🧩 Astuces Mémo — Objectif 20/20

Pré-traitement des données — Tout mémoriser sans effort

📌 Comment utiliser ce document ?

Ce document contient des **mnémoniques** (trucs pour mémoriser) pour chaque concept du cours de Pré-traitement des données. Le **pré-traitement** c'est tout ce qu'on fait aux données *avant* de les donner à un algorithme de Machine Learning : nettoyer, encoder, mettre à l'échelle, gérer les trous, etc.

🏷️ ENCODAGE — Transformer le texte en chiffres

Les modèles ML ne comprennent que les chiffres. L'encodage convertit les catégories textuelles en nombres.

Concept	Astuce mémo
Ordinale = a un ordre	Ordinale = Ordre — même début ! (Bac < Licence < Master)
Nominale = pas d'ordre	Nominale = juste un nom, une étiquette.
One-Hot = plein de colonnes	Imagine un hôtel 🏨 : chaque modalité a sa propre chambre (colonne).
150 modalités → PAS One-Hot	150 chambres = trop cher → Target Encoding
Target Encoding	Target = Trop de modalités. Les deux commencent par T.
Label sur nominal = PIÈGE	Rouge=1, Bleu=2, Vert=3 → le modèle croit Vert > Rouge. Faux !

🔍 OUTLIERS — Valeurs anormalement éloignées

Un outlier est une valeur qui s'écarte fortement du reste. Ex : tout le monde gagne 3000-8000 DH sauf un qui gagne 500 000 DH.

Concept	Astuce mémo
IQR = Q3 - Q1	I Q uitte la R oute — un outlier sort de la route !
Bornes = ± 1.5×IQR	Retiens juste "1.5" — c'est toujours 1.5.
Z-score = (X-μ)/σ	Z éro au centre — à combien de pas (σ) tu es du centre (μ).
Z > 3 = outlier	À 3 pas du groupe = perdu = outlier.
Outlier légitime	→ R especter → R obustScaler.
Outlier illégitime	= erreur → poubelle 🗑️

❓ VALEURS MANQUANTES — Les trous dans les données

Des cases vides dans le dataset. La question clé : POURQUOI sont-elles vides ?

- = "M'en e" →
- = "il che vers quelqu'un" → dépend d'un
- = "M cissique" → dépend de

Concept	Astuce mémo
MCAR	C = Complètement aléatoire. Bug capteur, café renversé.
MAR	A = dépend d'une Autre variable (ex: l'âge).
MNAR	N = Non aléatoire. Riches cachent leur revenu = manque à cause de lui-même.
Taux < 5%	→ Supprimer les lignes.
Taux 5-30%	→ Imputer (moyenne, médiane, mode, KNN).
Taux > 30%	→ Variable indicatrice (0/1) + modèle prédictif.
Taux > 90%	→ Recoder en binaire.
Variable numérique	→ Moyenne (symétrique) ou médiane (asymétrique).
Variable catégorielle	→ Mode. JAMAIS la moyenne sur du texte !

📏 SCALING — Mettre tout sur la même échelle

Si l'âge va de 0 à 80 et le revenu de 0 à 500 000, le revenu écrase l'âge. Le scaling remet tout au même niveau.

- : (X−min)/(max−min) → entre
- : (X−μ)/σ → centré en
- : (X−médiane)/IQR →

Concept	Astuce mémo
Quand scaler ?	"KSK" — NN, VM, -Means = algos à distance → obligatoire.
Arbre / Random Forest	Pas besoin — coupent par seuils, pas de distances.
Min-Max sensible	Un seul outlier détruit le min ou max → tout l'échelle bouge.
RobustScaler	obust = ésiste. Médiane + IQR.

📊 ACP — Compresser les variables

L'ACP transforme N variables corrélées en 2-3 nouvelles variables résumant l'essentiel. Comme une photo 2D d'un objet 3D.

Concept	Astuce mémo
Quand ?	"CRV" — Corrélation + Réduction + Visualisation.
Règle des 80%	Comme une note de passage — cumuler les PC jusqu'à passer 80%.
PC1 = la plus grande	PC1 = patron 👔. PC2 = adjoint. PC3 = stagiaire.
Perte d'info	= 100% − variance cumulée retenue.
Standardiser avant	Sinon variable en millions domine.
Perte interprétabilité	PC1 = smoothie 🥤 : tu sais plus ce qu'il y a dedans.

🔧 FEATURE ENGINEERING — Créer de nouvelles variables

Inventer de nouvelles colonnes à partir des existantes pour aider le modèle.

Type	Astuce	Exemple
Ratio	Diviser deux variables	prix / surface = prix_m2
Date	Extraire depuis une date	mois, jour_semaine, ancienneté
Groupe	Agrégation par catégorie	moyenne_ventes_par_magasin
Binaire	Transformer en oui/non	est_weekend, est_rez_de_chaussée

📝 PIPELINE NLP — Traiter du texte pour le ML

Les modèles ML ne lisent pas. Le pipeline NLP transforme du texte brut en chiffres exploitables.

« ettoie on alon, ers e in »
ettoyage → okenisation → top words → temming/ emma → ectorisation

Concept	Astuce
Tokenisation	Token = jeton 🎰 — couper le texte en petits jetons.
Stop words	Mots "stop" 🛑 — inutiles (le, la, de, est).
Stemming	Hache 🪓 (rapide, brutal) : "mangeons" → "mang".
Lemmatisation	Scalpel 🔪 (précis, lent) : "mangeons" → "manger".
TF-IDF	Très fréquent ici, incroyablement dur à trouver ailleurs.

🚨 Anti-sèche — Les pièges de l'exam

Piège	Comment l'éviter
Confondre MAR et MNAR	MAR = **A**utre. MNAR = **N**arcissique.
Moyenne sur du catégoriel	Moyenne de "Rouge, Bleu" = aucun sens → **MODE**.
Oublier standardiser avant ACP	Toujours mentionner "il faut standardiser avant".
Normalisation vs Standardisation	**N**orm = [0,1]. **S**tand = **S**igma (Z-score).
One-Hot sur 150 modalités	150 colonnes = explosion → **Target Encoding**.
"supprimer" quand taux > 30%	Trop de perte → **imputer ou recoder**.