

# Prédiction de la Présence des Employés via Deep Learning

**Sujet :** Analyse de séries temporelles et modélisation prédictive (Smart Workplaces).

**Données :** [GitHub - Prediction of Employee Presence](#)

---

## Phase 1 : Exploration et Préparation

- **Acquisition des données :** Cloner le repo et charger les données (CSV/Excel).
- **EDA (Exploratory Data Analysis) :**
  - Visualisation de la série temporelle (tendance, saisonnalité journalière/hebdomadaire).
  - Détection des anomalies.
  - Analyse des corrélations
- **Preprocessing :**
  - Nettoyage : Imputation des valeurs manquantes.
  - Feature Engineering : Création de variables temporelles (heure de la journée, jour de la semaine, indicateur week-end).
  - **Normalisation :** Étape cruciale pour le Deep Learning (MinMaxScaling ou StandardScaler).

## Phase 2 : Modélisation - Baselines & Machine Learning

- Avant de faire du Deep Learning, il faut une référence.
- **Machine Learning Classique (Optionnel mais recommandé) :** Random Forest ou XGBoost en utilisant des fenêtres glissantes (lags) comme entrées.

## Phase 3 : Deep Learning

- **Mise en forme des données :** Création des tenseurs 3D (Samples, TimeSteps, Features).
- **Architectures à tester :**
  - **RNN Simple :** Pour comprendre les limites (vanishing gradient).
  - **LSTM (Long Short-Term Memory) :** Le standard pour ce type de problème.
  - **GRU (Gated Recurrent Unit) :** Alternative plus légère au LSTM.
  - **TimeXer** et autres modèles de fondation pour séries temporelles

- **Expérimentation** : Varier la longueur de la séquence d'entrée (look-back window), le nombre de neurones, et le Dropout pour éviter l'overfitting.

#### Phase 4 : Analyse, Rapport et Soutenance

- Comparaison des performances (Tableaux récapitulatifs).
- Visualisation des prédictions vs Réalité sur l'ensemble de test.
- Rédaction du rapport et création des slides.

---

#### A. Le Rapport Technique (50% de la note)

Le rapport (max 15 pages) doit contenir :

1. **Introduction** : Contexte du Smart Workplace et définition du problème.
2. **Analyse des Données** : Graphiques pertinents montrant les patterns de présence (pics le matin, creux le midi, absence le week-end).
3. **Méthodologie** :
  - Comment les données ont été divisées (Train/Val/Test) ? **Attention** : Interdiction de mélanger les données (shuffling) sur des séries temporelles.
  - Justification du choix des hyperparamètres.
4. **Résultats** : Comparaison des modèles (Naïf vs LSTM) avec métriques (MAE, RMSE).
5. **Conclusion** : Quelle est la meilleure approche ? Quelles sont les limites ?

#### B. La Soutenance Orale (50% de la note)

- **Format** : 15 minutes de présentation + 10 minutes de Questions/Réponses.
- **Support** : Slides visuels (éviter les murs de texte).
- **Attentes** :
  - Clarté de l'exposition.
  - Capacité à expliquer *pourquoi* un modèle fonctionne mieux qu'un autre.
  - Démonstration de la compréhension des mécanismes du Deep Learning (ex: expliquer le rôle d'une porte d'oubli dans un LSTM).