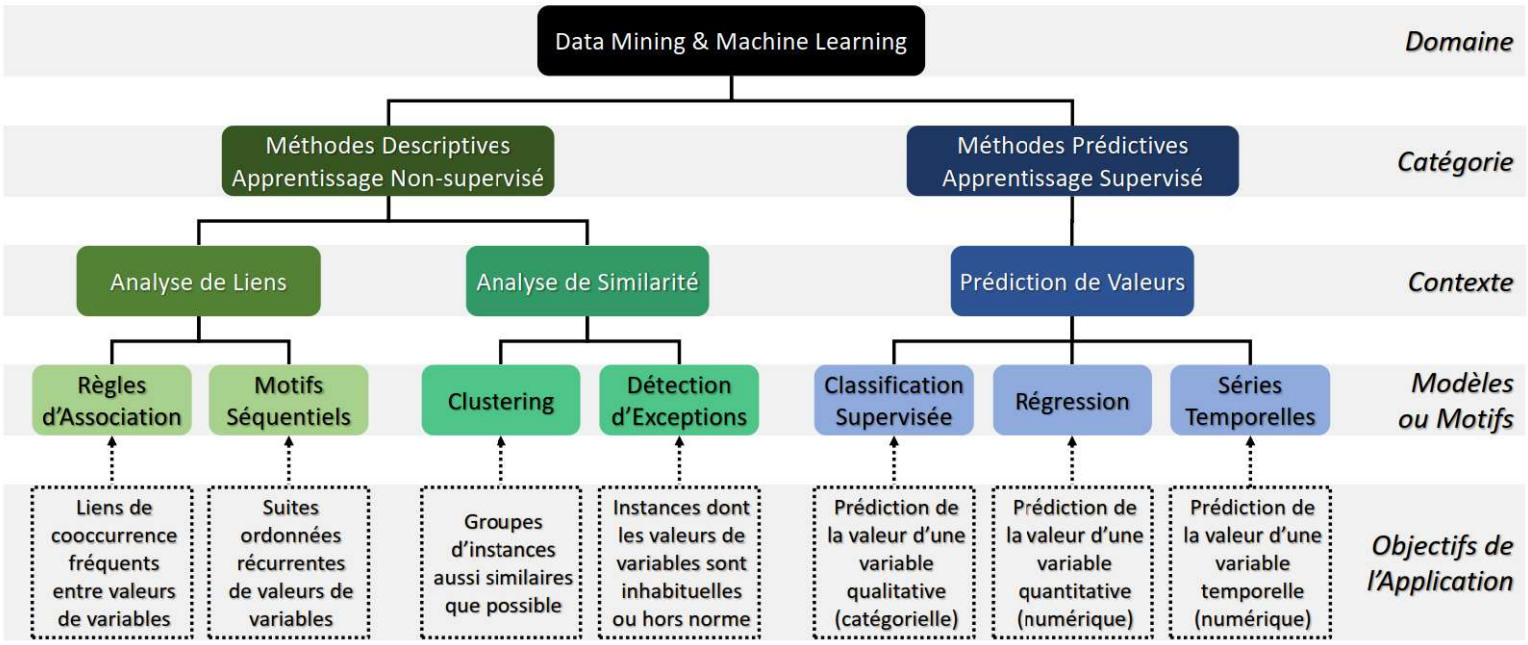
Introduction à la Classification Supervisée

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

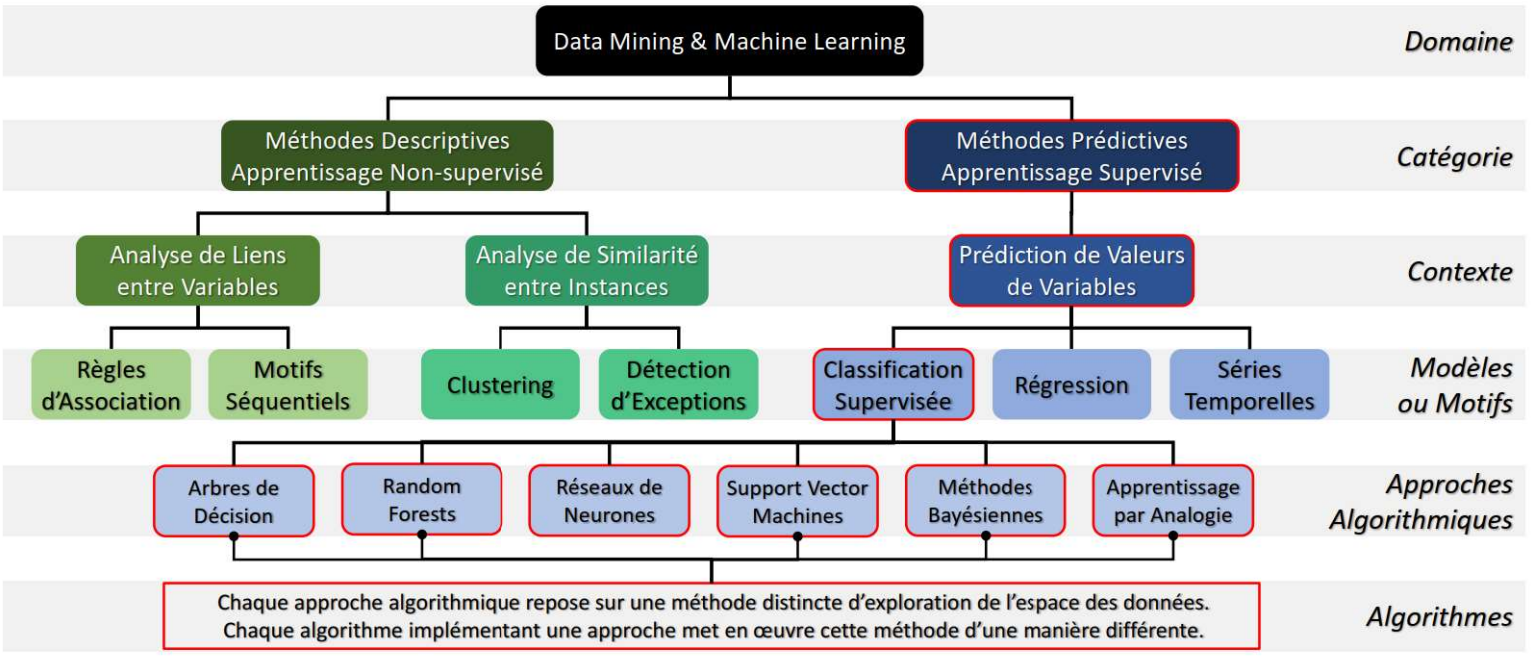
Taxonomie des Méthodes d'Extraction de Modèles de Connaissances



Définition de la Classification Supervisée

- Classification supervisée :
  1. Apprendre un modèle de prédiction de la classe d'exemples (instances) à partir d'exemples dont la classe est connue (expériences passées)
  2. Appliquer le modèle pour prédire la classe de nouveaux exemples de classe inconnue
- Classe d'un exemple : valeur d'une variable qualitative (catégorielle) appelée variable de classe ou variable cible
  - Ex : variable indiquant si le client a acheté le produit qui lui a été proposé
- Classifieur : modèle de prédiction de la classe d'un exemple en fonction de ses caractéristiques représentées par des variables qualitatives (catégorielles) ou quantitatives (numériques) appelées variables prédictives
  - Ex : prédiction de l'achat ou non du produit par le client en fonction de ses caractéristiques sociodémographiques (âge, revenus, etc.)
- L'algorithme apprend le classifieur à partir d'une matrice de données dont les lignes (instances) sont les exemples et les colonnes (attributs) sont les variables prédictives et la variable cible (à prédire)

Méthodes d'Extraction de Modèles de Connaissances



Exemple : Application de Prédiction d'Appétence

- Prédiction de le propension des clients à acheter le produit
- Les exemples de la matrice de données décrivent les clients déjà connus
- La variable Buyer indique si le client a acheté (True) ou non (False) le produit

Dictionnaire des données

Variable	Description	Valeurs
ID	Numéro identifiant du client	[1, 20]
Age	Age en nombre d'année	[19, 54]
Income	Catégorie de revenus	Low, Medium, High
Student	Le client est-il étudiant?	Yes, No
Credit	Capacité d'emprunt du client	Excellent, Intermediate
Buyer	Le client a-t-il acheté le produit?	True, False

Matrice de Données

ID	Age	Income	Student	Credit	Buyer
1	28	High	No	Intermediate	False
2	24	High	No	Excellent	False
3	39	High	No	Intermediate	True
4	47	Medium	No	Intermediate	True
5	41	Low	Yes	Intermediate	True
6	52	Low	Yes	Excellent	False
7	35	Low	Yes	Excellent	True
8	19	Medium	No	Intermediate	False
9	22	Low	Yes	Intermediate	True
10	54	Medium	Yes	Intermediate	True
11	23	Medium	Yes	Excellent	True
12	34	Medium	No	Excellent	True
13	37	High	Yes	Intermediate	True
14	44	High	No	Intermediate	True
15	19	Low	No	Intermediate	False
16	32	Medium	No	Excellent	True
17	47	Medium	No	Excellent	True
18	23	Low	Yes	Intermediate	True
19	40	Medium	No	Excellent	False
20	25	High	Yes	Excellent	True

Exemple : Application de Prédiction d'Appétence

Matrice de Données

ID	Age	Income	Student	Credit	Buyer
1	28	High	No	Intermediate	False
2	24	High	No	Excellent	False
3	39	High	No	Intermediate	True
4	47	Medium	No	Intermediate	True
5	41	Low	Yes	Intermediate	True
6	52	Low	Yes	Excellent	False
7	35	Low	Yes	Excellent	True
8	19	Medium	No	Intermediate	False
9	22	Low	Yes	Intermediate	True
10	54	Medium	Yes	Intermediate	True
11	23	Medium	Yes	Excellent	True
12	34	Medium	No	Excellent	True
13	37	High	Yes	Intermediate	True
14	44	High	No	Intermediate	True
15	19	Low	No	Intermediate	False
16	32	Medium	No	Excellent	True
17	47	Medium	No	Excellent	True
18	23	Low	Yes	Intermediate	True
19	40	Medium	No	Excellent	False
20	25	High	Yes	Excellent	True

- Le classifieur définira un modèle d'affectation d'une classe (prédiction **Buyer=True** ou **Buyer=False**) à un client en fonction de ses caractéristiques (Age, Income, Student et Credit)
- Paramétrage de l'apprentissage du classifieur :
  - Variable cible :

    Buyer variable de classe à prédire
  - Variables prédictives :

    Age, **Income**, **Student**, **Credit** décrivant les caractéristiques testées
  - Variables ignorées :

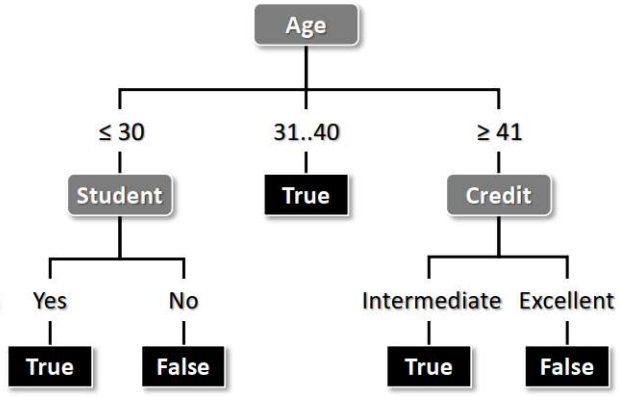
    ID identifiant unique (e.g. nom, prénom, numéro de client, téléphone)

Exemple : Classifieur de Type Arbre de Décision

Matrice de Données

ID	Age	Income	Student	Credit	Buyer
1	28	High	No	Intermediate	False
2	24	High	No	Excellent	False
3	39	High	No	Intermediate	True
4	47	Medium	No	Intermediate	True
5	41	Low	Yes	Intermediate	True
6	52	Low	Yes	Excellent	False
7	35	Low	Yes	Excellent	True
8	19	Medium	No	Intermediate	False
9	22	Low	Yes	Intermediate	True
10	54	Medium	Yes	Intermediate	True
11	23	Medium	Yes	Excellent	True
12	34	Medium	No	Excellent	True
13	37	High	Yes	Intermediate	True
14	44	High	No	Intermediate	True
15	19	Low	No	Intermediate	False
16	32	Medium	No	Excellent	True
17	47	Medium	No	Excellent	True
18	23	Low	Yes	Intermediate	True
19	40	Medium	No	Excellent	False
20	25	High	Yes	Excellent	True

Arbre de Décision



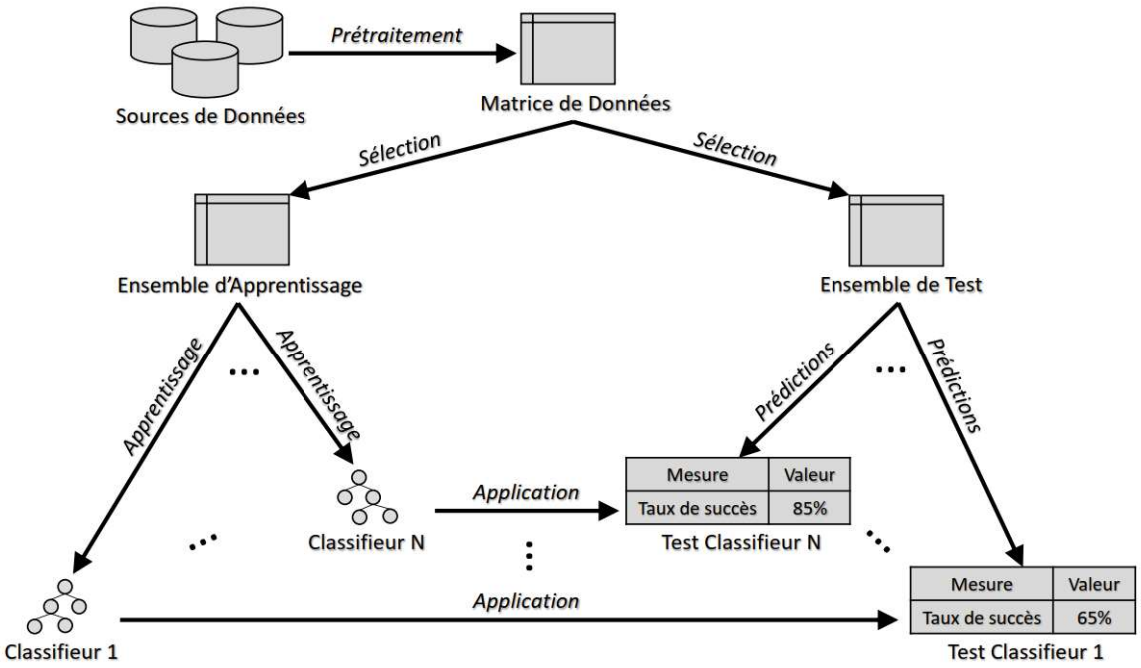
Structure de l'arbre de décision :

- Nœuds internes : variable prédictive testée
- Arcs entre nœuds : valeur (ou intervalle de valeurs numériques) de la variable prédictive testée
- Nœuds feuilles : valeur prédite pour la variable de classe (classe Buyer=True ou Buyer=False)

Processus d'Apprentissage Supervisé

- Identification des variables (cible, prédictives, ignorées)
- Définition des sous-matrices utilisées par l'algorithme d'apprentissage
  - Ensemble d'apprentissage : matrice sur laquelle le classifieur sera appris
  - Ensemble de test : matrice sur laquelle le classifieur sera testé afin d'évaluer ses performances (qualité des prédictions)
- Définition de la configuration algorithmique pour l'apprentissage
  - Choix de l'algorithme
    - Ex : apprentissage d'arbre de décision par l'algorithme C5.0, CART, etc.
  - Définition d'un paramétrage (valeurs des paramètres) pour l'algorithme
    - Ex : profondeur maximale de l'arbre de décision, etc.
- Application de la configuration algorithmique à l'ensemble d'apprentissage
- Test du classifieur appris en comparant la classe prédite avec la classe réelle (classe dans l'ensemble de test) pour tous les exemples de l'ensemble de test

Processus d'Apprentissage et de Test



Exemple : Application de Prédiction d'Appétence

Matrice de Données

ID	Age	Income	Student	Credit	Buyer
1	28	High	No	Intermediate	False
2	24	High	No	Excellent	False
3	39	High	No	Intermediate	True
4	47	Medium	No	Intermediate	True
5	41	Low	Yes	Intermediate	True
6	52	Low	Yes	Excellent	False
7	35	Low	Yes	Excellent	True
8	19	Medium	No	Intermediate	False
9	22	Low	Yes	Intermediate	True
10	54	Medium	Yes	Intermediate	True
11	23	Medium	Yes	Excellent	True
12	34	Medium	No	Excellent	True
13	37	High	Yes	Intermediate	True
14	44	High	No	Intermediate	True
15	19	Low	No	Intermediate	False
16	32	Medium	No	Excellent	True
17	47	Medium	No	Excellent	True
18	23	Low	Yes	Intermediate	True
19	40	Medium	No	Excellent	False
20	25	High	Yes	Excellent	True

Sélection

Sélection

Ensemble d'Apprentissage

ID	Age	Income	Student	Credit	Buyer
1	28	High	No	Intermediate	False
2	24	High	No	Excellent	False
3	39	High	No	Intermediate	True
4	47	Medium	No	Intermediate	True
5	41	Low	Yes	Intermediate	True
6	52	Low	Yes	Excellent	False
7	35	Low	Yes	Excellent	True
8	19	Medium	No	Intermediate	False
9	22	Low	Yes	Intermediate	True
10	54	Medium	Yes	Intermediate	True
11	23	Medium	Yes	Excellent	True
12	34	Medium	No	Excellent	True
13	37	High	Yes	Intermediate	True
14	44	High	No	Intermediate	True

Ensemble de Test

ID	Age	Income	Student	Credit	Buyer
15	19	Low	No	Intermediate	False
16	32	Medium	No	Excellent	True
17	47	Medium	No	Excellent	True
18	23	Low	Yes	Intermediate	True
19	40	Medium	No	Excellent	False
20	25	High	Yes	Excellent	True

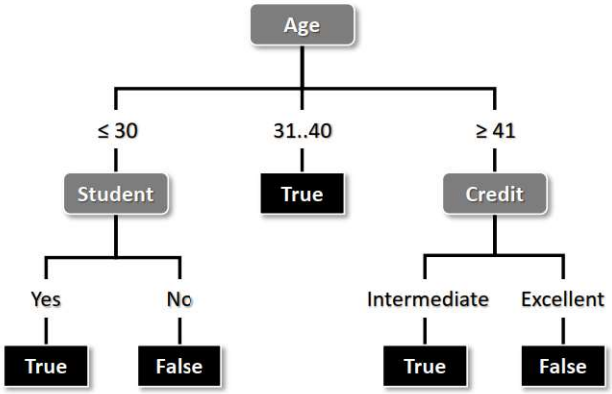
Exemple : Phase d'Apprentissage du Classifieur

Ensemble d'Apprentissage

ID	Age	Income	Student	Credit	Buyer
1	28	High	No	Intermediate	False
2	24	High	No	Excellent	False
3	39	High	No	Intermediate	True
4	47	Medium	No	Intermediate	True
5	41	Low	Yes	Intermediate	True
6	52	Low	Yes	Excellent	False
7	35	Low	Yes	Excellent	True
8	19	Medium	No	Intermediate	False
9	22	Low	Yes	Intermediate	True
10	54	Medium	Yes	Intermediate	True
11	23	Medium	Yes	Excellent	True
12	34	Medium	No	Excellent	True
13	37	High	Yes	Intermediate	True
14	44	High	No	Intermediate	True

Analyse des cooccurrences de valeurs entre les variables prédictives et chacune des deux classes Buyer=True et Buyer=False afin d'identifier les meilleurs critères (valeurs de variables) pour prédire la classe

Arbre de Décision



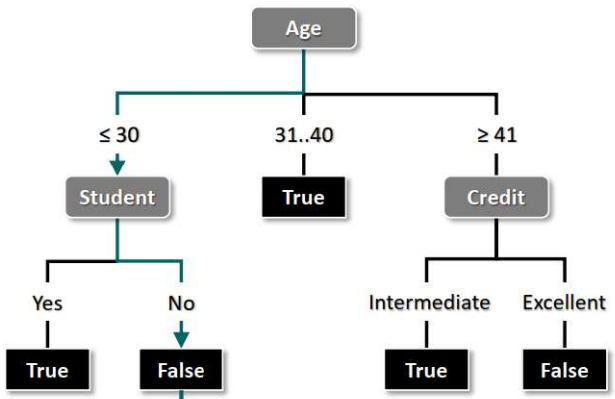
Exemple : Phase de Test du Classifieur

Ensemble de Test

ID	Age	Income	Student	Credit	Buyer	Prediction
15	19	Low	No	Intermediate	False	?
16	32	Medium	No	Excellent	True	?
17	47	Medium	No	Excellent	True	?
18	23	Low	Yes	Intermediate	True	?
19	40	Medium	No	Excellent	False	?
20	25	High	Yes	Excellent	True	?

Identification de la branche (chemin du nœud racine à un nœud feuille) correspondant aux valeurs des variables prédictives qui décrivent l'exemple de test

Arbre de Décision



Ensemble de Test

ID	Age	Income	Student	Credit	Buyer	Prediction
15	19	Low	No	Intermediate	False	False
16	32	Medium	No	Excellent	True	?
17	47	Medium	No	Excellent	True	?
18	23	Low	Yes	Intermediate	True	?
19	40	Medium	No	Excellent	False	?
20	25	High	Yes	Excellent	True	?

Classe prédite

Exemple : Évaluation du Classifieur

- Objectif : évaluer la fiabilité des prédictions sur les exemples de test
- Comparaison pour chaque exemple entre la classe prédite et la classe réelle

Ensemble de Test

ID	Age	Income	Student	Credit	Buyer	Prediction	Test
15	19	Low	No	Intermediate	False	False	Succès
16	32	Medium	No	Excellent	True	True	Succès
17	47	Medium	No	Excellent	True	False	Échec
18	23	Low	Yes	Intermediate	True	True	Succès
19	40	Medium	No	Excellent	False	True	Échec
20	25	High	Yes	Excellent	True	True	Succès

- Comptage des nombres de succès et d'échecs de prédiction
  - 2 échecs : 1 prédiction « True » et 1 prédiction « False » incorrectes
  - 4 succès : 3 prédictions « True » et 1 prédictions « False » correctes
- Estimation de la probabilité de bien ou mal classer un exemple
  - Précision du classifieur (*Classification Accuracy*) = 4/6 ≈ 67%
  - Taux d'erreur (*Error Rate*) = 2/6 ≈ 33%

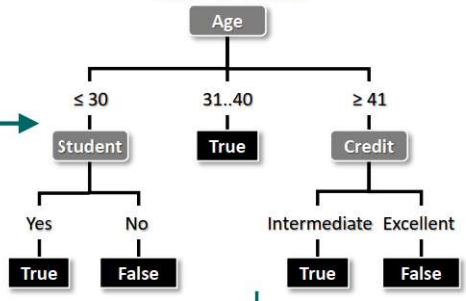
Exemple : Phase de Mise en Œuvre du Classifieur Choisi

Ensemble de Prospects

ID	Age	Income	Student	Credit	Prediction
50	24	High	No	Intermediate	?
51	33	High	No	Excellent	?
52	54	Medium	Yes	Intermediate	?
53	23	Medium	Yes	Excellent	?
54	25	Medium	No	Intermediate	?
55	23	Medium	Yes	Excellent	?
56	23	Low	No	Intermediate	?
57	40	Medium	No	Excellent	?

Prédiction de la classe Buyer=True ou Buyer=False des nouveaux exemples de classe inconnue

Arbre de Décision



Prospects prédits acheteurs probables

ID	Age	Income	Student	Credit	Prediction
51	33	High	No	Excellent	True
52	54	Medium	Yes	Intermediate	True
53	23	Medium	Yes	Excellent	True
55	23	Medium	Yes	Excellent	True
57	40	Medium	No	Excellent	True

Prospects prédits acheteurs improbables

ID	Age	Income	Student	Credit	Prediction
50	24	High	No	Intermediate	False
54	25	Medium	No	Intermediate	False
56	23	Low	No	Intermediate	False

Définition d'une offre commerciale adaptée à la propension d'achat prédite pour chaque client prospect

Ensemble de Prospects

ID	Age	Income	Student	Credit	Prediction
50	24	High	No	Intermediate	False
51	33	High	No	Excellent	True
52	54	Medium	Yes	Intermediate	True
53	23	Medium	Yes	Excellent	True
54	25	Medium	No	Intermediate	False
55	23	Medium	Yes	Excellent	True
56	23	Low	No	Intermediate	False
57	40	Medium	No	Excellent	True

Exemple : Prédiction d'Appétence des Clients Prospects

- Dans le cadre d'une campagne de marketing par e-mail pour l'optimisation des ventes basée sur la prédiction d'appétence (propension à acheter) du client
- Les individus prédits dans la classe Buyer=True pourrons par exemple recevoir un message publicitaire de rappel

Prospects prédits acheteurs probables

ID	Age	Income	Student	Credit	Prediction
51	33	High	No	Excellent	True
52	54	Medium	Yes	Intermediate	True
53	23	Medium	Yes	Excellent	True
55	23	Medium	Yes	Excellent	True
57	40	Medium	No	Excellent	True

- Les individus prédits dans la classe Buyer=False pourront par exemple recevoir une offre promotionnelle (e.g. rabais)

Prospects prédits acheteurs improbables

ID	Age	Income	Student	Credit	Prediction
50	24	High	No	Intermediate	False
54	25	Medium	No	Intermediate	False
56	23	Low	No	Intermediate	False

Références et Bibliographie

- Sites Internet
  - KDNuggets: Business Analytics, Big Data, Data Mining, Data Science, and Machine Learning. https://www.kdnuggets.com/
  - R and Data Mining: Documents, examples, tutorials and resources on R and data mining. http://www.rdatamining.com/
  - CRAN Task View: Machine Learning & Statistical Learning. https://cran.r-project.org/web/views/MachineLearning.html
- Bibliographie
  - R and Data Mining - Examples and Case Studies. Yanchang Zhao. Academic Press, Elsevier, 2012. ISBN 978-0-123-96963-7
  - Data Classification: Algorithms and Applications. Charu C. Aggarwal. Chapman and Hall/CRC, 2014. ISBN 978-1-466-58674-1
  - Data Science : Fondamentaux et Études de Cas – Machine Learning avec Python et R. Éric Biernat, Michel Lutz & Yann LeCun, Eyrolles, 2015. ISBN 978-2-212-14243-3