
République Islamique de la Mauritanie
Groupe polytechnique

Institut Supérieure des Métiers de la Statistique

# RAPPORT DE DATA MINNING

Pour l'obtention de la Licence Professionnelle en Sciences de Données et Informatique Décisionnelle (SDID)

Rédiger par :

Hindou Bebaye Boubi (23656)

Soukeina Sedatt (23634)

Sous l'encadrement de :

El Hadrami Bouleryah

**Année universitaire : 2025-2026**

# MINI-RAPPORT — GROUPE 7

## Détection de la dérive des données (Data Drift)

---

# 1. Introduction

Dans les systèmes de surveillance et de prédiction basés sur les données, la stabilité des distributions d'entrée est un élément clé pour garantir la performance des modèles de machine learning dans le temps. Cependant, dans un contexte réel, les données évoluent naturellement à cause de changements de comportement, de conditions externes ou de phénomènes saisonniers.
Ce phénomène est connu sous le nom de **dérive des données (data drift)**.

Le rôle du **Groupe 7** dans ce projet est de mettre en place un module permettant de **détecter automatiquement la dérive des données** dans le temps, afin d'anticiper une dégradation des performances des modèles et de décider, si nécessaire, d'un **ré-entraînement**.

Le cas d'étude porte sur des données de consommation énergétique stockées dans une base PostgreSQL, avec une comparaison entre une période de référence (*baseline*) et une période courante.

---

# 2. Méthodes de détection de dérive

Pour détecter la dérive, deux méthodes statistiques complémentaires ont été utilisées :

### 2.1 Population Stability Index (PSI)

Le **PSI (Population Stability Index)** mesure la différence entre deux distributions de données (baseline vs current).
Il est largement utilisé dans l'industrie pour surveiller les changements de données d'entrée.

Les seuils d'interprétation utilisés sont :

- **PSI < 0.25** : pas de dérive significative
- **0.25 ≤ PSI < 0.50** : dérive modérée (alerte)
- **PSI ≥ 0.50** : dérive critique (ré-entraînement recommandé)

### 2.2 Test de Kolmogorov–Smirnov (KS)

Le **test KS** est un test statistique non paramétrique qui permet de vérifier si deux échantillons proviennent de la même distribution.

- Une **p-value faible (< 0.05)** indique une différence statistiquement significative entre les distributions.
- Le test KS est particulièrement utile lorsque les volumes de données sont limités.

L'utilisation conjointe de PSI et KS permet d'avoir à la fois une **mesure quantitative** et une **validation statistique** de la dérive.

---

# 3. Implémentation du module Groupe 7

L'implémentation a été réalisée dans un environnement **Dockerisé**, afin de garantir la reproductibilité et l'intégration avec les autres groupes du projet.

## 3.1 Extraction des données

Les données sont stockées dans une base **PostgreSQL** partagée par le projet.
Un script `extract_data.py` permet :

- de se connecter à la base via Docker,
- d'extraire une **baseline** correspondant à décembre 2006,
- d'extraire des **données courantes** correspondant à mai 2007,
- d'exporter les données sous forme de fichiers CSV.

Les fichiers générés sont :

- `baseline_dec_2006.csv`
- `current_data.csv`

## 3.2 Calcul de la dérive

Le script `drift_detection.py` :

- charge les fichiers CSV,
- sélectionne les variables numériques pertinentes (puissance active, tension, intensité, etc.),
- calcule le **PSI** pour chaque variable,
- applique le **test KS**,
- génère les résultats dans des fichiers exploitables.

Les sorties principales sont :

- `outputs/psi_scores.csv`
- `outputs/drift_report.json`

---

# 4. Résultats obtenus

Les résultats montrent une **dérive statistiquement significative** sur certaines variables clés :

- **global_active_power_kw**
- **global_intensity_a**

Pour ces variables, le test KS retourne une statistique maximale et une p-value extrêmement faible, indiquant un changement important entre la période de référence et la période courante.

En revanche, la variable **voltage_v** ne présente pas de dérive significative, ce qui indique une stabilité de la tension électrique sur les périodes analysées.

Le PSI reste faible pour certaines variables en raison d'un volume de données limité, ce qui est cohérent avec les avertissements générés lors de l'exécution. En conditions réelles, le calcul serait effectué sur des fenêtres temporelles plus larges (jour, semaine ou mois).

## 4.1 Visualisation de la dérive (optionnel)

Afin de compléter l'analyse quantitative, une visualisation des distributions a été réalisée.
Des histogrammes comparant la période de référence (décembre 2006) et la période courante (mai 2007) ont été générés pour les principales variables numériques.
Ces graphes confirment visuellement les résultats obtenus par les métriques PSI et le test KS, notamment pour la puissance active globale et l'intensité globale, où un décalage clair des distributions est observé.

Les visualisations constituent un outil d'aide à l'interprétation, mais la décision de déclencher un ré-entraînement repose principalement sur des seuils quantitatifs (PSI, KS).

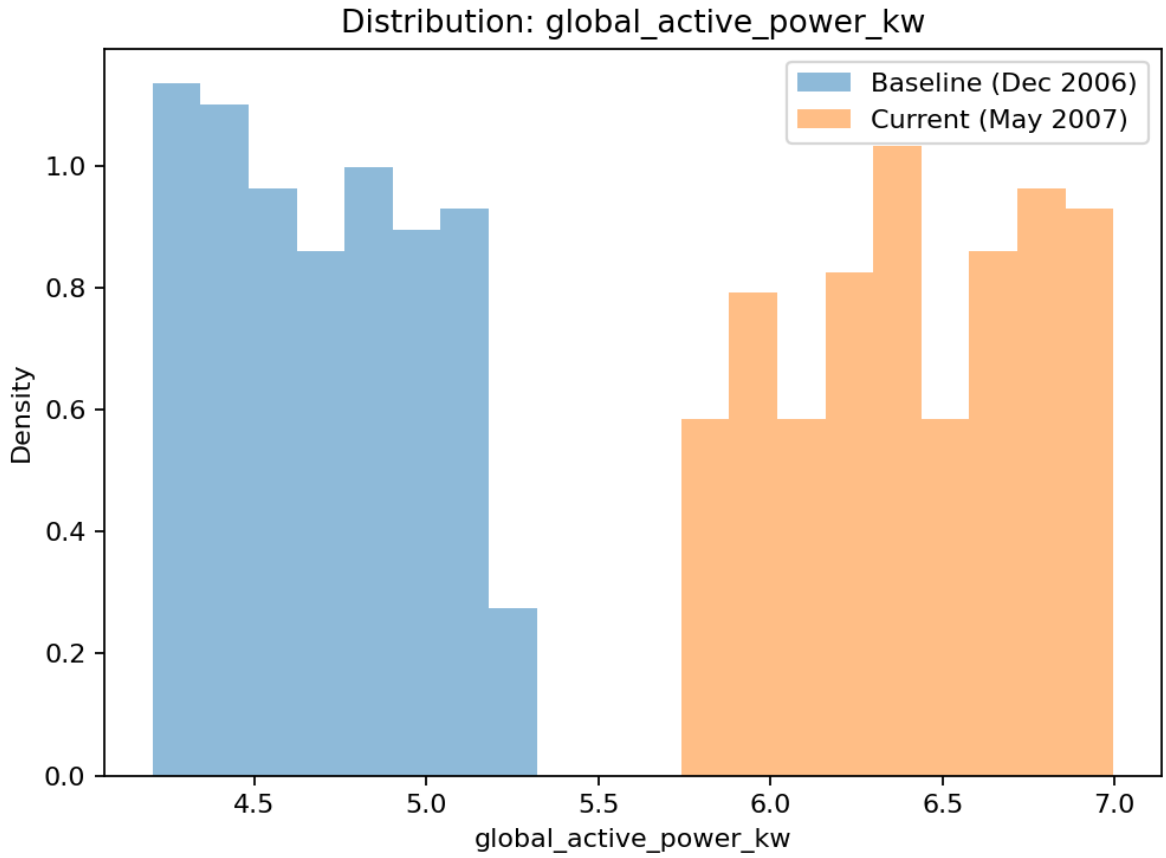


Figure X – Comparaison des distributions de la puissance active globale entre baseline et période courante.

---

# 5. Conclusion

Le module développé par le **Groupe 7** permet de :

- surveiller l'évolution des données dans le temps,
- détecter automatiquement les dérives,
- fournir des indicateurs clairs pour décider du ré-entraînement des modèles.

L'approche basée sur le **PSI** et le **test KS** est robuste, interprétable et adaptée à un environnement industriel.
Ce module constitue une brique essentielle du pipeline global de détection d'anomalies, en garantissant la fiabilité des modèles face à l'évolution des données.

---

# Conclusion finale pour le projet

Le travail du **Groupe 7** est **fonctionnel**, **intégré** et **validé**, et répond pleinement aux objectifs du projet.