# 📝 Examen de Rattrapage — Pré-traitement des données

2ème année Génie Informatique et Digitalisation — Pr. YOUNES DAOUI

Durée : 1h30 | Documents non autorisés

## 📌 De quoi parle ce document ?

Le **Pré-traitement des données** est l'ensemble des étapes qu'on applique aux données brutes **avant** de les donner à un algorithme de Machine Learning. Un modèle ML ne peut pas travailler avec des données sales (valeurs manquantes, incohérentes, pas à la même échelle...). Ce document est un **examen simulé** avec son corrigé détaillé, couvrant les 6 grands thèmes du cours :

- **Encodage** : transformer les catégories (texte) en chiffres exploitables
- **Outliers** : détecter et traiter les valeurs aberrantes
- **Valeurs manquantes** : gérer les trous dans les données
- **Scaling** : mettre toutes les variables sur la même échelle
- **ACP** : réduire le nombre de variables en gardant l'essentiel
- **Feature Engineering** : créer de nouvelles variables intelligentes

## Q1. Encodage des variables catégoriques (3 points)

Vous disposez d'un dataset contenant les variables suivantes :

- `niveau_etude` : {"Bac", "Licence", "Master", "Doctorat"}
- `couleur_produit` : {"Rouge", "Bleu", "Vert", "Jaune", "Noir", "Blanc"}

1. Quel type d'encodage proposez-vous pour **chaque** variable ? Justifiez. **(1,5 pts)**
2. Appliquez l'encodage choisi sur les 3 premières valeurs de chaque variable et montrez le résultat sous forme de tableau. **(1,5 pts)**

## Q2. Détection et traitement des outliers (3 points)

On vous donne le dataset suivant (variable `salaire` en milliers de DH) :

| Obs. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Salaire | 5 | 7 | 6 | 8 | 5 | 50 | 7 | 6 |

1. Calculez Q1, Q3 et l'IQR. **(1 pt)**
2. Déterminez les bornes inférieures et supérieures. **(0,5 pt)**
3. Identifiez le(s) outlier(s) et proposez **deux** stratégies de traitement. **(1,5 pts)**

## Q3. Valeurs manquantes — Analyse et traitement (4 points)

### Situation A — Dataset hospitalier

La variable `pression_arterielle` est manquante à 8%.

- Les données manquantes sont **totalement aléatoires**, sans lien avec aucune autre variable.
- Les valeurs existantes suivent une distribution **normale**.

### Situation B — Dataset hospitalier

La variable `traitement_prescrit` (catégorielle) est manquante à 25%.

- La probabilité qu'un traitement soit manquant dépend de l'**âge du patient**.
- La variable contient 5 modalités.

### Situation C — Dataset hospitalier

La variable `revenu_patient` est manquante à 45%.

- Les patients à **faible revenu** sont ceux qui ne déclarent pas leur revenu.

Pour chaque situation :

1. Identifiez le **type de valeur manquante** (MCAR, MAR ou MNAR). **(1,5 pts)**
2. Proposez une **méthode de gestion** adaptée et justifiez. **(2,5 pts)**

## Q4. Mise à l'échelle — Scaling (3 points)

Vous avez un dataset avec les variables suivantes :

| Variable | Min | Max | Moyenne | Écart-type |
|---|---|---|---|---|
| Âge | 18 | 75 | 35 | 12 |
| Revenu (DH) | 2000 | 500 000 | 45 000 | 85 000 |
| Score_credit | 300 | 850 | 620 | 110 |

1. Expliquez la différence entre la **Normalisation (Min-Max)** et la **Standardisation (Z-score)**. Donnez la formule de chacune. **(1,5 pts)**
2. Si vous utilisez un algorithme **KNN**, quel scaling choisissez-vous ? Justifiez. **(0,75 pt)**
3. Appliquez la **standardisation** sur la valeur `Âge = 50`. **(0,75 pt)**

## Q5. ACP — Analyse en Composantes Principales (4 points)

Un data scientist travaille sur un dataset de **satisfaction client** contenant 8 variables. Après analyse, il observe que plusieurs variables sont **fortement corrélées** et souhaite **visualiser** les données en 2D.
Après application d'une ACP :

| Composante | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 |
|---|---|---|---|---|---|---|---|---|
| Variance | 45% | 25% | 15% | 8% | 4% | 2% | 0.7% | 0.3% |

1. Combien de composantes pour expliquer **≥ 80%** de la variance ? **(1 pt)**
2. Quelle est la **perte d'information** avec PC1 + PC2 seulement ? **(1 pt)**
3. Expliquez l'utilité de l'ACP, ses **avantages** et ses **limites**. **(2 pts)**

## Q6. Feature Engineering (3 points)

Vous disposez d'un dataset de **locations immobilières** avec : `surface_m2`, `nb_chambres`, `prix_loyer`, `ville`, `etage`, `date_construction`, `date_mise_en_location`.

1. Proposez au moins **4 nouvelles features pertinentes**. **(2 pts)**
2. Justifiez **pourquoi** chacune améliorerait le modèle. **(1 pt)**

## Barème — Total : /20 points

| Question | Thème | Points |
|---|---|---|
| Q1 | Encodage | 3 |
| Q2 | Outliers | 3 |
| Q3 | Valeurs manquantes | 4 |
| Q4 | Scaling | 3 |
| Q5 | ACP | 4 |
| Q6 | Feature Engineering | 3 |
| **Total** | | **20** |

# 📗 CORRIGÉ DÉTAILLÉ

# ✅ Corrigé Q1 — Encodage

## 1. Type d'encodage (1,5 pts)

`niveau_etude` → **Label Encoding ordonné**

- Cette variable est **ordinale** — il y a un ordre naturel : Bac < Licence < Master < Doctorat
- Le Label Encoding respecte cette hiérarchie en attribuant des entiers croissants

`couleur_produit` → **One-Hot Encoding**

- Cette variable est **nominale** — il n'y a pas d'ordre entre les couleurs
- Le Label Encoding créerait une **fausse relation d'ordre** (Rouge=0 < Bleu=1 < Vert=2)
- Avec 6 modalités (< 100), le One-Hot reste raisonnable en termes de dimensionnalité

## 2. Application (1,5 pts)

`niveau_etude` (Ordinal) :

| Valeur | Encodé |
|---|---|
| Bac | 0 |
| Licence | 1 |
| Master | 2 |

`couleur_produit` (One-Hot) :

| Valeur | Rouge | Bleu | Vert | Jaune | Noir | Blanc |
|---|---|---|---|---|---|---|
| Rouge | 1 | 0 | 0 | 0 | 0 | 0 |
| Bleu | 0 | 1 | 0 | 0 | 0 | 0 |
| Vert | 0 | 0 | 1 | 0 | 0 | 0 |

## ✅ Corrigé Q2 — Outliers

Données triées : 5, 5, 6, 6, 7, 7, 8, 50

### 1. Calcul (1 pt)

- Q1 (1er quartile) = médiane de {5, 5, 6, 6} = (5+6)/2 = 5,5
- Q3 (3ème quartile) = médiane de {7, 7, 8, 50} = (7+8)/2 = 7,5
- IQR = Q3 − Q1 = 7,5 − 5,5 = 2

### 2. Bornes (0,5 pt)

- Borne inf. = Q1 − 1,5 × IQR = 5,5 − 3 = 2,5
- Borne sup. = Q3 + 1,5 × IQR = 7,5 + 3 = 10,5

### 3. Outlier + stratégies (1,5 pts)

Outlier : 50 (car 50 > 10,5)

Stratégie 1 — Suppression : supprimer l'observation, car 50 est très éloigné et faussera par la suite l'analyse.

Stratégie 2 — Capping : remplacer 50 par la borne supérieure (10,5). Cela conserve l'observation tout en limitant l'impact de la valeur extrême.

## ✅ Corrigé Q3 — Valeurs manquantes

### Situation A — `pression_arterielle` (8%)

**Type : MCAR** (Missing Completely At Random)

- Les données manquantes sont totalement aléatoires, sans lien avec aucune variable ni avec la valeur elle-même.

**Méthode : Imputation par la moyenne**

- Taux faible (8%) et distribution normale → la moyenne est un bon estimateur central.
- Le fait que ce soit MCAR garantit que l'imputation ne biaise pas les résultats.
- Alternative : suppression des lignes (acceptable car 8% seulement).

### Situation B — `traitement_prescrit` (25%, catégorielle)

**Type : MAR** (Missing At Random)

- La probabilité de valeur manquante dépend d'une autre variable observée (l'âge), et non de la valeur manquante elle-même.

**Méthode : Imputation par le mode conditionnel** (mode par groupe d'âge)

- Comme la donnée manquante dépend de l'âge, on groupe par tranches d'âge et on impute le **mode** (valeur la plus fréquente) de chaque groupe.
- 25% est trop élevé pour supprimer les lignes.
- C'est une variable catégorielle → on **ne peut pas** utiliser la moyenne.

### Situation C — `revenu_patient` (45%)

**Type : MNAR** (Missing Not At Random)

- La probabilité de valeur manquante dépend de la valeur elle-même — les patients à faible revenu ne déclarent pas.

**Méthode : Variable indicatrice + modèle prédictif**

- Créer `revenu_manquant` (0/1) — l'absence est elle-même une information précieuse.
- Imputer via un modèle (KNN ou régression) — les méthodes simples (moyenne, médiane) seraient biaisées car les valeurs manquantes ne sont pas aléatoires.

- Taux de 40% → la suppression est impossible

## ✅ Corrigé Q4 — Scaling

### 1. Différence et formules (1,5 pts)

**Normalisation (Min-Max) :**

$$X_{norm} = (X - X_{min}) / (X_{max} - X_{min}) \rightarrow \text{résultat dans } [0, 1]$$

- Sensible aux outliers (min et max tirés par les extrêmes)

**Standardisation (Z-score) :**

$$X_{std} = (X - \mu) / \sigma \rightarrow \text{centré en 0, écart-type de 1}$$

- Plus robuste que le Min-Max, adaptée aux distributions normales

### 2. Choix pour KNN (0,75 pt)

**Standardisation (Z-score)** — KNN utilise des calculs de distance. Sans scaling, la variable Revenu (2000–30 000) dominerait complètement le calcul de distance par rapport à Âge (18–75). Le scaling est obligatoire pour que chaque variable contribue équitablement.

### 3. Standardisation de Âge = 50 (0,75 pt)

$$X_{std} = (50 - 35) / 12 = 15 / 12 = 1{,}25$$

→ L'âge 50 est situé 1,25 écart-type au-dessus de la moyenne.

## ✅ Corrigé Q5 — ACP

### 1. Nombre de composantes pour ≥ 80% (1 pt)

| Composantes retenues | Variance cumulée |
|---|---|
| PC1 | 45% |
| PC1 + PC2 | 70% |
| PC1 + PC2 + PC3 | 85% ✅ |

→ Il faut retenir **3 composantes** pour dépasser 80%.

### 2. Perte d'information avec PC1 + PC2 (1 pt)

Variance retenue = 45% + 25% = 70%

**Perte = 100% − 70% = 30%**

### 3. Utilité, avantages et limites (2 pts)

**Utilité :** 8 variables fortement corrélées → redondance. L'ACP réduit de 8 à 2-3 variables en combinant les variables corrélées.

**Avantages :**

- **Réduction dimensionnelle** : 8 → 3 variables pour en garder 85% de l'info
- **Élimination de la multi-colinéarité** : les composantes sont orthogonales (non-corrélées)
- **Visualisation** : projection en 2D/3D possible
- **Performance** : réduit le surapprentissage

**Limites :**

- **Perte d'interprétabilité** : chaque PC est un mélange de toutes les variables
- **Linéarité** : ne capture que les relations linéaires
- **Sensibilité aux outliers et à l'échelle** : il faut standardiser avant

## ✅ Corrigé Q6 — Feature Engineering

| # | Feature | Formule | Justification |
|---|---|---|---|
| 1 | `prix_par_m2` | `prix_loyer / surface_m2` | Normalise le prix par la surface → permet de comparer des biens de tailles différentes, indicateur clé du marché immobilier |
| 2 | `ancienneté_batiment` | `date_location − date_construction` | L'âge du bâtiment influence son état et donc le loyer. Un bâtiment neuf → loyer plus élevé |
| 3 | `surface_par_chambre` | `surface_m2 / nb_chambres` | Taille moyenne des pièces. Grandes chambres → loyer plus élevé |
| 4 | `est_rez_de_chaussee` | `1 si etage = 0 sinon 0` | Le RDC a souvent un loyer différent des étages supérieurs. Discrétisation d'une |

💡 Autres features acceptables : `etage_eleve` , `ville_encodee` , `annee_construction` , `mois_mise_en_location`

## 📌 Conseils pour le rattrapage

> ⚠️ **Thèmes les plus probables :**
>
> 1. **Encodage** — Distinguer Label, One-Hot, Target Encoding
> 2. **Outliers** — Calcul IQR + Z-score, 3 stratégies de traitement
> 3. **Valeurs manquantes** — MCAR / MAR / MNAR et la méthode adaptée
> 4. **Scaling** — Formules Min-Max et Z-score par cœur
> 5. **ACP** — Interpréter variance cumulée, avantages et limites
> 6. **Feature Engineering** — Créer des features (ratio, date, groupe, binaire)
> 7. **Pipeline NLP** — 6 étapes dans l'ordre avec le rôle de chacune

> 🚨 **Q3 (valeurs manquantes, 4 pts) et Q5 (ACP, 4 pts)** portent le plus de points. Ce sont les questions les plus discriminantes — préparez-les en priorité !