



An abstract background featuring a dense, colorful pattern of overlapping circles in shades of orange, yellow, green, blue, and purple, creating a textured, liquid-like effect.

ANALYSE ET VISUALISATION DES DONNÉES

Alexandra-Maria DOBRESCU

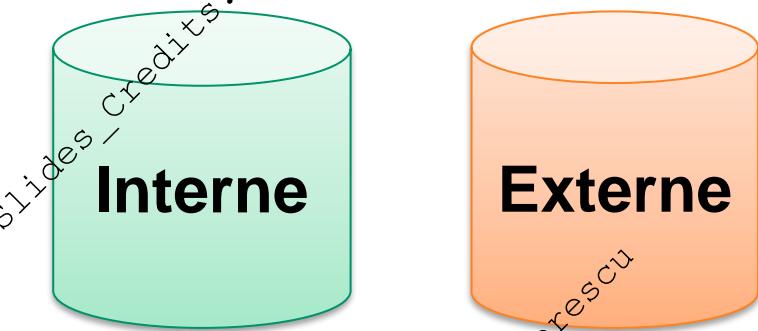
| Cours 4 |

Collection et préparation des données

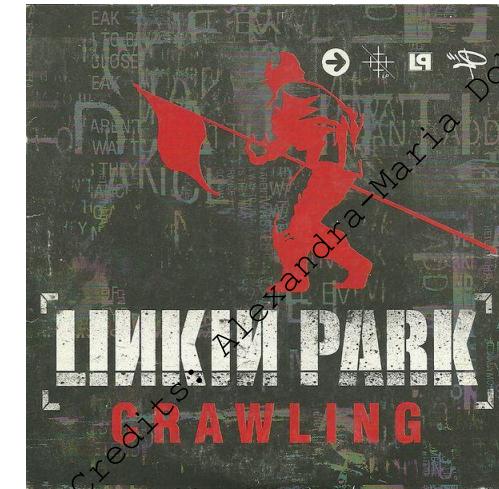
- SOURCES ET ACQUISITION DES DONNÉES
- NETTOYAGE ET PRÉTRAITEMENT DES DONNÉES
- TYPES DE DONNÉES
- LA MESURE

Sources et acquisition des données

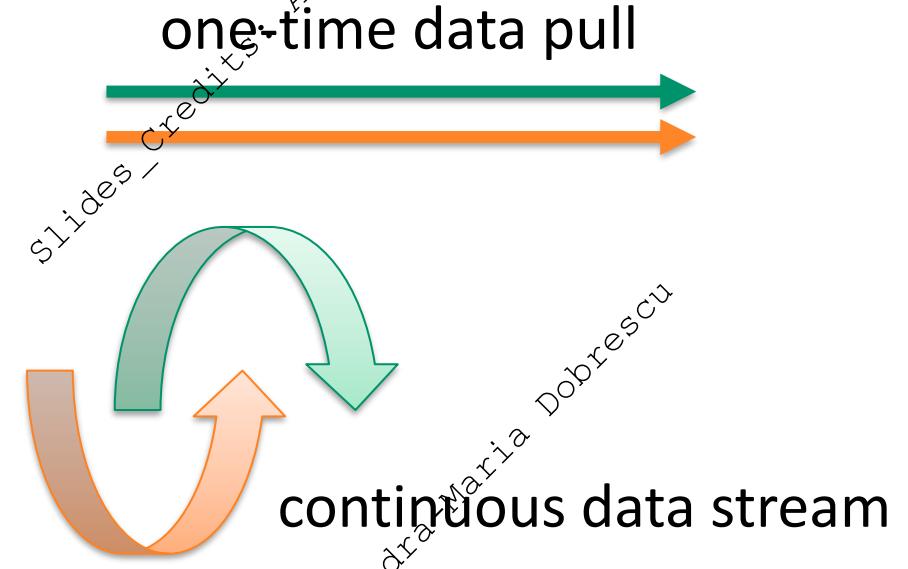
SOURCES DE DONNÉES



- les bases de données,
- le web scraping,
- les senseurs,
- les enquêtes,
- les médias sociaux, etc.



ACQUISITION DE DONNÉES

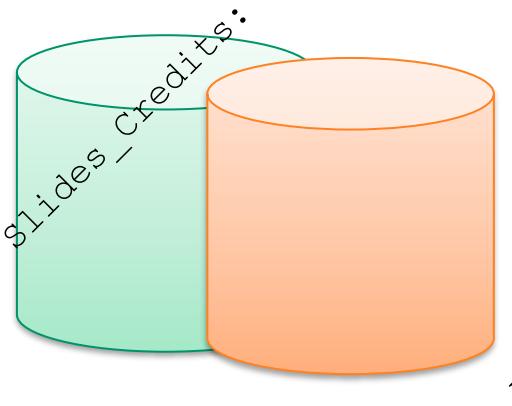


On doit collecter des données à partir:

- des sources choisies,
- en fonction des besoins.

Nettoyage et prétraitement des données

NETTOYAGE DES DONNÉE



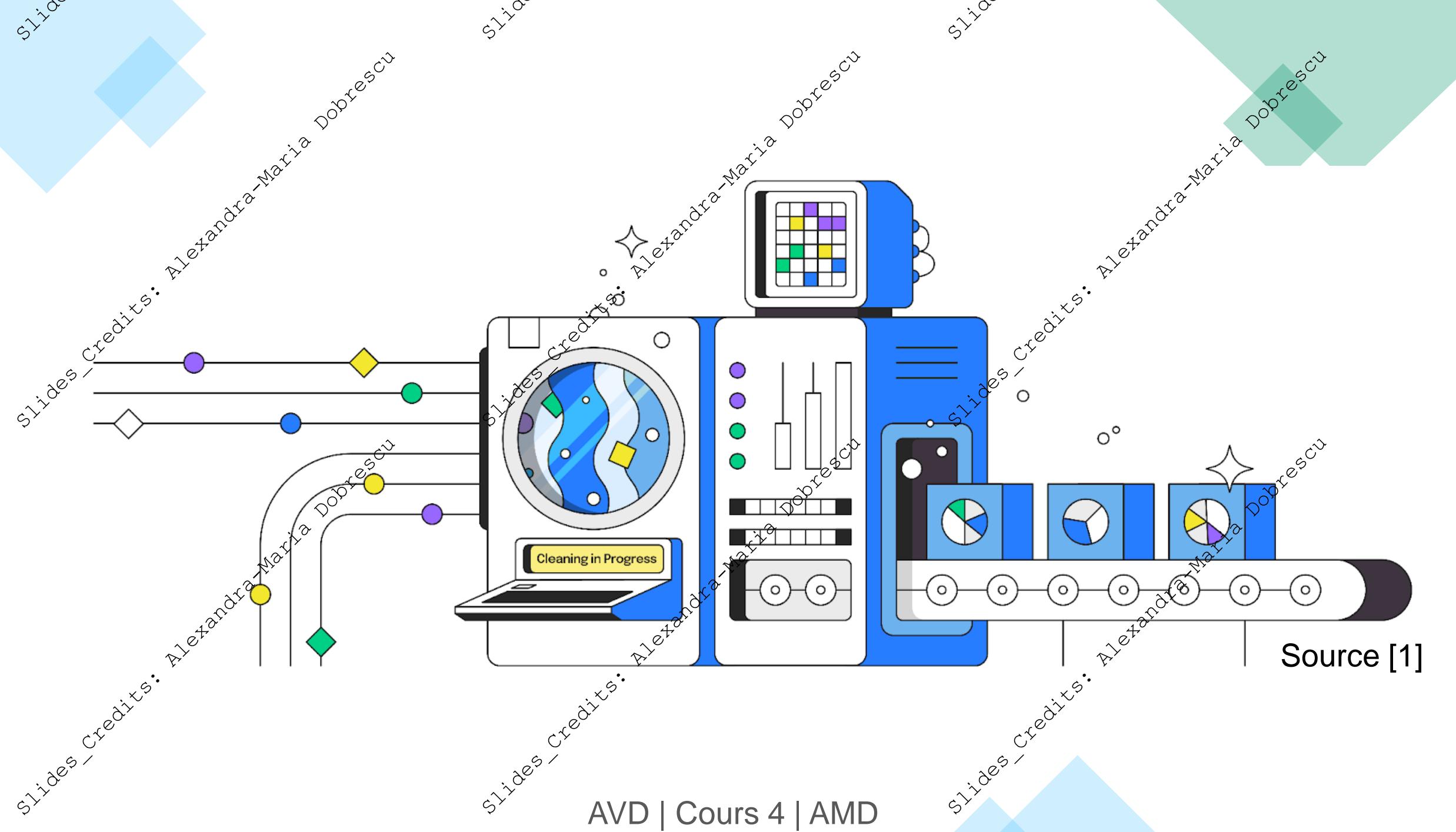
- des erreurs,
- des incohérences,
- des valeurs manquantes

implique

- la suppression des doublons,
- la correction des erreurs,
- la suppression ou le remplacement des valeurs manquantes,
- L'uniformisation des données bruyantes (smooth noisy data),
- la suppression ou l'identification des valeurs aberrantes (outliers),
- La suppression des incohérences.

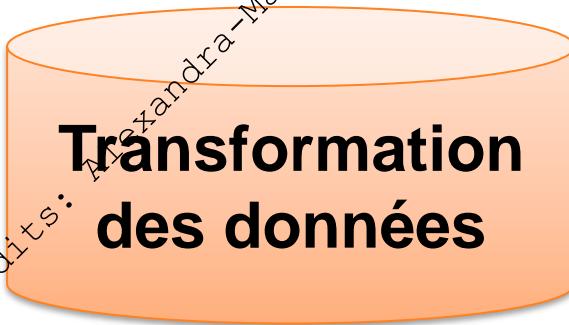
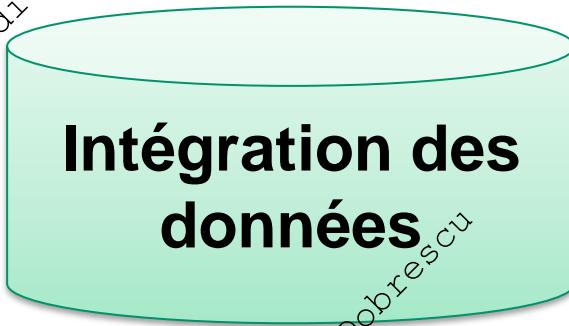
Q: POURQUOI ?

R: POUR GARANTIR LA COHÉRENCE DES DONNÉES.



Nettoyage et prétraitement des données

PRÉTRAITEMENT DES DONNÉES

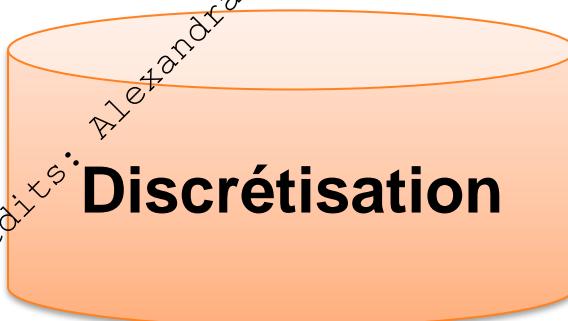


intégration de données provenant de sources multiples, avec éventuellement des types et des structures de données différents, et traitement des données en double ou incohérentes.

normalisation (ou standardisation) des données, résumés, généralisation, construction de nouveaux attributs.

Nettoyage et prétraitement des données

PRÉTRAITEMENT DES DONNÉES



1) Tous les attributs ne sont pas nécessaires pour le processus particulier d'exploration de données que nous voulons effectuer.

2) Seuls **les attributs pertinents** sont sélectionnés pour un traitement ultérieur, ce qui réduit la taille totale de l'ensemble de données (et le temps nécessaire à l'exécution de l'algorithme).

Certains algorithmes ne fonctionnent que sur des données discrètes. C'est pourquoi les valeurs des attributs continus doivent être remplacées par des valeurs discrètes provenant d'un ensemble limité.

Example: L'âge → [jeune, d'âge moyen, vieux]

Discrétisation des données

ALGORITHMES

Classificateur Naïve Bayes

DANS LE DOMAINE DE L'APPRENTISSAGE AUTOMATIQUE ET DE L'EXPLORATION DE DONNÉES (**DATA MINING**), LES CLASSIFICATEURS DE NAÏFS BAYES SONT UNE FAMILLE DE «CLASSIFICATEURS PROBABILISTES» SIMPLES BASÉS SUR L'APPLICATION DU THÉORÈME DE BAYES AVEC DE FORTES HYPOTHÈSES D'INDÉPENDANCE ENTRE LES CARACTÉRISTIQUES.

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)} \text{ ou,}$$

$P(C|X)$ est la probabilité postérieure de la classe C étant donné les caractéristiques de X

$P(C)$ est la probabilité *a priori* de la classe C

$P(X|C)$ est la vraisemblance (*likelihood*) des caractéristiques X pour la classe C

$P(X)$ est la probabilité de la classe X

Discrétisation des données

ALGORITHMES

Classificateur Naïve Bayes

LE CLASSIFICATEUR CALCULE LA PROBABILITÉ A POSTERIORI POUR CHAQUE CLASSE ET ATTRIBUE LA CLASSE AYANT LA PROBABILITÉ LA PLUS ÉLEVÉE COMME CLASSE PRÉDITE POUR L'ENSEMBLE DONNÉ DE CARACTÉRISTIQUES X.

$$\text{postérieure} = \frac{\text{priori} \times \text{la vraisemblance}}{\text{evidence}}$$

Exemple : Dans le cadre de la classification des courriers électroniques indésirables, il est possible de discrétiser des caractéristiques continues telles que le nombre de mots uniques ou la longueur du courrier électronique afin de les adapter à l'algorithme de Naïve Bayes.

Le terme «naïf» est utilisé parce que certaines hypothèses simplificatrices sont formulées.

Discrétisation des données

ALGORITHMES

Arbres de décision

LES ARBRES DE DÉCISION, COMME L'ALGORITHME C4.5, PEUVENT FONCTIONNER AVEC DES DONNÉES CONTINUES ET DISCRÈTES. LA DISCRÉTISATION DES ATTRIBUTS CONTINUS PEUT SIMPLIFIER LA STRUCTURE DE L'ARBRE ET LA RENDRE PLUS FACILE À INTERPRÉTER.

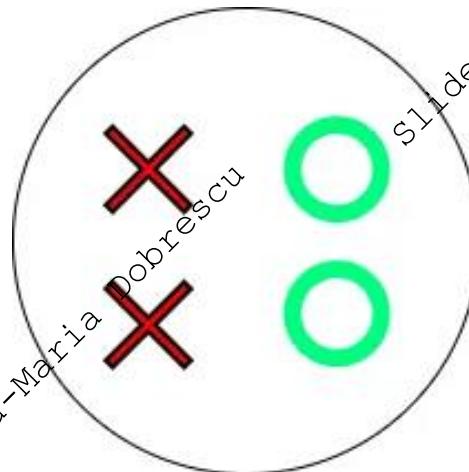
L'entropie de l'information: la mesure de l'impureté d'un exemple donné.

$$\text{Entropie_Shannon} = - \sum_{i=1}^n P(x_i) \log_{\text{base}}(P(x_i))$$

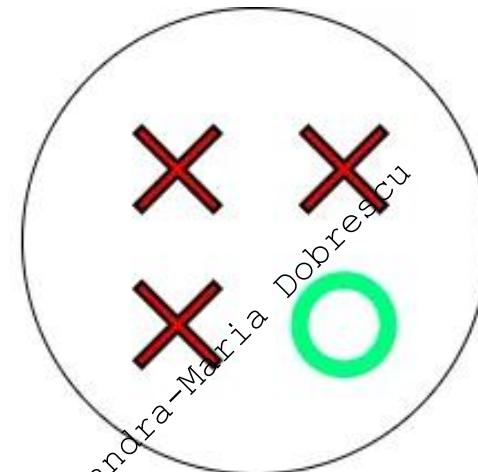
Discrétisation des données

ALGORITHMES

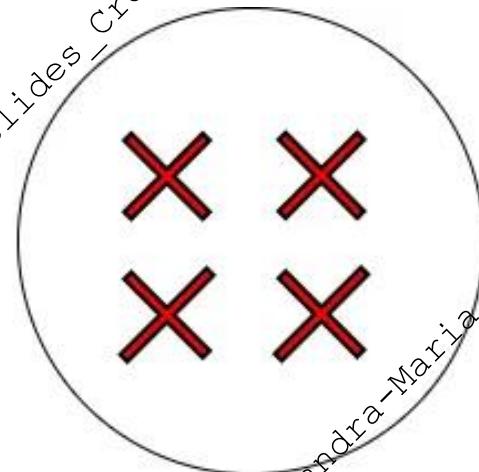
Arbres de décision



Entropy = 1
The sample is impure



Entropy = 0.811
The sample is relatively pure



Entropy = 0
The sample is pure

Source [2]

Discrétisation des données

ALGORITHMES

Arbres de décision

L'entropie d'un ensemble de données est égale à 0 :

- l'ensemble de données est pur
- tous les points de données de l'ensemble appartiennent à la même classe ou catégorie

Il n'y a pas d'incertitude ou de variabilité dans l'ensemble de données, ce qui le rend pur par rapport à la variable cible considérée.

L'entropie augmente:

- plus l'incertitude est grande
- plus les étiquettes de classe sont mélangées dans l'ensemble de données

Discrétisation des données

ALGORITHMES

Arbres de décision

ON VEUT DÉTERMINER LE MEILLEUR ATTRIBUT POUR DIVISER L'ENSEMBLE DE DONNÉES EN SOUS-ENSEMBLES AFIN DE MAXIMISER LE GAIN D'INFORMATIONS, CE QUI SE TRADUIT PAR DES SOUS-ENSEMBLES PLUS PURS À CHAQUE NŒUD DE L'ARBRE.

$$Gain = Entropie(parent) - \sum_{i=1}^n \frac{Count(parent_i)}{Count(parent)} Entropie(parent)$$

Le gain d'information (Information Gain) :

- aide à la sélection des caractéristiques
- Il s'agit d'un paramètre utilisé pour calculer le changement d'entropie d'un ensemble de données avant et après une transformation.

Discrétisation des données

ALGORITHMES

Arbres de décision

**entropie combinée des sous-ensembles
disjoints de l'ensemble de données**

$$Gain = \text{Entropie}(\text{parent}) - \sum_{i=1}^n \frac{\text{Count}(\text{parent}_i)}{\text{Count}(\text{parent})} \text{Entropie}(\text{parent}_i)$$

Discrétisation des données

ALGORITHMES

Arbres de décision

entropie combinée des sous-ensembles
disjoints de l'ensemble de données

$$Gain = \text{Entropie}(\text{parent}) - \sum_{i=1}^n \frac{\text{Count}(\text{parent}_i)}{\text{Count}(\text{parent})} \text{Entropie}(\text{parent}_i)$$

calculée comme une moyenne pondérée

Exercice Classique



...

Attributes				Classes
Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

Source [3]

Discrétisation des données

ALGORITHMES

Arbres de décision

L'algorithme C4.5:

- C'est une version améliorée d'ID3 (développé par Ross Quinlan).
 - **Les attributs sont choisis en fonction du rapport de gain, et non pas simplement sur le gain.**
 - **Post-pruning permet la réduction de la taille de l'arbre ⇔ s'il réduit l'erreur estimée.**
- On peut traiter d'une manière appropriée les valeurs manquantes.

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

Bibliographie

- [1] <https://www.obviously.ai/post/data-cleaning-in-machine-learning>
- [2] https://miro.medium.com/v2/resize:fit:1100/format:webp/1*awa3gie6RSsOQEyOH-CQuw.png
- [3] https://miro.medium.com/v2/resize:fit:1100/format:webp/0*TtRjZOvYr7uSuWQn.jpg