Guide Pratique des Transformations de Données

Data Science - Aide-mémoire
Stratégie

Janvier 2026

1 Logique de Décision (Le "Cheat Sheet")

Avant de coder, utilisez ce tableau pour choisir la transformation appropriée :

Type de Variable	Condition / Problème	Transformation	Exemple Projet
Numérique Continue	Pas d'outliers / Bornes fixes	**Min-Max Scaling**	Score crédit (0-100)
Numérique Continue	Distribution Normale	**Standardisation**	Taille, Poids
Numérique Continue	**Outliers importants**	**Robust Scaling**	Salaire, Fortune
Catégorielle	Hiérarchie Logique	**Ordinal Encoding**	Niveau d'étude, Satisfaction
Catégorielle	Pas de hiérarchie	**One-Hot Encoding**	Ville, Sexe, Couleur
Numérique	Relations non-linéaires	**Discrétisation**	Tranches d'âge, Revenus

2 Mise à l'échelle (Scaling)

2.1 Normalisation (Min-Max Scaling)

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Exemple : Un score d'examen de 0 à 100. **Pourquoi ?** Utile quand on veut que toutes les données soient dans l'intervalle $[0, 1]$. Très utilisé en Deep Learning.

2.2 Standardisation (Z-score)

$$x_{std} = \frac{x - \mu}{\sigma}$$

Exemple : La taille humaine (en cm). **Pourquoi ?** On centre sur 0 avec un écart-type de 1. Indispensable pour l'ACP car elle traite toutes les variables avec la même importance statistique sans être bloquée par des unités différentes.

2.3 Robust Scaling (Quartiles)

$$x_{robust} = \frac{x - \text{médiane}}{Q_3 - Q_1}$$

Exemple : Le prix des maisons. **Pourquoi ?** La médiane et l'IQR ne sont pas influencés par les valeurs extrêmes (le château à 50M€ n'impacte pas le calcul du "centre" des données).

3 Transformations Catégorielles

3.1 Encodage Ordinal

Exemple : Taille de vêtement (S, M, L, XL). On attribue $S = 0, M = 1, L = 2, XL = 3$. **Règle :** L'ordre mathématique doit refléter une réalité (Grandeur, Niveau, Priorité).

3.2 One-Hot Encoding

Exemple : Ville (Paris, Lyon, Marseille). **Règle :** On crée N colonnes binaires. On ne peut pas dire que Lyon (2) est plus grand que Paris (1). Sans cela, le modèle inventerait une hiérarchie imaginaire.

4 Discrétisation (Binning)

Exemple : L'âge. Au lieu d'utiliser 18, 19, 20... on crée des sacs (Bins) : "Junior", "Senior". **Pourquoi ?** Parfois, la différence entre 32 et 33 ans est nulle, mais la différence entre "Actif" et "Retraité" est majeure pour une banque.

5 Analyse Exploratoire des Données (EDA)

L'EDA est le processus systématique pour comprendre la structure et la qualité des données avant la modélisation.

1. **Data Profiling :** Vérification des dimensions, des types de colonnes (numérique vs texte) et du taux de valeurs manquantes (NaN).
2. **Analyse Univariée :** Étude de la distribution de chaque variable (Histogrammes pour le numérique, diagrammes en barres pour le catégoriel).
3. **Analyse Bivariée :** Recherche de corrélations (Matrice de corrélation) et de relations entre les caractéristiques et la cible (y).
4. **Détection d'Outliers :** Utilisation de Boxplots pour identifier les points extrêmes.

6 Traitement des Valeurs Manquantes (Imputation)

Lorsqu'une donnée manque, trois stratégies principales s'offrent au Data Scientist :

6.1 Suppression (Deletion)

— **Quand ?** Si moins de 5% des données manquent et que c'est aléatoire, ou si une colonne est vide à plus de 60%.
— **Risque :** Perte d'information importante si les données manquantes ne sont pas dues au hasard.

6.2 Imputation Statistique (Simple)

Méthode	Type de Variable	Cas d'usage
Moyenne	Numérique	Distribution normale, pas d'outliers.
Médiane	Numérique	Présence d'outliers (méthode robuste).
Mode	Catégorielle	Remplacer par la catégorie la plus fréquente.
Valeur Fixe	Toutes	Créer une catégorie "Inconnu" ou mettre à 0.

6.3 Imputation Avancée (Algorithmique)

— **KNN Imputer :** On remplace la valeur manquante par la moyenne des k voisins les plus proches.
— **Iterative Imputer :** Utilise un modèle de régression pour prédire la valeur manquante en fonction des autres colonnes.

7 Exemples d'Imputation : Quand faire quoi ?

— **Exemple 1 (Salaire) :** S'il manque 2 salaires sur 1000, et qu'il y a des multimillionnaires, on utilise la **médiane**.
— **Exemple 2 (Ville) :** Si la ville manque, on utilise le **mode** (la ville la plus représentée) ou on note "Inconnu".
— **Exemple 3 (Capteur IoT) :** Si un capteur de température s'éteint 1 min, on fait une **interpolation** (moyenne entre la valeur juste avant et juste après).

7.1 Gestion du Risque lié aux Données Creuses

Pour les variables présentant un taux de valeurs manquantes supérieur à 10% (ex : Ville, Niveau d'étude), nous rejetons l'imputation par le mode qui introduirait un biais de fréquence artificiel.
— **Catégories "Inconnu" :** Le manque d'information est traité comme une modalité à part entière. Cela permet au système expert de capturer une corrélation éventuelle entre la rétention d'information et le risque de crédit.
— **Missing Indicators (Variables Fantômes) :** Pour les données numériques, chaque imputation est doublée d'une variable binaire indiquant l'absence initiale de la donnée, préservant ainsi l'intégrité du signal pour l'algorithme.

8 Perspective : Optimisation par Imputation Algorithmique (EM)

Dans une phase d'évolution du *Système Expert*, **nous pourrions envisager** l'intégration d'une stratégie d'imputation multivariée basée sur l'algorithme **Expectation-Maximization (EM)** ou l'**Iterative Imputer**. Cette approche permettrait de ne plus traiter les données manquantes comme des points isolés, mais comme des variables dépendantes du contexte global du client.

8.1 Critères de déclenchement d'une telle imputation

L'implémentation de cette couche algorithmique **pourrait être conditionnée** par le profil de vacuité du dataset :
— **Seuil de Robustesse (5% - 40%) :** Si une variable présente un taux de données manquantes significatif, **on pourrait privilégier** l'algorithme EM pour reconstruire le signal sans introduire le biais de centralité propre à la médiane.
— **Cohérence Multidimensionnelle :** Cette méthode **pourrait être activée** prioritairement pour les variables corrélées (ex : estimer un *Salaire_Annuel* manquant en fonction du *Niveau_Etude* et de l'*Age* observés).

8.2 Apports potentiels pour le Scoring

L'adoption de cette méthode **permettrait** au *Système Expert* de :

1. **Réduire le bruit statistique :** En évitant d'injecter des valeurs "artificielles" (moyennes) qui écrasent les variances.
2. **Affiner la séparation des classes :** Une imputation plus proche de la réalité terrain **pourrait faciliter** le travail de l'ACP et des modèles de classification pour distinguer les profils à risque.