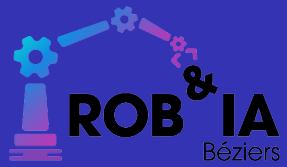
Recherche de variables descriptives/explicatives Sartorius



R4.13 : Étude séries temporelles

February 13, 2025

Retour sur la confidentialité des données

- Dans le cas où vous utilisez un PC de l'IUT, vérifier que les données ne sont pas directement accessibles pour d'autres utilisateurs.
- Penser à arrêter VS Code correctement (Close Folder).
- Ne jamais stocker les données sur un cloud public non sécurisé.
- Vérifier que les scripts et notebooks ne contiennent pas de données sensibles en clair.
- Supprimer les fichiers temporaires après usage.

Méthodologie d'analyse des données

- Collecte organisation des données (pandas R_.07)
- Réduction de la dimension → NS-Learning (ACP)
- Clustering si aucune piste → NS-Learning (recherche soit de clusters soit d'éléments isolés KMeans puis t-SNE /UMAP)
- Si étiquetage possible, Classification ou Régression possible → S-Learning (MLP par exemple)
- Si beaucoup de données, on peut essayer un LLM (Large Language Model) = Deep Learning mais plus technique et gourmand en ressource.

Retour sur les données tabulées en Machine Learning

Feature

Sample

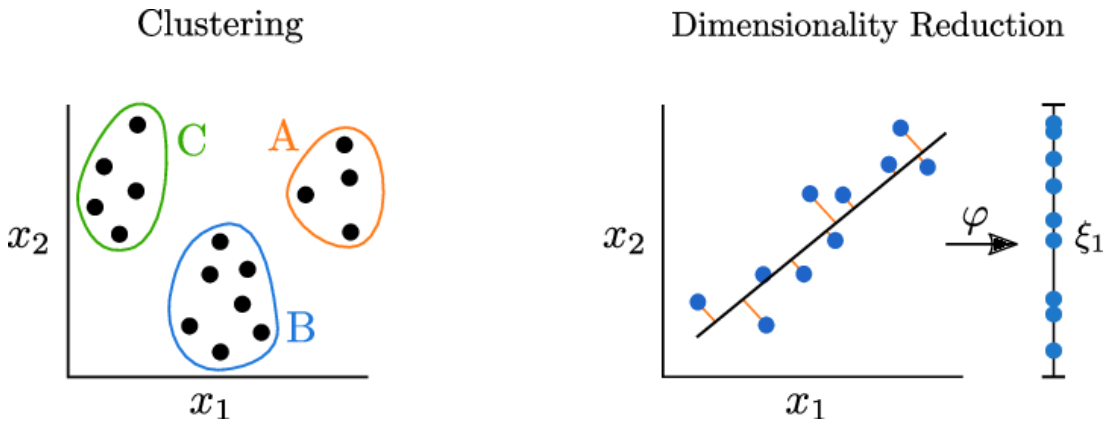
$$\mathbf{X} = \begin{bmatrix} 1.2 & 3.4 & 5.6 & \cdots & 2.1 \\ 4.5 & 1.2 & 3.3 & \cdots & 0.9 \\ 2.3 & 4.4 & 1.1 & \cdots & 3.2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 3.1 & 2.2 & 4.5 & \cdots & 1.7 \end{bmatrix}$$

$\rightarrow$ Un jeu de données est défini par une matrice comportant n lignes et m colonnes numériques.

Et pour notre jeu de données ?

Apprentissage non supervisé

- Données non étiquetées.
- Objectifs :
  - Réduction de dimension.
  - Clustering.



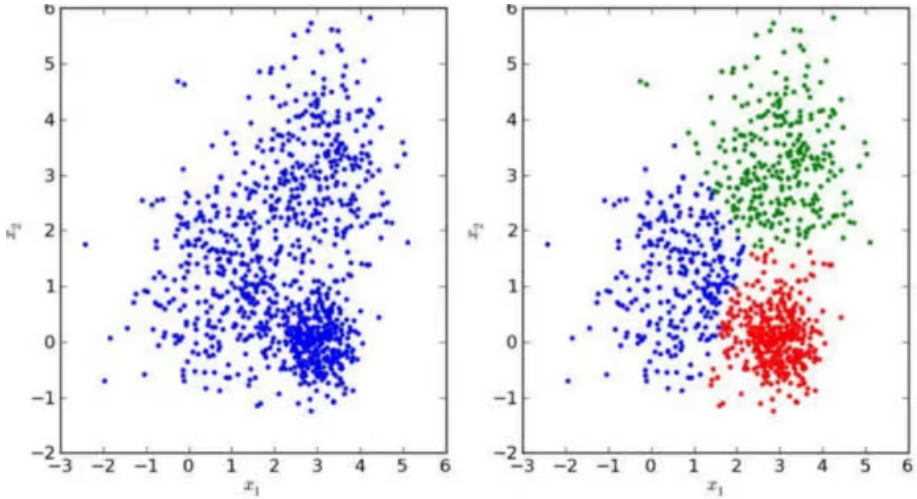
Stratégie pour notre jeu de données : Identifier les variables descriptives/explicatives corrélées si possible à une intervention de maintenance.

Analyse en Composantes Principales (ACP)

- Technique de réduction de dimension utilisée pour extraire l'information essentielle d'un jeu de données tout en minimisant la perte d'information.
- transforme un jeu de données d'un espace de m dimensions à p dimensions (avec $p < m$), tout en conservant un maximum de la variance des données d'origine.
- Étapes :
  1. Centrage-Réduction des données
  2. Calcul de la matrice de covariance
  3. Diagonalisation $\rightarrow$ (valeurs propres : variance expliquée, vecteur propre : axes principaux)
- Lien avec la ressource R3.13 de J. Azé.

Apprentissage supervisé (autre piste)

- Classification et régression.
- Exemple pour notre cas :
  - Étiquetage possible : *fonctionnement*, *arrêt normal*, *panne* extrait de GMAO.



  - Utilisation de KNN pour identifier les instances qui sont trop éloignées de leur cluster ou d'un MLP pour mapper les 60 variables sur une sortie *intervention curative* (True/False).

Travail à faire pour la prochaine fois

- Créer un notebook intitulé **identification_variables_descriptives_Sartorius.ipynb** à déposer d'ici au 3/03.
- Travailler sur le DataFrame produit dans les dernières séances.
- Présentation des éléments du DataFrame : shape, dtypes, columns, head(), etc.
- Mise en forme pour traitement ACP : pas de données object ou str, uniquement int ou float.
- Travail similaire à R3.13 sur les données météo (Scikit-learn, puis Fanalysis).
- Analyse des résultats et questions pour le commanditaire.
- **Attention** : l'utilisation d'une IA générative est déconseillée, l'objectif étant d'appliquer la méthodologie vue en TP.