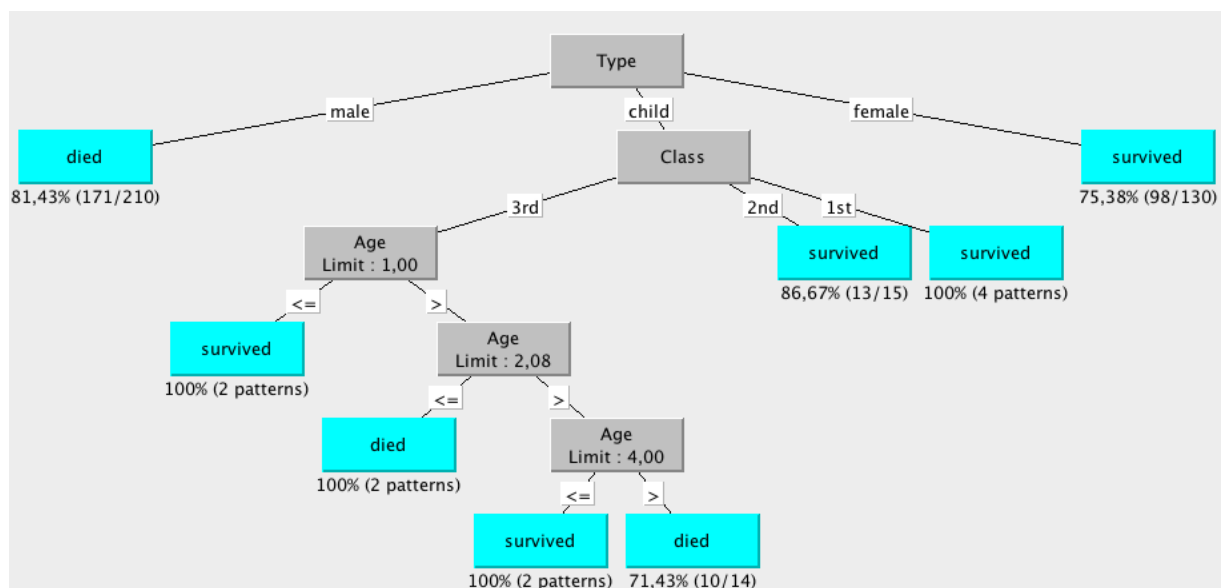


FIGURE 2 – Modèle obtenu sur l'ensemble d'apprentissage de la base de données relative aux passagers du Titanic. Les informations de type "(13/15)" données sous les feuilles précisent le mélange. Le premier nombre correspond au nombre d'individus de la classe et le second au nombre total d'individus.

Répondez aux questions suivantes en vous servant des informations disponibles dans la table 3 et la figure 2.

Q. 1 : De quelle tâche s'agit-il : régression, discrimination, classification, classement ? Justifiez.

Q. 2 : Supprimez-vous des variables ? Justifiez.

Q. 3 : Supprimez-vous des individus ? Justifiez.

Q. 4 : Quel est le taux de bonnes classifications du modèle sur l'ensemble d'apprentissage ?

Q. 5 : Quelle est la probabilité de prédire la bonne classe par hasard ?

Q. 6 : Quelle est la probabilité de prédire la bonne classe en tenant compte des probabilités *a priori* des classes au regard de l'ensemble d'apprentissage ?

Q. 7 : Quelle est la matrice de confusion du modèle sur l'ensemble d'apprentissage ?

Q. 8 : Calculez le rappel et la précision de la classe "survived" obtenus par le modèle d'après l'ensemble d'apprentissage ?

Q. 9 : Quel est l'intérêt du corpus de test ?

Q. 10 : Ce modèle est-il utile ?

Q. 11 : Supposons que l'algorithme utilisé ne gère que les variables quantitatives. Quelles transformations faudrait-il faire ?