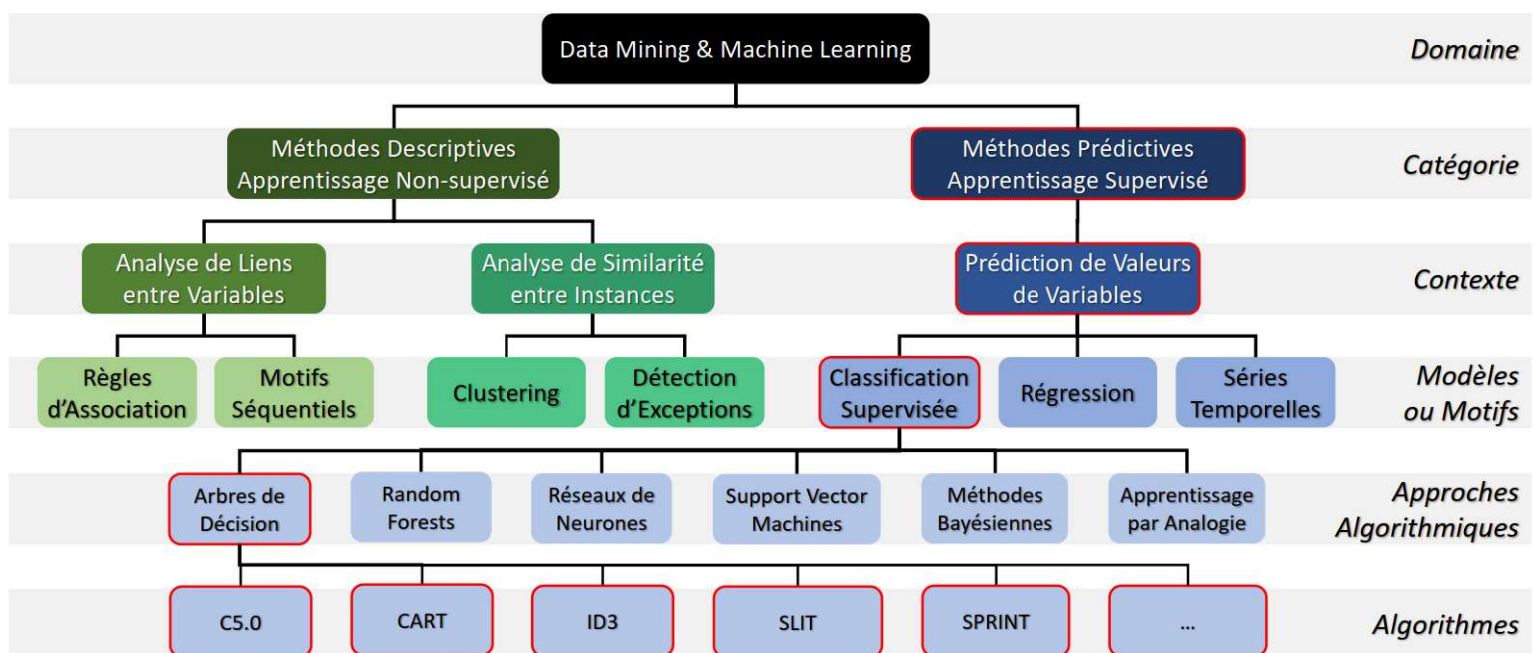


# Apprentissage d'Arbres de Décision

Nicolas PASQUIER  
Université Côte d'Azur  
Département Informatique  
Laboratoire I3S (UMR-7271 UCA/CNRS)  
<http://www.i3s.unice.fr/~pasquier>



## Méthodes d'Extraction de Modèles de Connaissances

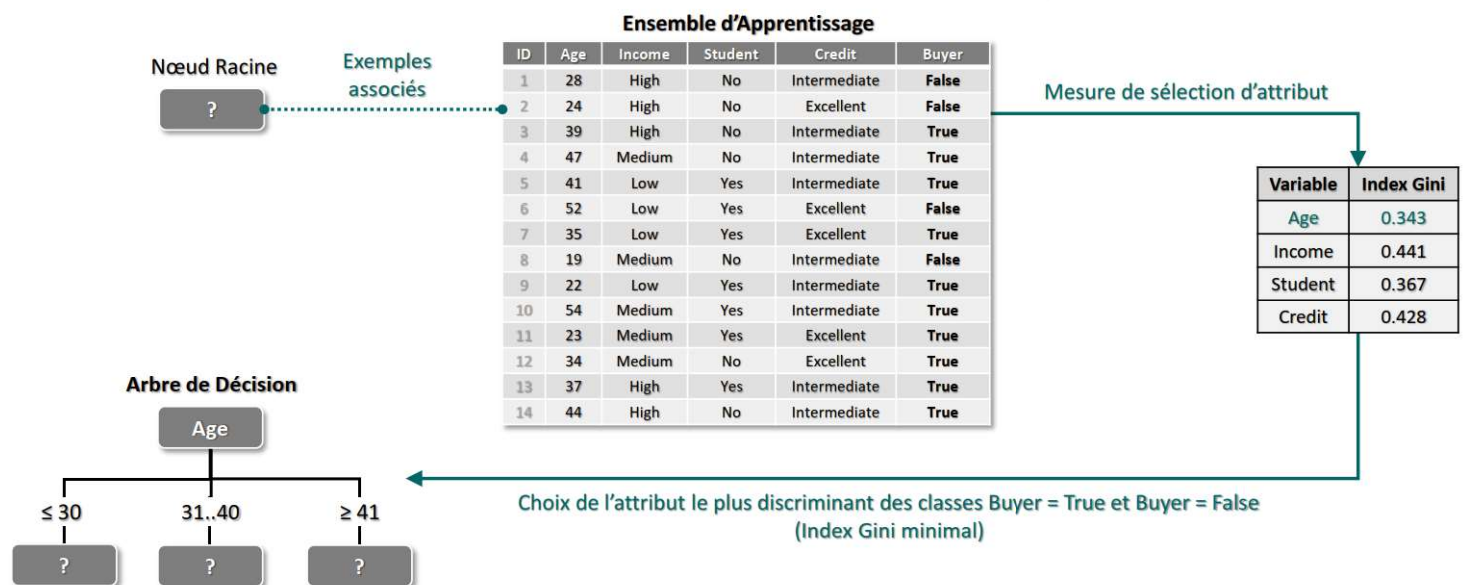


# Configuration Algorithmique : Algorithme et Paramétrage

- Les différentes approches reposent chacune sur une manière (formalisme mathématique) de traiter l'espace des données (probabilités, notion d'hyperplans, etc.)
- Les différents algorithmes et leurs paramètres correspondent chacun à une manière de mettre en œuvre l'approche
  - Paramètres généraux communs aux algorithmes d'une approche
  - Parfois des paramètres spécifiques à l'implémentation
- Chaque configuration algorithmique peut générer un modèle différent
- Apprentissage d'un arbre de décision à partir d'exemples (instances) dont la classe est connue : induction de l'arbre
  - Objectif : apprendre à partir des exemples les critères permettant de déterminer la classe d'un exemple
  - Ces critères sont les tests de valeur des variables qui permettent de discriminer (différencier) les classes

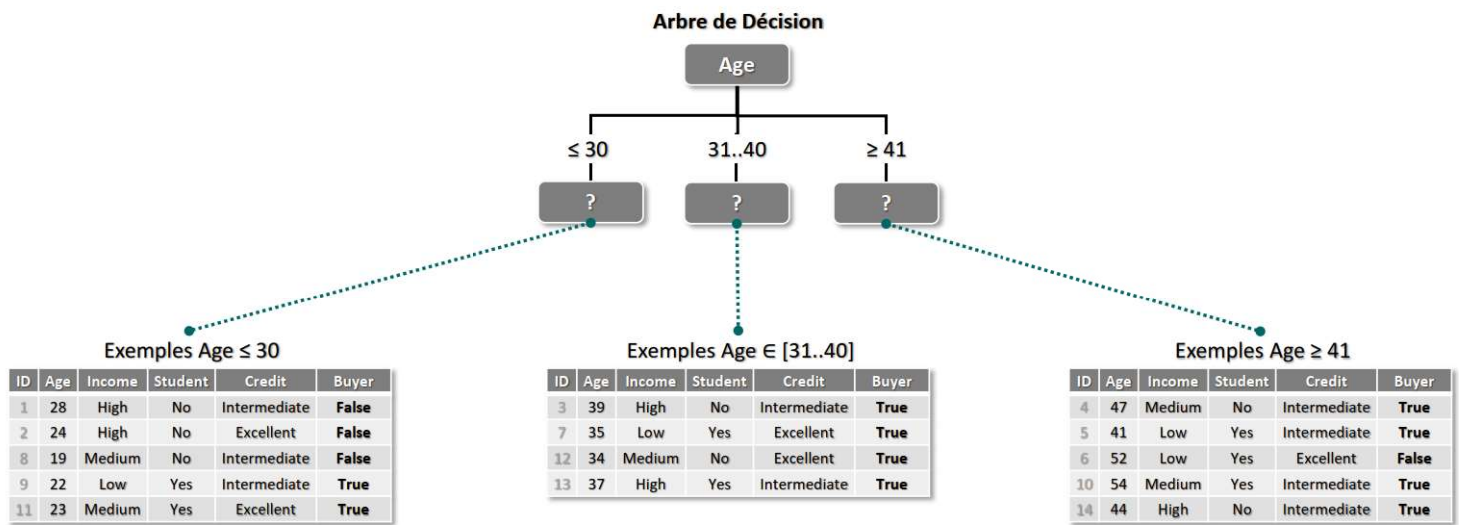
## Exemple d'Induction d'Arbre de Décision

### 1. Création du nœud racine auquel sont associés les exemples d'apprentissage



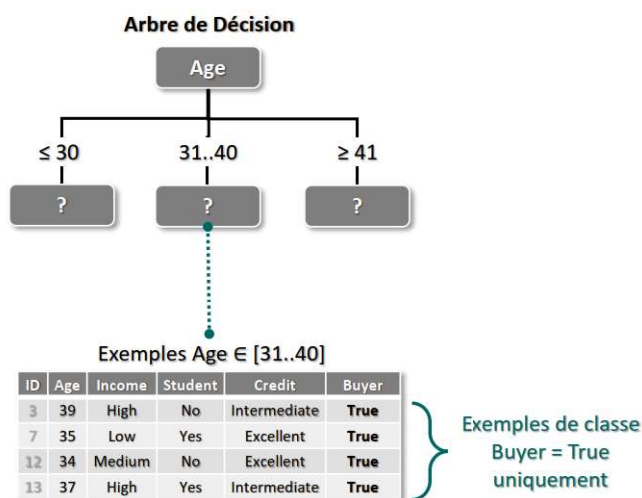
## Exemple d'Induction d'Arbre de Décision

### 2. Association des exemples correspondant à chaque nœuds fils créé



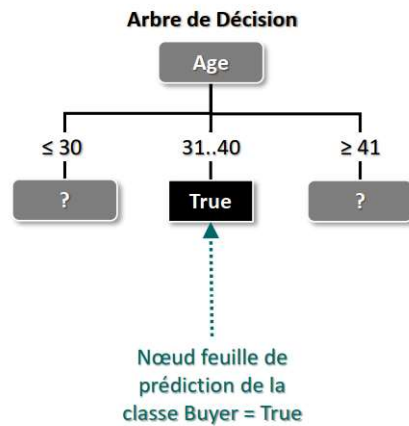
## Exemple d'Induction d'Arbre de Décision

### 3. Pour chaque nœud fils créé, on teste l'unicité de la classe des exemples



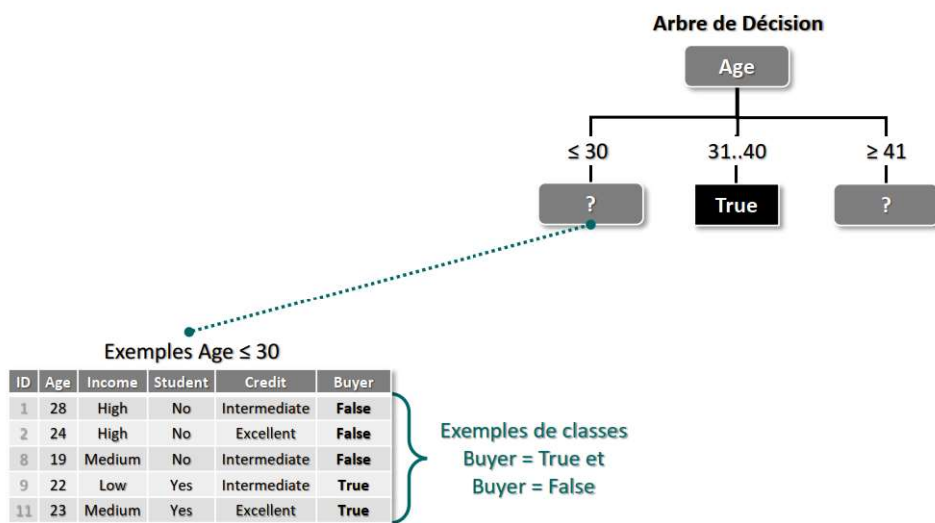
## Exemple d'Induction d'Arbre de Décision

4. Si la classe est unique, le nœud devient un nœud feuille prédisant cette classe



## Exemple d'Induction d'Arbre de Décision

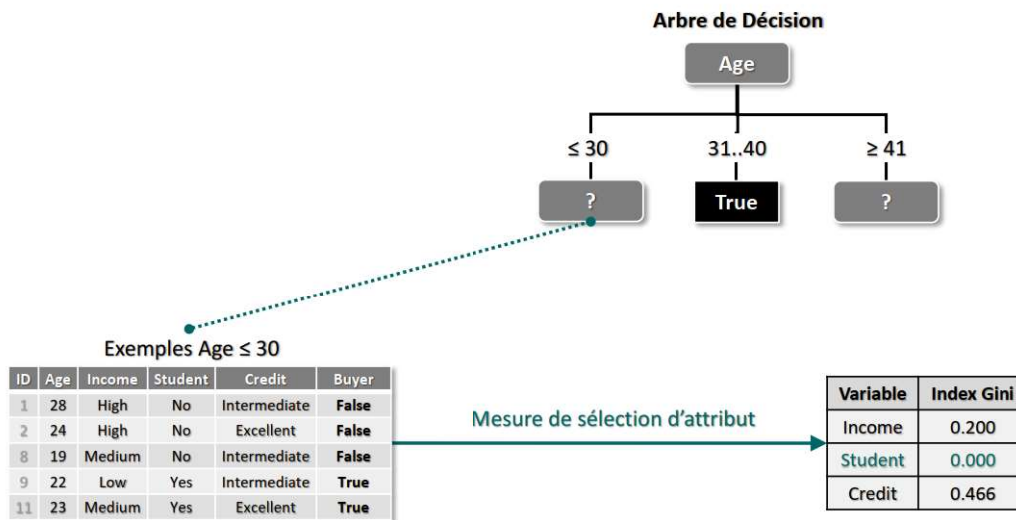
5. Si la classe n'est pas unique, on applique la mesure de sélection d'attribut aux exemples associés au nœud





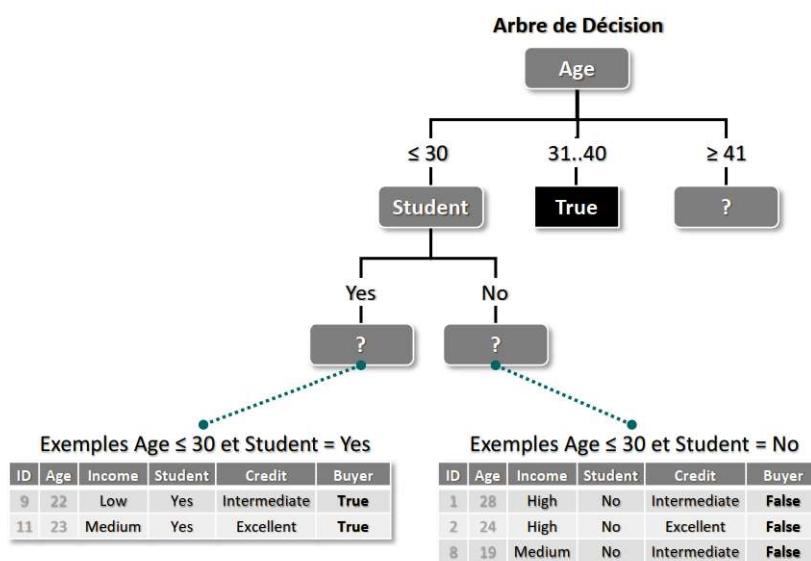
## Exemple d'Induction d'Arbre de Décision

### 6. Choix de l'attribut le plus discriminant identifié par la mesure utilisée



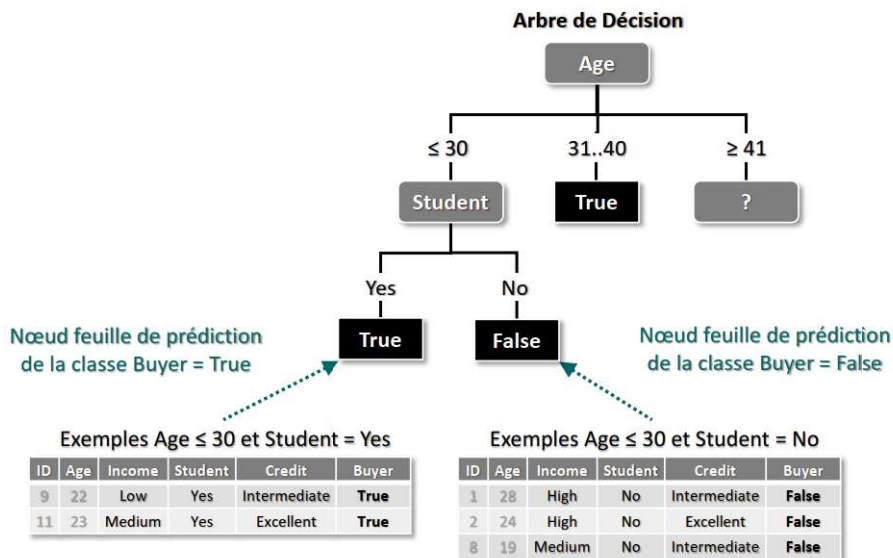
## Exemple d'Induction d'Arbre de Décision

### 7. Création d'un arc et d'un nœud fils pour chaque valeur de l'attribut sélectionné



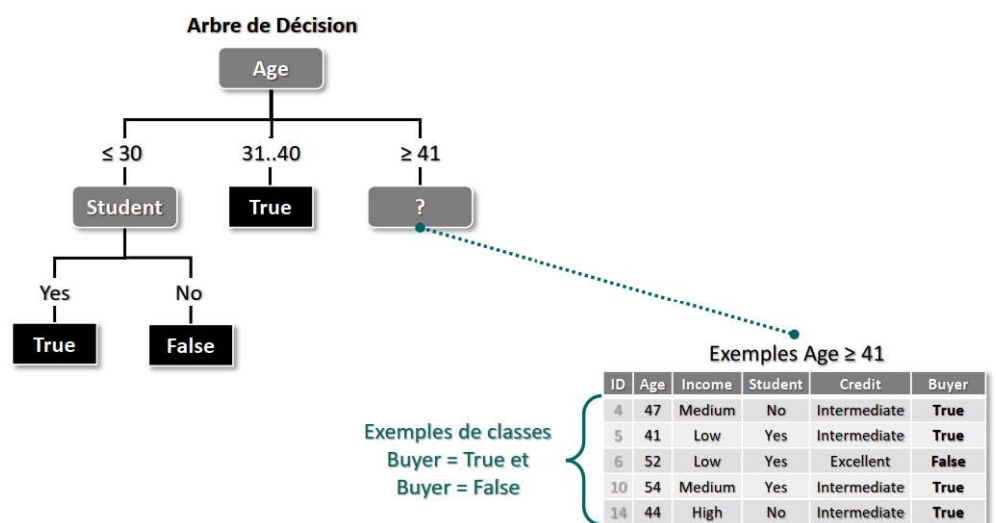
## Exemple d'Induction d'Arbre de Décision

8. Application récursive du processus test de classe unique et sélection d'attribut si ce n'est pas le cas



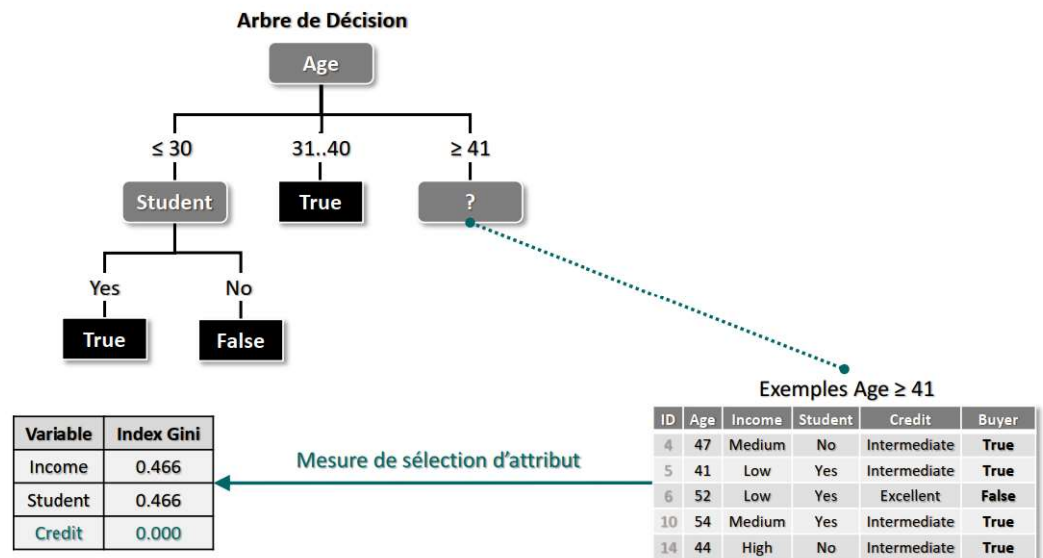
## Exemple d'Induction d'Arbre de Décision

9. Lorsqu'une branche est complète, on passe à la suivante



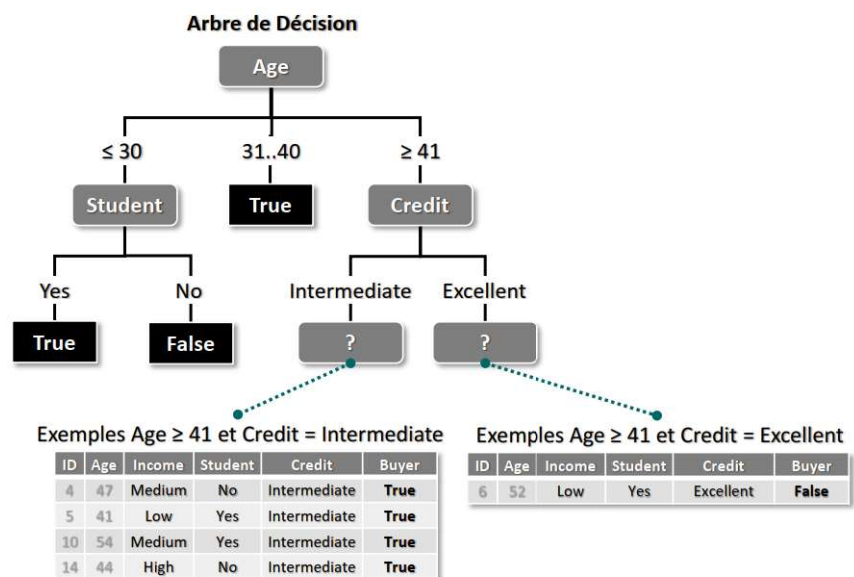
## Exemple d'Induction d'Arbre de Décision

### 10. Application récursive du processus test de classe unique et sélection d'attribut si ce n'est pas le cas



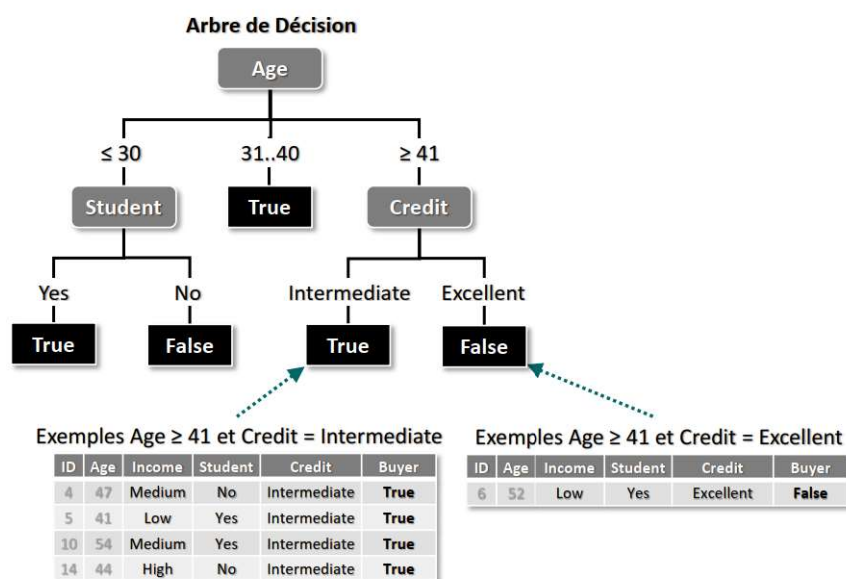
## Exemple d'Induction d'Arbre de Décision

### 11. Application récursive du processus test de classe unique et sélection d'attribut si ce n'est pas le cas



## Exemple d'Induction d'Arbre de Décision

12. L'algorithme s'arrête lorsque tous les exemples d'apprentissage ont été traités



## Variantes Algorithmiques d'Induction d'Arbres de Décision

- Les différents algorithmes (C5.0, CART, etc.) sont des variantes de l'approche
- Arbre binaire (division en deux nœuds fils seulement) ou non
  - Dans le cas d'un arbre non binaire, une variable testée n'est plus considérée par la suite
  - Dans le cas d'un arbre binaire, une même variable sera testée plusieurs fois
- Mesure de sélection d'attribut utilisée par défaut (Index Gini, Information Gain, etc.)
  - Certaines implémentations permettent de paramétrer ce choix
- Critère d'arrêt du développement d'une branche (division en nœuds fils)
  - Effectif minimal d'un nœud feuille (e.g. au moins 10 instances)
  - Effectif minimal d'un nœud pour qu'il soit « divisé » en nœuds fils (e.g. au moins 20 instances)
  - Proportion minimale d'une classe dans le nœud pour qu'il devienne un nœud feuille prédisant cette classe (e.g. 80%)



# Références et Bibliographie

---

- Principales Librairies R
  - [rpart](#) : arbres de décision CART
  - [tree](#) : variante des arbres de décision CART
  - [party](#) : sélection de variables et critère d'arrêt statistiques
  - [C50](#) : arbres de décision C5.0
  - [RWeka](#) : variantes J4.8 de l'algorithme C5.0 et algorithme M5
- Bibliographie
  - R and Data Mining - Examples and Case Studies. Chapter 4 (Decision Trees and Random Forest) and Chapter 13 (Case Study II: Customer Response Prediction and Profit Optimization). Yanchang Zhao. Academic Press, Elsevier, 2012. ISBN 978-0-123-96963-7
  - Data Classification: Algorithms and Applications. Chapter 4 (Decision Trees: Theory and Algorithms). Charu C. Aggarwal. Chapman and Hall/CRC, 2014. ISBN 978-1-466-58674-1