Université Paris Cité

Machine Learning for Data Science

Rapport du Projet de Classification Non Supervisé

Réalisé par :

- BELKAID Meryem (AMSD)
- BOANANI Hafsa (AMSD)
- OUMGHAR Abir (AMSD)

Année universitaire 2023-2024

Table des matières

1. Introduction

La disparité marquée de la consommation électrique entre différents foyers résulte d'une combinaison complexe de paramètres, notamment le nombre et l'usage spécifique des équipements électroménagers, répartis entre le chauffage, l'eau chaude, la cuisson et divers autres appareils électriques. Cette diversité est également influencée par la fréquence et la durée d'utilisation de ces équipements. La nature non stockable de l'électricité à grande échelle souligne l'importance cruciale de maintenir un équilibre constant entre la production et la consommation pour assurer la stabilité du système électrique et prévenir d'éventuelles pannes.

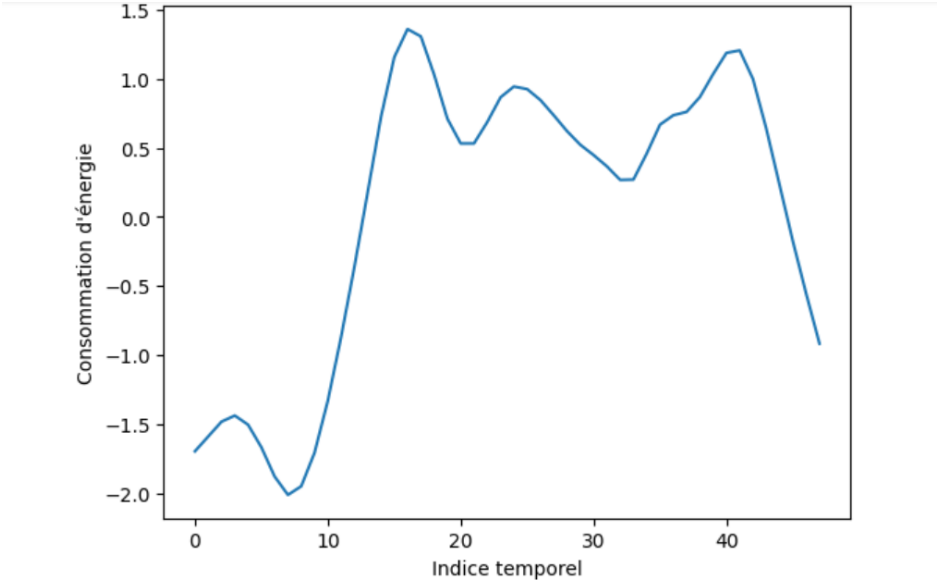
La variabilité des comportements des consommateurs pose un défi majeur, et regrouper ces comportements en catégories homogènes peut fournir des avantages significatifs. Cela permet non seulement une meilleure compréhension des habitudes de consommation, mais ouvre également la voie à la proposition d'offres personnalisées et à des prévisions plus précises. En effet, prévoir la consommation de groupes homogènes facilite la compréhension globale en agrégeant ces prévisions.

Ainsi, ce projet vise à répondre à ces défis en appliquant la classification non supervisée pour segmenter un ensemble de séries temporelles représentant la consommation électrique de 100 appartements sur une période de 91 jours consécutifs. L'objectif particulier est de détecter de manière automatique des ruptures, symbolisant des changements notables dans les habitudes de consommation, et ce, de manière uniforme à travers toutes les séries temporelles. En établissant ces segments homogènes, le projet aspire à apporter des éclairages significatifs sur les dynamiques complexes de la consommation énergétique résidentielle, ouvrant la voie à des stratégies de gestion plus précises et durables.

2. Exploration des données

Pour initier notre exploration des données, nous débutons par l'importation des informations cruciales. Les données, collectées toutes les 30 minutes au cours de 91 jours consécutifs auprès de 100 appartements, sont organisées en trois tableaux : X, APPART, et JOUR. Ces éléments, respectivement représentant la consommation énergétique, le numéro de l'appartement, et le numéro du jour, fournissent la base de notre analyse, nous établissons alors une première compréhension de la structure et de la dimensionnalité de nos données.

Une fois les données importées, nous procédons à leur transformation pour les rendre plus lisibles et manipulables. Cela inclut le renommage des colonnes, la fusion des données supplémentaires avec le tableau principal, et le calcul de statistiques descriptives pour chaque appartement. Ces étapes permettent de créer une représentation plus claire des profils de consommation énergétique individuels, visualisée à travers des graphiques spécifiques à chaque appartement.

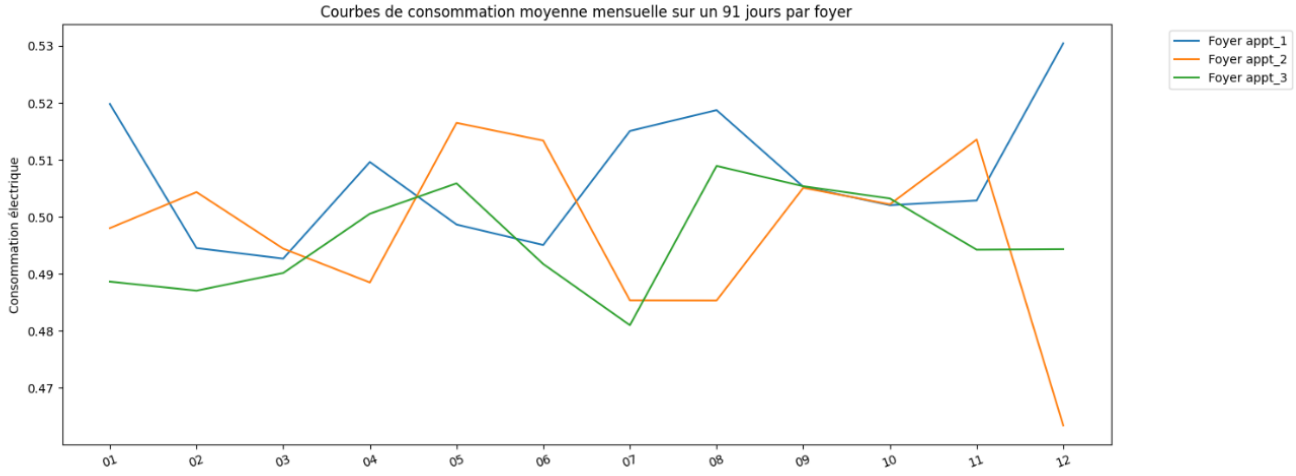


Consommation d'énergie - Appartement 3

Une étape clé de notre exploration consiste à agréger les données à des échelles temporelles significatives, à savoir mensuelle et journalière. Cette agrégation offre un moyen efficace d'observer les variations comportementales des ménages, en mettant en évidence les moments les plus sensibles de consommation, tels que le matin, le soir, la nuit et les week-ends. Cette visualisation initiale, bien que susceptible d'atténuer la variabilité des données, constitue une première technique d'agrégation, nous permettant d'esquisser les tendances générales.

Par la suite, nous envisageons une exploration plus fine en visualisant les courbes représentant la consommation journalière moyenne de certains ménages pendant une semaine. Cette approche détaillée

nous offre des indications précieuses sur les habitudes spécifiques de chaque foyer, notamment en ce qui concerne les variations de consommation liées aux jours de la semaine.



Afin d'affiner notre analyse et de préparer nos données pour une segmentation avancée, nous considérons la normalisation des courbes de consommation. Cette étape stratégique vise à garantir que l'algorithme de clustering se concentre sur les différences de comportement plutôt que sur les niveaux absolus de consommation énergétique. L'ensemble de ces étapes constitue un socle robuste pour une exploration plus approfondie, notamment en vue de la segmentation avancée des séries temporelles à l'aide de techniques de clustering.

En résumé, notre exploration des données comprend les étapes cruciales d'importation, de transformation pour une lisibilité accrue, d'agrégation mensuelle et journalière, ainsi que de visualisation à différentes échelles temporelles. Ces fondations solides préparent le terrain pour des analyses plus avancées et une compréhension approfondie des motifs de consommation énergétique des 100 appartements étudiés.

3. Réduction des dimensions

Pour diminuer la complexité de nos données temporelles tout en préservant les caractéristiques significatives, nous avons identifié trois méthodes distinctes :

3.1. Analyse en Composantes Principales (ACP) :

L'introduction de l'Analyse en Composantes Principales (ACP) marque une étape décisive dans notre démarche d'exploration des données. L'application de l'ACP représente une avancée significative, offrant une optimisation mathématique de la réduction de la variance des séries temporelles de consommation. Ce processus, cependant, n'est pas sans compromis, car il entraîne une perte partielle d'interprétabilité des

caractéristiques. Néanmoins, il permet de condenser nos données en un nombre restreint de caractéristiques, tout en préservant au moins 90% de la variabilité des données.

Une fois les données normalisées et consolidées dans un DataFrame dédié (df_scaled_to_acp), nous passons à l'étape suivante en initialisant l'objet PCA. Le paramètre n_components est spécifié avec une valeur de 0.95, exprimant notre intention de préserver 95% de la variance des données. L'ACP est ensuite appliquée aux données normalisées à l'aide de la méthode fit_transform(). Les composantes principales résultantes sont alors stockées dans un nouveau DataFrame nommé df_acp.

Cette phase cruciale de réduction dimensionnelle est essentielle pour concentrer l'information tout en conservant une représentation significative des données normalisées, facilitant ainsi la suite de notre analyse.

3.2. Création de Caractéristiques Personnalisées :

En adoptant cette approche, nous avons la liberté de concevoir nos propres caractéristiques en utilisant des techniques d'analyse spécifiques aux séries temporelles. Bien que ces caractéristiques soient facilement interprétables, elles ne sont pas nécessairement optimisées pour le clustering. Nous pouvons extraire des statistiques descriptives telles que la moyenne, l'écart-type, ou d'autres mesures de tendance centrale et de dispersion pour chaque série temporelle, afin de capturer les aspects les plus significatifs de nos données.

3.3. Transformation de Fourier :

En recourant à la Transformation de Fourier, nous sommes en mesure d'extraire les caractéristiques fréquentielles inhérentes à nos séries temporelles. Cette approche représente nos données sous forme de coefficients fréquentiels, offrant ainsi une perspective utile pour détecter des motifs ou des tendances périodiques. Les coefficients fréquentiels les plus pertinents peuvent être sélectionnés pour représenter les informations les plus significatives de nos séries temporelles.

En employant ces méthodes, notre objectif est de réduire la dimensionnalité de nos données et d'extraire les caractéristiques les plus informatives, tout en préservant un niveau élevé d'interprétabilité. Cette approche facilitera non seulement un clustering plus efficace, mais également une compréhension approfondie des disparités de comportement au sein de nos données temporelles.

4. Clustering

Cette partie de notre travail est dédiée à la technique de clustering. Le clustering est un processus qui permet de regrouper des données similaires dans des groupes ou des clusters. Dans cette section, nous avons exploré différentes méthodes de clustering, notamment la classification hiérarchique, les méthodes basées sur l'extraction des caractéristiques et les méthodes directes telles que le K-means, la carte de

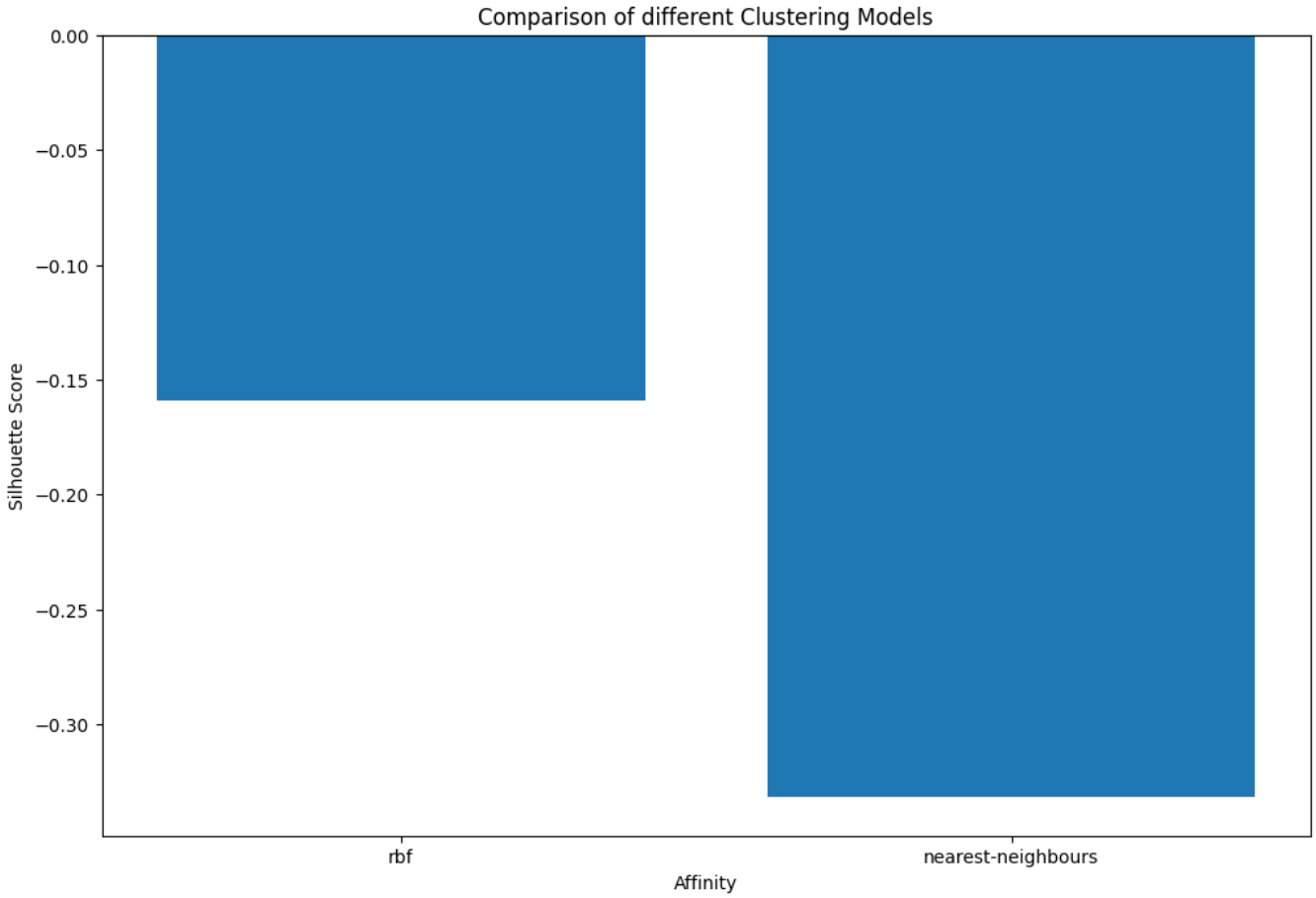
Kohonen et la méthode DBA. Chaque méthode a ses propres avantages et inconvénients, et le choix de la méthode dépend du type de données et des objectifs de l'analyse. Une fois les clusters obtenus, nous avons effectué une analyse supplémentaire pour comprendre les caractéristiques distinctes de chaque cluster.

Dans cette section, nous avons exploré différentes méthodes de clustering.

- Classification Hiérarchique : La classification hiérarchique consiste à construire une hiérarchie de clusters en regroupant les données de manière ascendante ou descendante.

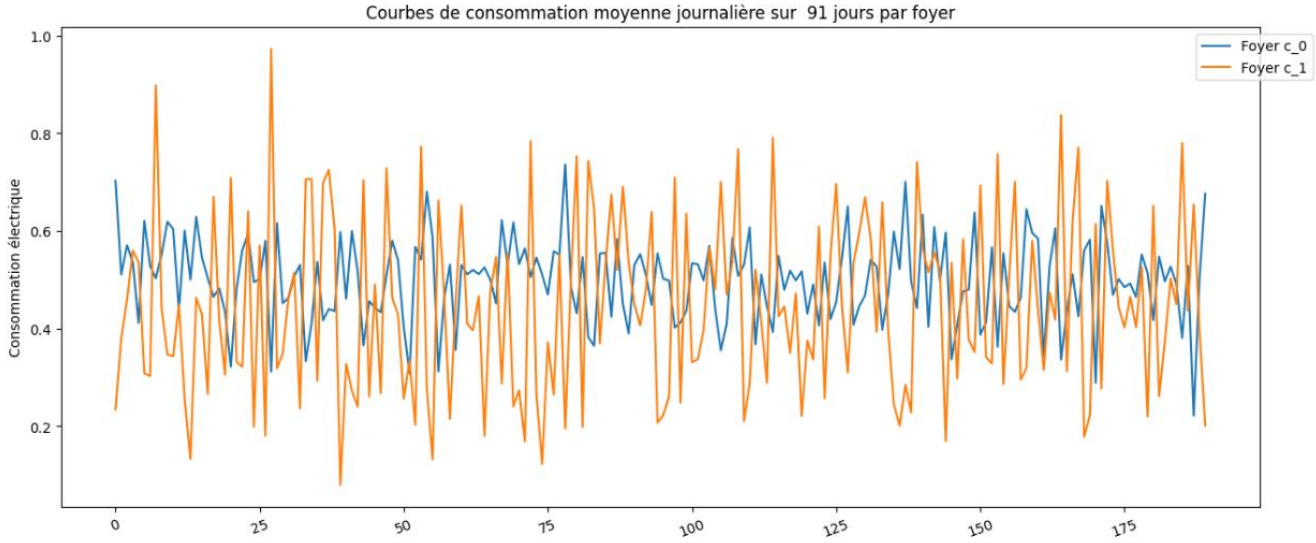
- Méthodes basées sur l'extraction des caractéristiques : Nous avons également exploré des méthodes qui utilisent des techniques d'extraction de caractéristiques pour regrouper les données en fonction de leurs propriétés spécifiques.

- Méthodes directes : Nous avons étudié des méthodes directes telles que le K-means, la carte de Kohonen (Self Organizing Map SOM) et la méthode DBA (Dynamic Time Warping Barycenter Averaging).



Une fois les clusters obtenus à l'aide de ces différentes méthodes, nous avons effectué une analyse supplémentaire pour comprendre les caractéristiques distinctes de chaque cluster.

Lors de notre analyse, nous avons tracé les courbes de consommation moyenne journalière sur une période de 91 jours par foyer. Cette approche nous a permis de visualiser les tendances de consommation sur une base quotidienne et d'identifier des modèles récurrents.



5. Segmentation

Dans cette partie de notre notebook, nous nous sommes concentrés sur la segmentation des données. La segmentation consiste à diviser les données en sous-groupes homogènes en fonction de certains critères. Dans notre cas, nous avons utilisé la segmentation pour regrouper les profils de consommation électrique similaires. Nous avons utilisé la méthode de clustering spectral pour construire des clusters de profils de consommation ayant des caractéristiques similaires. En analysant les profils moyens de chaque cluster, nous avons pu identifier les différents comportements de consommation. Cette segmentation nous permet d'adapter les stratégies de gestion de l'énergie en fonction des besoins spécifiques de chaque groupe.

Nous nous sommes concentrés sur la segmentation des données. La segmentation consiste à diviser les données en sous-groupes homogènes en fonction de certains critères.

Nous avons utilisé la méthode de clustering spectral pour construire des clusters de profils de consommation ayant des caractéristiques similaires. Cette méthode identifie automatiquement le nombre de clusters à partir des données.

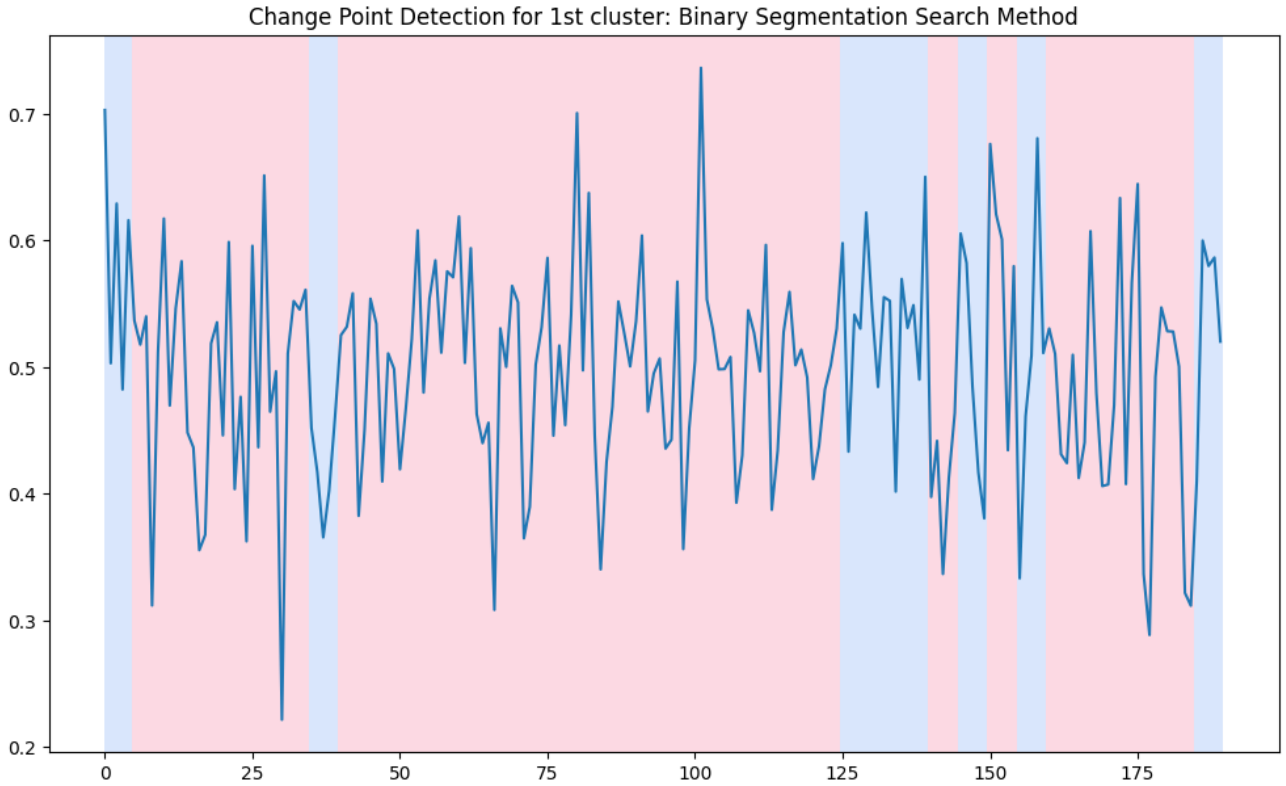
Une fois les clusters formés, nous avons calculé les profils de consommation électrique moyenne de chaque cluster. En analysant ces profils moyens, nous avons pu identifier les différents comportements de consommation.

6. Binary Segmentation Search Method

Cette partie de notre travail introduit la méthode de Binary Segmentation Search (BSS). Cette méthode est utilisée pour détecter les points de changement dans une séquence de données. Elle fonctionne en divisant itérativement la séquence en sous-séquences afin d'identifier les points où un changement significatif se produit. Dans notre cas, nous avons appliqué la méthode de BSS aux profils de consommation électrique moyenne de chaque cluster. Cela nous a permis de prédire les points de changement dans les profils de consommation électrique, ce qui peut être utile pour comprendre les tendances et les évolutions dans les comportements de consommation des ménages. Toutefois, il est important de noter que la méthode de BSS est une méthode approximative et peut ne pas détecter tous les points de changement réels.

Dans notre travail Nous avons appliqué la méthode de BSS aux profils de consommation électrique moyenne de chaque cluster. Cette méthode divise itérativement les profils en sous-séquences afin d'identifier les points où un changement significatif se produit.

Nous avons utilisé la méthode de détection de points de changement appelée Binary Segmentation Search Method pour analyser les profils de consommation électrique moyenne des clusters obtenus à partir du clustering spectral.



En utilisant la méthode de BSS, nous sommes en mesure de prédire les points de changement dans les profils de consommation électrique moyenne, ce qui peut être utile pour comprendre les tendances et les évolutions dans les comportements de consommation des ménages. Toutefois, il est important de noter que la méthode de BSS est une méthode approximative et peut ne pas détecter tous les points de changement réels.

7. Conclusion

En conclusion, notre exploration des techniques d'apprentissage non supervisé a révélé des perspectives passionnantes pour l'analyse des données de consommation électrique des ménages. En utilisant des méthodes telles que le clustering, la segmentation et le Binary Segmentation Search (BSS), nous avons pu découvrir des schémas et des comportements sous-jacents qui auraient pu rester invisibles autrement.

Ces techniques nous ont permis de regrouper les consommateurs en fonction de leurs habitudes de consommation, de segmenter les profils électriques pour identifier des tendances spécifiques et de détecter les points de changement significatifs. Ces informations sont cruciales pour comprendre les dynamiques de consommation, anticiper les fluctuations de la demande et concevoir des stratégies d'efficacité énergétique ciblées.

L'apprentissage non supervisé a prouvé sa valeur en révélant des insights précieux et en permettant une prise de décision éclairée dans le domaine de l'énergie. En combinant ces techniques avec des approches créatives et innovantes, nous pouvons exploiter pleinement le potentiel des données de consommation électrique pour promouvoir des pratiques durables, réduire les coûts énergétiques et contribuer à la préservation de l'environnement.

En conclusion, l'apprentissage non supervisé offre une porte ouverte vers une meilleure compréhension des comportements de consommation, une optimisation de la gestion de l'énergie et une transition vers une société plus consciente de son empreinte énergétique. Grâce à ces avancées, nous pouvons esquisser un avenir où l'efficacité énergétique devient une réalité pour tous, avec des impacts positifs à la fois sur le plan environnemental et économique.

8. Lien de Notebook

- ✓ Colab Link : https://drive.google.com/file/d/1e822AARQM8qfqII_V6zPR6JdaYA1GLB-/view?usp=sharing