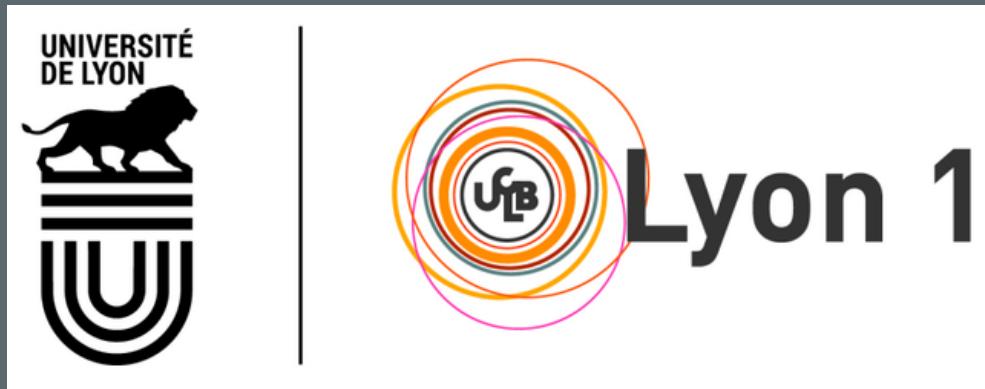


Rapport de stage

Université Claude Bernard Lyon 1

Kévin Tang

Année 2024



Rapport de stage

Université Claude Bernard Lyon 1

01/04/2024 - 30/08/2024

Kévin Tang

Etudiant en M2 Data Science
Année scolaire : 2023-2024

Enseignant référent

M. Rémy CAZABET

Encadrants de stage

Mme. Parisa GHODOUS
M. Mohamed-Essaid KHANOUCHE

Établissement :

Université Claude Bernard Lyon 1

43 Boulevard du 11 Novembre 1918
69622 Villeurbanne cedex
www.univ-lyon1.fr

Remerciements

Je souhaite exprimer ma profonde gratitude à toutes les personnes qui ont contribué à ce stage.

Tout d'abord, je remercie chaleureusement M. Mohamed-Essaid KHANOUCHÉ et Mme. Parisa GHODOUS, mes encadrants durant ce stage, pour leur disponibilité, leur accompagnement et leurs précieux conseils tout au long de cette expérience. Leur soutien constant et leurs remarques constructives m'ont permis de progresser et d'approfondir mes connaissances dans le domaine de la recherche en consommation énergétique.

Un grand merci à M. Christian Obrecht pour avoir essayé de faciliter l'accès aux données, malgré les contraintes rencontrées et à M. Ahror BELAID pour l'aide reçue pour la rédaction de l'article.

Enfin, je remercie mes collègues et amis pour leur soutien moral et leurs encouragements, qui ont été essentiels pour surmonter les défis rencontrés pendant ce stage.

Merci à toutes et à tous pour votre aide précieuse.

Sommaire

05 | Introduction

06 | Description de la structure

09 | Environnement de travail

10 | Les missions confiées

11 | Etat de l'Art

14 | Analyse de RECS2020

29 | Présentation des résultats

31 | Les problèmes rencontrés, les solutions
envisagées

32 | Bilan du stage

33 | Conclusion

34 | Références bibliographiques

Introduction

1

J'ai eu l'opportunité d'intégrer un stage qui s'inscrit dans le domaine émergent de l'Internet du comportement (IoB : Internet of Behavior), un paradigme innovant visant à analyser les tendances comportementales des utilisateurs pour atteindre des objectifs spécifiques. Le stage s'est concentré sur l'application de ce concept à la gestion de la consommation énergétique dans les espaces résidentiels, un domaine de plus en plus crucial à mesure que les préoccupations en matière de durabilité énergétique s'intensifient.

L'IoB permet de collecter et d'analyser des données comportementales provenant de divers dispositifs intelligents. Ces données, qui reflètent les habitudes et modes de vie des utilisateurs, sont devenues essentielles pour comprendre comment les comportements individuels influencent la consommation énergétique. Les dépenses énergétiques des résidences représentent en effet une part significative de la consommation mondiale d'énergie, ce qui souligne l'importance de cette thématique.

Le stage visait à examiner l'impact de différents comportements sur la consommation d'énergie, en identifiant les modèles de consommation et en comprenant les facteurs clés qui les sous-tendent. Pour ce faire, j'ai mené une analyse approfondie des données issues des ménages, y compris des statistiques descriptives et des techniques de regroupement. L'objectif était de découvrir les relations entre les comportements des occupants et leur consommation énergétique.

À partir des résultats obtenus, l'étude cherchait à fournir des informations précises sur la manière dont ces comportements contribuent à la consommation d'énergie, et à formuler des recommandations ciblées pour réduire cette consommation.

Ainsi, ce stage représentait une opportunité unique d'explorer les interactions entre comportement humain et durabilité énergétique, tout en m'immergeant dans les méthodologies de recherche appliquées à un domaine d'une grande importance pour l'avenir.

Description de la structure

2

L'Université Claude Bernard Lyon 1

L'Université

Située au cœur de la ville dynamique de Lyon, l'Université Claude Bernard Lyon 1, souvent abrégée en UCBL, se distingue comme un établissement d'enseignement supérieur et de recherche de premier plan en France. Fondée en 1971, elle s'inscrit dans une longue tradition d'excellence académique, ayant notamment donné naissance à de nombreux prix Nobel et personnalités illustres.

L'UCBL propose un éventail complet de formations en licence, master et doctorat, couvrant un large spectre de disciplines, des sciences fondamentales aux sciences humaines et sociales, en passant par les sciences de l'ingénieur, les sciences de la santé et les sciences économiques et de gestion. Avec plus de 47 000 étudiants répartis sur ses campus, l'université offre un environnement d'apprentissage stimulant et multiculturel.

L'université est également reconnue pour son dynamisme en matière de recherche. Ses 85 laboratoires et plateformes de recherche mènent des travaux de pointe dans des domaines d'investigation variés, contribuant ainsi à l'avancement des connaissances et à l'innovation. L'UCBL se classe régulièrement parmi les meilleures universités françaises et internationales en termes de performance de recherche.

L'université joue également un rôle crucial dans le développement économique et social de la région Auvergne-Rhône-Alpes. Elle collabore étroitement avec les entreprises et les acteurs socio-économiques locaux pour favoriser l'innovation et l'entrepreneuriat. L'UCBL est ainsi un atout majeur pour le territoire, contribuant à son dynamisme et à son attractivité.

Le campus LyonTech-La Doua

Le campus LyonTech-la Doua, situé à Villeurbanne au cœur de la métropole lyonnaise, se distingue par son effervescence et son dynamisme. Avec ses 22 000 étudiants, 1 500 chercheurs et 1 300 doctorants, il constitue un pôle d'enseignement supérieur et de recherche de premier plan.

S'étendant sur environ 100 hectares, le campus LyonTech-la Doua offre un cadre propice à l'épanouissement des étudiants. La présence de la faculté des Sciences et Technologies, l'UFR STAPS, l'IUT Lyon 1, l'INSA Lyon, CPE Lyon, l'ENSSIB et de Polytech Lyon garantit une diversité de formations de qualité, répondant ainsi aux aspirations et aux besoins d'un large éventail d'étudiants.

Le campus réunit aussi les antennes régionales du CNRS, de l'INRAE et de l'INRIA. Cette concentration d'acteurs majeurs de la recherche favorise la synergie et l'émergence de collaborations innovantes dans des domaines d'investigation variés.



Le Nautibus

Situé sur le campus LyonTech-la Doua à Villeurbanne, le Nautibus est un bâtiment emblématique de l'Université Claude Bernard Lyon 1 (UCBL). Il abrite principalement des unités de formation et de recherche en Sciences et Technologies, en particulier le département Informatique de l'université.

Le département d'informatique du Nautilus accueille près de 800 étudiants, répartis dans 11 formations de premier, deuxième et troisième cycles. Ces cursus, dispensés par une équipe d'une cinquantaine d'enseignants-rechercheurs expérimentés, couvrent un large spectre de domaines, allant de l'informatique fondamentale à l'informatique appliquée, en passant par l'intelligence artificielle, les réseaux et systèmes informatiques, et la robotique.

Quatre laboratoires de recherche de pointe, animés par des chercheurs passionnés et reconnus dans leurs domaines respectifs, complètent l'écosystème du Nautibus. Ces laboratoires mènent des travaux innovants sur des thématiques d'actualité telles que la cybersécurité, l'internet des objets, le big data et l'intelligence artificielle, contribuant ainsi à l'excellence de l'UCBL en matière de recherche en informatique.

Cette synergie entre enseignement de qualité, recherche de pointe et infrastructures modernes permet aux étudiants du Nautibus d'acquérir des compétences solides et de développer leur expertise dans les domaines de l'informatique qui les passionnent. Ils sont ainsi préparés à relever les défis du monde numérique de demain et à s'insérer avec succès dans le marché du travail.

Le LIRIS

Né en 2003 sous la tutelle d'institutions prestigieuses telles que le CNRS, l'INSA Lyon, l'Université Claude Bernard Lyon 1, l'Université Lumière Lyon 2 et l'Ecole Centrale de Lyon, le Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS) s'impose comme un pôle d'excellence en informatique et sciences de l'information. Fort de ses 330 membres, le laboratoire rayonne à l'échelle nationale et internationale, porté par un dynamisme et une expertise incontestables.

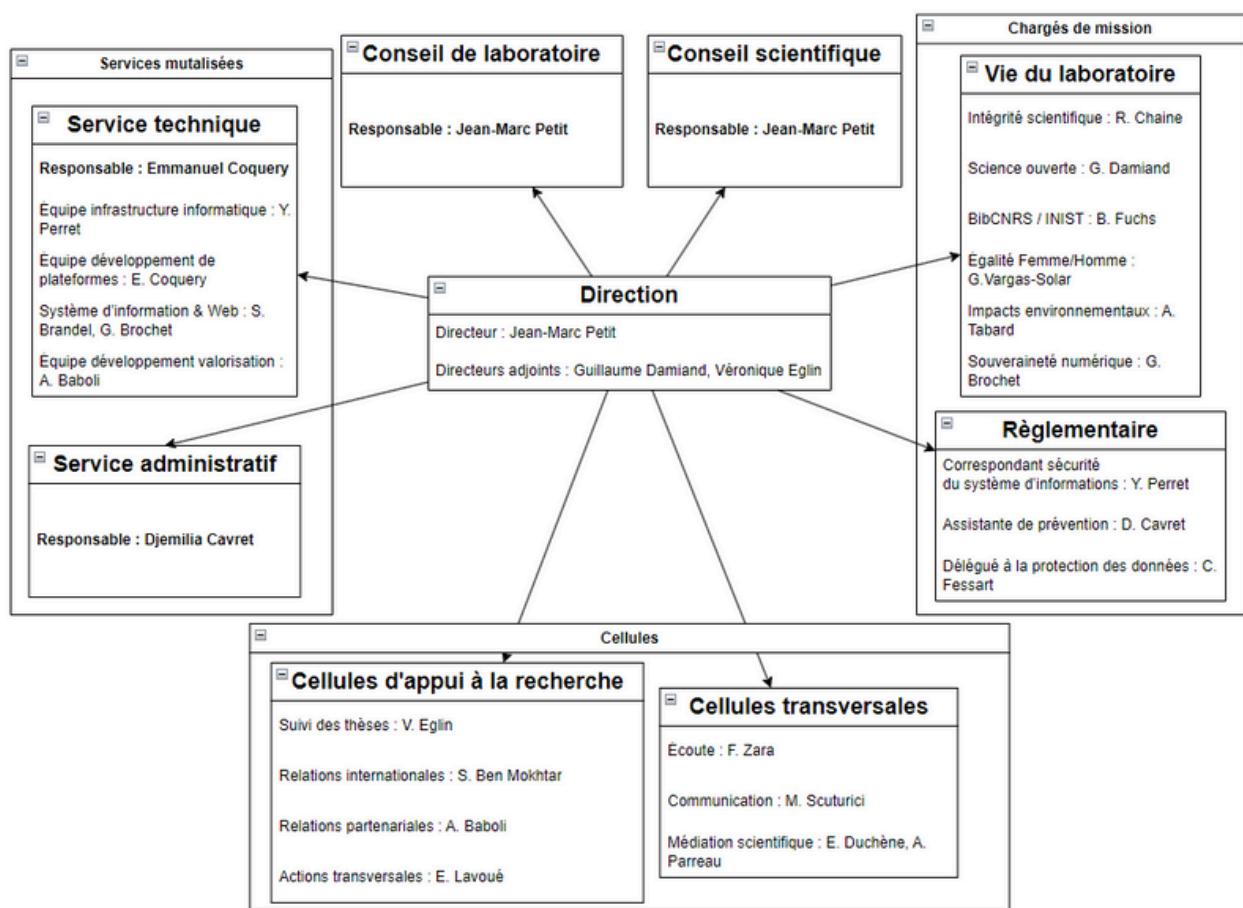
Sous la houlette de Jean-Marc Petit, son directeur actuel, le LIRIS explore un vaste champ scientifique, englobant l'informatique et plus largement les sciences et technologies de l'information. Ses recherches, structurées en six pôles de compétences, s'attaquent aux défis majeurs du monde numérique, de l'intelligence artificielle à l'analyse de données massives, en passant par la vision par ordinateur, la cybersécurité, la transformation digitale et l'apprentissage humain.

Le LIRIS se distingue par son approche résolument interdisciplinaire, tissant des liens étroits entre les sciences humaines et sociales, l'ingénierie, la médecine, les sciences de la vie et les sciences de l'environnement. Cette synergie d'expertises permet d'aborder les enjeux sociétaux contemporains dans une perspective globale et innovante, en particulier ceux liés à la souveraineté numérique et au développement durable.

Conscient de l'importance de la médiation scientifique, le LIRIS s'engage activement dans la diffusion de la culture informatique auprès du grand public. Il entretient également de nombreuses relations avec son environnement socio-économique et culturel, tant au niveau local et régional que national. Ces collaborations nourrissent ses recherches et permettent de valoriser ses travaux auprès d'un large public.

Les travaux du LIRIS trouvent des applications concrètes dans des domaines variés, tels que la santé, le calcul scientifique, l'apprentissage humain et l'intelligence ambiante. Ses recherches contribuent ainsi à améliorer la qualité de vie et à relever les défis majeurs de notre société.

Organigramme du LIRIS



Environnement de travail

3

Lieu et matériel

Lieu de travail

Mon stage s'est effectué au sein du laboratoire LIRIS, situé au 2ème étage du bâtiment Nautibus. L'ensemble de l'étage est dédié au laboratoire, offrant un environnement de travail spacieux et propice à la recherche. L'espace est intelligemment agencé, divisé entre des bureaux individuels occupés par les enseignants-chercheurs et des "BOX", des espaces de travail collaboratifs fréquemment utilisés par les stagiaires. Des espaces communs tels qu'un coin café, une cuisine et des sanitaires viennent compléter l'aménagement, favorisant ainsi les échanges et le bien-être des occupants.

Soucieux de mon confort et de ma productivité, mes encadrants m'ont également accordé la possibilité d'effectuer une partie de mon travail en télétravail. Cette flexibilité m'a permis de gérer mon temps de manière autonome et d'optimiser mon organisation personnelle, contribuant ainsi à mon épanouissement professionnel.



Outils et logiciels

Lors de ma présence au sein du laboratoire LIRIS, j'étais équipé d'un ordinateur portable Dell Latitude 5520, doté d'un processeur i5-1135G7 et de 16 Go de RAM, mis à disposition par l'université. Ce matériel performant me permettait d'accomplir mes tâches quotidiennes avec fluidité et efficacité. En télétravail, j'utilisais mon ordinateur personnel, un PC fixe ASUS, qui offrait des ressources suffisantes pour poursuivre mes travaux dans un environnement familier.

Pour assurer une collaboration efficace et un partage fluide des informations, j'utilisais la suite d'outils Google Workspace. Cette plateforme cloud me permettait d'accéder à mes documents, feuilles de calcul et présentations depuis n'importe quel appareil, en ligne ou hors ligne. Cette flexibilité m'a permis de travailler de manière continue et productive, que je sois au laboratoire ou en télétravail.

En outre, pour le développement et l'exécution de codes, j'ai régulièrement utilisé Jupyter Notebook et Google Colab. Ces outils étaient essentiels pour le traitement des données et le développement de modèles analytiques. Pour les présentations, j'ai fréquemment utilisé Google Slides, ce qui m'a permis de communiquer efficacement mes résultats et mes avancées. Enfin, pour la rédaction de l'article final, j'ai utilisé Overleaf, une plateforme LaTeX, afin de produire des documents bien formatés et conformes aux normes académiques.

Les missions confiées

4

Mon stage dans l'entreprise

Mes missions

1

Etat de l'art

Réaliser un état de l'art est une étape cruciale dans tout projet de recherche, qu'il soit scientifique, technique ou entrepreneurial. Il s'agit d'une analyse approfondie et critique de la littérature existante sur un sujet donné, permettant de dresser un panorama complet des connaissances actuelles et d'identifier les avancées les plus récentes.

L'état de l'art aide à identifier les lacunes dans les recherches actuelles, à souligner les questions non résolues, et à repérer les opportunités d'innovation. En fin de compte, cette étape assure que le projet est pertinent, bien informé et orienté vers des contributions originales et significatives.

2

Analyse de RECS2020

Durant mon stage, j'ai réalisé une analyse approfondie de la consommation énergétique des ménages en utilisant le dataset RECS2020, j'ai étudié comment les comportements et les caractéristiques des ménages influencent leur consommation d'énergie. En appliquant des techniques de sélection de caractéristiques et des algorithmes de clustering, j'ai identifié les variables clés et regroupé les ménages en segments significatifs pour mieux comprendre les facteurs déterminants et proposer des stratégies énergétiques plus efficaces.

3

Présentation de résultats

Tout au long de mon stage, j'ai régulièrement présenté les résultats de mes analyses lors de réunions avec mon tuteur. Ces présentations visaient à partager les progrès, discuter des résultats et recueillir des retours pour affiner les approches. J'utilisais PowerPoint pour structurer mes exposés, ce qui facilitait la communication sur les données, méthodes et conclusions. Après chaque présentation, je réexécutais les expérimentations en tenant compte des modifications suggérées, afin d'améliorer les analyses et affiner les résultats.

Les missions confiées

4

L'Etat de l'Art

L'état de l'art que j'ai réalisé dans le cadre de ce stage s'articule autour de quatre axes principaux : une exploration des articles sur l'Internet of Behaviors (IoB), une analyse des travaux sur l'extraction de features dans le domaine énergétique, une étude des articles traitant de la prédition de la consommation électrique, et enfin une revue des moyens d'influence de l'Internet of Things (IoT). Ces quatre perspectives permettent de dresser un panorama complet des enjeux et des avancées technologiques liés à l'interaction entre les comportements humains, la consommation d'énergie, et les dispositifs intelligents.

Articles sur l'Internet of Behaviors (IoB)

Étudier les articles sur l'Internet des Comportements m'a permis de bien me situer dans le contexte de cette technologie émergente et de comprendre ses implications profondes. Des publications récentes telles que "Internet of Behaviors: A Survey" (juillet 2023) et "Internet of Behaviors A literature review of an emerging technology" (2023) offrent un aperçu détaillé des développements récents dans le domaine. Ces articles montrent comment l'IoB, en s'appuyant sur l'Internet des Objets (IoT), intègre des méthodes avancées d'analyse des données comportementales pour influencer et prédire les comportements humains dans divers contextes, comme la santé, le marketing, et la gestion des ressources.

La compréhension de ces concepts a été enrichie par l'étude du travail de Gote Nyman, qui a introduit le concept d'IoB en associant les comportements humains à des adresses IoB spécifiques, chaque adresse faisant référence à un comportement. Cette approche permet une analyse plus précise des comportements et des intentions humaines, ouvrant la voie à des applications diverses.

Les articles comme "Internet of Behavior IoB - an alternative for diff" (2022) et "A Tutorial on Internet of Behaviors Concept Architecture Technology Applications and Challenges" (2022) m'ont permis de saisir les défis techniques liés à la collecte et à l'analyse des données comportementales, ainsi que les enjeux éthiques associés à la gestion des informations personnelles. Ces lectures ont également mis en lumière les différentes applications possibles de l'IoB, des systèmes de santé aux solutions de gestion des comportements dans les environnements commerciaux et domestiques.

En somme, la revue de ces articles a été cruciale pour comprendre non seulement les avancées et les applications de l'IoB, mais aussi les défis qu'il pose, me permettant ainsi de me positionner dans le cadre plus large de cette technologie en pleine évolution.

Articles sur l'extraction de features dans le domaine énergétique

Dans le cadre de l'état de l'art, j'ai analysé plusieurs articles scientifiques qui abordent l'extraction de caractéristiques dans le domaine de la consommation énergétique des ménages. Ces travaux de recherche présentent diverses approches méthodologiques pour identifier et extraire les variables les plus pertinentes, afin de mieux comprendre les facteurs influençant la consommation d'énergie résidentielle.

L'article de Sani et al. (2019) propose une méthode innovante combinant plusieurs techniques d'analyse pour sélectionner les variables les plus influentes en matière de consommation énergétique, à partir des données de consommation des ménages. Cette approche permet de regrouper les ménages selon des critères énergétiques similaires, facilitant ainsi l'identification de patterns de consommation.

De son côté, Heinrich et al. (2022) se concentrent sur la création d'archétypes comportementaux, visant à mieux comprendre les habitudes de consommation d'énergie dans le secteur résidentiel. En utilisant des techniques avancées de clustering et d'analyse des variables, leur étude met en lumière les comportements types des ménages, offrant des perspectives pour adapter les politiques énergétiques en fonction des profils identifiés.

Enfin, l'étude de Sanquist et al. (2011) explore la relation entre les styles de vie des ménages et leur consommation électrique. En recourant à l'analyse factorielle, les auteurs identifient les facteurs comportementaux les plus déterminants, permettant une réduction des dimensions et une meilleure compréhension des influences sur la consommation énergétique.

Ces diverses méthodologies m'ont permis de développer une compréhension approfondie des stratégies d'extraction de features appliquées à l'analyse de la consommation énergétique des ménages, ainsi que de leurs implications pour l'élaboration de modèles prédictifs et de politiques de gestion de l'énergie.

Articles sur la prédiction de la consommation électrique

Dans le domaine de la prévision de la consommation énergétique, une exploration des méthodes récentes révèle une tendance marquée vers l'utilisation des réseaux neuronaux profonds et des approches hybrides pour améliorer la précision des prédictions. L'analyse de plusieurs études a mis en évidence les avantages de ces techniques avancées pour capturer les schémas complexes de consommation.

La thèse de Mouna Labiad, une doctorante ayant précédemment mené ses recherches au sein de ce laboratoire, aborde la modélisation prédictive de la consommation énergétique des bâtiments. L'objectif principal de son travail est de développer une méthodologie capable de prédire la consommation énergétique des bâtiments, même lorsqu'aucune donnée opérationnelle historique n'est disponible pour le bâtiment cible, comme c'est souvent le cas dans les bâtiments nouvellement construits ou récemment rénovés.

Des recherches comme celles de Wang et al. (2021) montrent comment l'intégration de modèles de réseaux neuronaux convolutifs avec des approches de fusion de caractéristiques peut significativement améliorer les prévisions pour les bâtiments non résidentiels. Cette approche est complétée par des modèles tels que ceux décrits par Hadjout et al. (2023), qui combinent le traitement des valeurs aberrantes avec des techniques de clustering pour affiner les prévisions malgré les variations extrêmes dans les données.

L'étude des Transformers pour la prévision de la consommation électrique, comme exploré par Chan et Yeo (2022), illustre les avantages de ces modèles en termes de parallélisation et de rapidité de traitement, surmontant ainsi certaines limitations des méthodes traditionnelles. Les modèles hybrides, comme celui proposé par Syed et al. (2021), qui combinent LSTM avec d'autres techniques, démontrent également une amélioration significative dans la précision des prévisions à court terme pour les maisons individuelles.

De plus, les modèles hybrides CNN-LSTM, comme ceux développés par Alhussein et al. (2020), montrent comment l'association de différentes architectures peut offrir des prévisions plus robustes en tenant compte des comportements spécifiques des clients. Cependant, des défis demeurent, notamment la prise en compte des facteurs contextuels plus larges, tels que les variables économiques ou météorologiques.

En conclusion, l'étude de ces articles m'a permis de me familiariser avec les diverses méthodes actuelles utilisées pour la prévision de la consommation électrique. Cette compréhension approfondie des techniques modernes, allant des réseaux neuronaux profonds aux approches hybrides, est essentielle pour appréhender les avancées récentes et les défis à venir dans ce domaine.

Articles sur les moyens d'influences de l'Internet of Things (IoT)

Dans le cadre de l'état de l'art, j'ai examiné plusieurs articles récents portant sur les moyens d'influencer le comportement des systèmes IoT (Internet of Things) à travers l'Internet of Behavior (IoB) et des techniques d'intelligence artificielle avancées. Ces recherches explorent des approches novatrices pour prédire la consommation énergétique des ménages et ajuster les comportements des appareils connectés afin de minimiser la consommation d'énergie.

L'article "Decentralized IoB for Influencing IoT-based Systems Behavior" de Elayan et al. (2024) propose un framework IoB décentralisé qui utilise des modèles de prédiction pour anticiper la consommation énergétique et influencer directement le comportement des appareils IoT. Cette approche permet non seulement de prédire la consommation avec précision, mais aussi de prendre des mesures préventives pour éviter une consommation excessive d'énergie.

De même, dans "Internet of Behavior and Explainable AI Systems for Influencing IoT Behavior" (2023), les mêmes auteurs présentent un système intégré de gestion énergétique qui combine des techniques de modélisation avec un système d'explication des décisions prises. Ce système favorise une meilleure compréhension des processus de décision par les utilisateurs, tout en optimisant la consommation d'énergie des appareils IoT à travers un contrôle dynamique et transparent.

Ces travaux soulignent l'importance croissante de l'IoB dans la gestion de l'énergie domestique, en particulier lorsqu'il est associé à des systèmes d'intelligence artificielle capables d'expliquer et de justifier les décisions prises pour influencer le comportement des appareils connectés.

Conclusion de l'Etat de l'Art

En conclusion, l'exploration approfondie des articles récents sur l'Internet des comportements (IoB) et la prévision de la consommation énergétique a fourni une perspective précieuse sur les avancées actuelles dans ces domaines. L'Internet des comportements, en tant qu'évolution de l'Internet des objets (IoT), se distingue par sa capacité à collecter et analyser des données comportementales pour influencer et prédire les actions humaines.

Les travaux de chercheurs comme Gote Nyman ont jeté les bases de ce concept en soulignant comment les comportements peuvent être déduits à partir de données riches, offrant ainsi des opportunités pour améliorer les services dans divers secteurs, tels que la gestion de l'énergie, la santé, et le marketing digital. L'IoB se présente comme une tendance technologique clé, avec des prévisions de croissance significatives, et ses applications vont des systèmes de surveillance de foule à la gestion des comportements dans les environnements éducatifs.

L'étude de ces travaux a non seulement permis de saisir les mécanismes et les tendances actuelles dans ces domaines, mais a également mis en lumière les défis et les opportunités futurs. Que ce soit pour prédire les comportements des utilisateurs dans un contexte numérique en pleine expansion ou pour anticiper la consommation énergétique avec une précision accrue, ces avancées sont cruciales pour le développement de solutions intelligentes et efficaces. Les approches émergentes, tout en promettant des améliorations substantielles, nécessitent encore des innovations et des recherches continues pour surmonter les limites actuelles et maximiser leur potentiel dans des applications réelles.

Analyse de RECS2020

Contexte et objectifs

La population mondiale a atteint 8,2 milliards d'habitants en 2024, soit un milliard de plus qu'en 2010, et devrait franchir la barre des 10 milliards avant la fin du siècle selon les prévisions des Nations Unies. Cette croissance rapide, combinée à l'urbanisation et à l'augmentation des besoins énergétiques, contribue à une augmentation significative de la consommation énergétique des ménages, qui représente environ 25% de la consommation énergétique mondiale. En France métropolitaine, les ménages représentent aujourd'hui 36% de la consommation électrique, faisant de ce secteur le plus énergivore en proportion, devant les secteurs tertiaire et industriel. Cette tendance suscite d'importantes inquiétudes quant à la durabilité énergétique, notamment au regard de la tension croissante sur le réseau électrique, de l'augmentation des émissions de carbone, des défis de la transition vers les énergies renouvelables et de la nécessité d'équilibrer la production d'électricité avec la demande croissante.

Parallèlement, l'essor des objets connectés (IoT), dont le nombre devrait atteindre 50 milliards d'unités d'ici 2025 et 100 milliards d'ici 2030, ouvre de nouvelles perspectives pour le suivi et l'optimisation de la consommation énergétique. Ces appareils collectent des quantités massives de données sur les habitudes, les préférences et les comportements des utilisateurs, offrant une opportunité sans précédent de comprendre et de gérer plus efficacement la consommation énergétique.

L'Internet du comportement (IoB) est un nouveau paradigme visant à analyser les données liées aux tendances comportementales des utilisateurs pour atteindre des objectifs spécifiques. Ces données, collectées par les appareils intelligents, reflètent le comportement, les habitudes et le mode de vie des utilisateurs dans divers domaines tels que la santé, les transports, l'éducation et la consommation d'énergie. L'IoB permet de développer des modèles de gestion de l'énergie dans les espaces résidentiels en observant comment le comportement des occupants influence la consommation d'énergie. En surveillant en continu le comportement des utilisateurs, l'IoB fournit des informations précieuses sur la manière dont la consommation d'énergie peut être optimisée.

Cependant, le défi demeure : comment interpréter ces données pour développer des modèles efficaces de gestion de l'énergie dans les espaces résidentiels, basés sur le comportement des occupants ? Alors que les approches existantes se concentrent principalement sur la sélection de caractéristiques liées aux appareils utilisés ou aux logements, notre approche se distingue par une focalisation sur les caractéristiques comportementales. La façon dont les membres du ménage utilisent leurs appareils, comme la fréquence d'utilisation et le mode de fonctionnement, peut grandement influencer la consommation globale d'énergie. Nous cherchons à comprendre comment les comportements humains influencent la consommation d'énergie, offrant ainsi une analyse plus détaillée et une opportunité d'améliorer l'efficacité énergétique globale.

Dans ce cadre, ce stage s'est fixé pour objectif d'analyser l'influence des comportements des ménages sur leur consommation d'énergie, en vue d'identifier des modèles de consommation distincts et de comprendre les facteurs clés qui les sous-tendent. Pour ce faire, nous avons travaillé sur le jeu de données RECS2020, qui contient des informations détaillées sur la consommation énergétique des ménages américains en 2020. Notre méthodologie a consisté à collecter et prétraiter les données comportementales, à sélectionner les caractéristiques pertinentes, puis à appliquer des algorithmes de clustering afin de segmenter les ménages en groupes significatifs.

Les objectifs spécifiques incluaient l'identification et la caractérisation des différents groupes de ménages en fonction de leurs comportements de consommation énergétique et de leurs caractéristiques socio-économiques. Nous avons également analysé les principaux facteurs influençant la consommation au sein de chaque groupe. Enfin, sur la base de ces analyses, nous avons formulé des recommandations ciblées pour réduire la consommation d'énergie, en tenant compte des spécificités de chaque groupe pour maximiser l'efficacité des mesures proposées.

Le dataset

L'enquête sur la consommation énergétique résidentielle (RECS) est une enquête exhaustive menée aux États-Unis pour recueillir des informations sur la consommation énergétique résidentielle des ménages américains. Réalisée par l'Energy Information Administration (EIA), cette enquête est l'une des sources de données les plus complètes et les plus fiables sur la consommation énergétique du secteur résidentiel.

L'édition 2020 de la RECS recueille des données sur divers aspects de la consommation énergétique des ménages, notamment les types et quantités de combustibles utilisés tels que l'électricité, le gaz naturel, le propane et le fioul, les caractéristiques du logement, les appareils électroménagers, les systèmes de chauffage et de climatisation, ainsi que les comportements récents tels que l'utilisation de l'énergie solaire ou l'emplacement de la recharge des véhicules électriques. L'enquête couvre ainsi un large éventail de 799 variables, permettant une analyse approfondie et multidimensionnelle des habitudes de consommation énergétique. La RECS utilise un échantillon représentatif de 18496 ménages américains pour garantir que les résultats peuvent être généralisés à l'ensemble de la population. Des enquêtes détaillées sont envoyées à des ménages sélectionnés et peuvent être administrées sous forme de questionnaires papier, en ligne, par entretiens téléphoniques ou lors de visites sur place.

Pour plus d'informations, vous pouvez consulter le site de l'EIA.

Pré-traitements

Afin de garantir la qualité et la pertinence des analyses, plusieurs étapes de prétraitement ont été appliquées aux données RECS 2020. Ces étapes sont cruciales pour nettoyer et préparer les données avant d'appliquer des techniques de sélection de caractéristiques et des algorithmes de clustering.

Suppression des variables concernant les énergies autres que l'électricité

Puisque notre recherche vise à identifier les comportements des utilisateurs influençant la consommation d'électricité, nous avons retiré 117 variables liées à d'autres types d'énergie, comme le gaz naturel, le propane, le fioul et le bois. En excluant ces variables, nous avons pu concentrer notre analyse spécifiquement sur les facteurs influençant la consommation d'électricité. Cette approche réduit la complexité de l'ensemble de données et facilite l'interprétation des résultats. Après ce retrait, nous avons conservé un total de 682 variables.

Suppression des variables indicatrices d'imputation et de calibration

Les variables indicatrices d'imputation montrent si les valeurs des autres variables de l'ensemble de données ont été imputées, ce qui signifie qu'elles représentent des estimations plutôt que des données réelles. Les variables d'étalonnage ajustent les pondérations des réponses pour garantir que l'échantillon reflète avec précision la population dans son ensemble. Pour affiner notre analyse et garantir la pertinence des données, nous avons supprimé 407 de ces variables utilisées par l'EIA. Bien que ces variables soient essentielles aux processus internes de l'EIA, elles ne sont pas nécessaires à notre étude, qui se concentre spécifiquement sur les comportements de consommation d'électricité. Après ces suppressions, il reste 275 variables pour notre analyse.

Encodage des variables catégorielles

Pour préparer notre ensemble de données à une analyse plus approfondie, nous avons converti les variables catégorielles en un format compréhensible par les algorithmes d'apprentissage automatique. Nous avons utilisé la méthode LabelEncoder de la bibliothèque scikit learn. Cette technique transforme chaque valeur catégorielle en une valeur numérique unique, indiquant la catégorie à laquelle elle appartient. Dans notre ensemble de données, nous avons identifié 7 variables catégorielles qui nécessitaient un codage. L'application de LabelEncoder à ces variables nous a permis de convertir les données catégorielles en une représentation numérique tout en préservant l'intégrité et la signification des informations d'origine.

Gestion des valeurs manquantes

Afin de garantir la qualité et l'exhaustivité de notre ensemble de données, nous avons adopté une approche pragmatique pour gérer les valeurs manquantes. La première étape a consisté à identifier toutes les lignes de l'ensemble de données qui contenaient une ou plusieurs valeurs manquantes. Nous avons identifié 290 échantillons contenant une valeur NaN, ce qui représente environ 1,57 % de l'ensemble de données. Cette proportion étant très faible, nous avons décidé de supprimer ces échantillons pour simplifier le traitement des données. Après avoir supprimé les lignes contenant des valeurs manquantes, le nombre de ménages dans notre ensemble de données est passé de 18 496 à 18 206. Cette approche garantit que notre analyse repose sur un ensemble de données complet, minimisant ainsi les biais et les erreurs potentielles liés à l'imputation des valeurs manquantes.

Réduction des variables cibles

Au début de notre étude, nous avons identifié trois principales variables cibles liées à la consommation d'électricité : la quantité totale d'électricité consommée en kilowattheures (kWh), la quantité totale d'électricité consommée en BTU (BTUEL) et le coût total de l'électricité consommée en dollars (DOLLAREL).

Cependant, après avoir effectué une analyse de corrélation de Pearson entre ces variables, nous avons constaté que certaines d'entre elles étaient fortement corrélées. En particulier, la variable BTUEL était fortement corrélée avec les kWh, car ces deux mesures sont interconvertibles (1 kWh équivaut à environ 3 412 BTU). Par conséquent, nous avons décidé de réduire le nombre de variables cibles à deux : kWh et DOLLAREL. Cette réduction simplifie l'analyse tout en conservant les mesures essentielles pour évaluer la consommation d'électricité. Suite à cette réduction, nous avons ajouté ces 2 variables au total de notre ensemble de données, portant le nombre de variables à 277.

Réduction du nombre de variables

Dans le cadre du prétraitement des données, nous avons réalisé une analyse de corrélation de Pearson afin d'identifier les paires de variables fortement corrélées dans notre jeu de données et de réduire la redondance. Cette analyse a conduit à la suppression de 69 variables. Par exemple, nous avons supprimé la variable "state_postal" et conservé "state_name" pour éviter la redondance entre les codes postaux des États et leurs noms. De même, nous avons supprimé la variable "TELLWORK", qui indique qu'un membre du ménage télétravaille, au profit de variables plus précises telles que "TELLEDAYS", "TLDESKTOP" ou "TLLAPTOP", qui apportent davantage d'informations sur les modalités de télétravail. Cette approche a également été appliquée à d'autres domaines, tels que les aides énergétiques (PAYHELP), les véhicules électriques (ELECVEH) ou le type de logement (STUDIO). La matrice de corrélation ci-dessous illustre les variables fortement corrélées et étaye nos décisions de réduction des variables en montrant comment certaines variables sont étroitement liées. Après ce processus de réduction, 69 variables ont été supprimées, laissant un total de 208 variables pour une analyse plus approfondie.

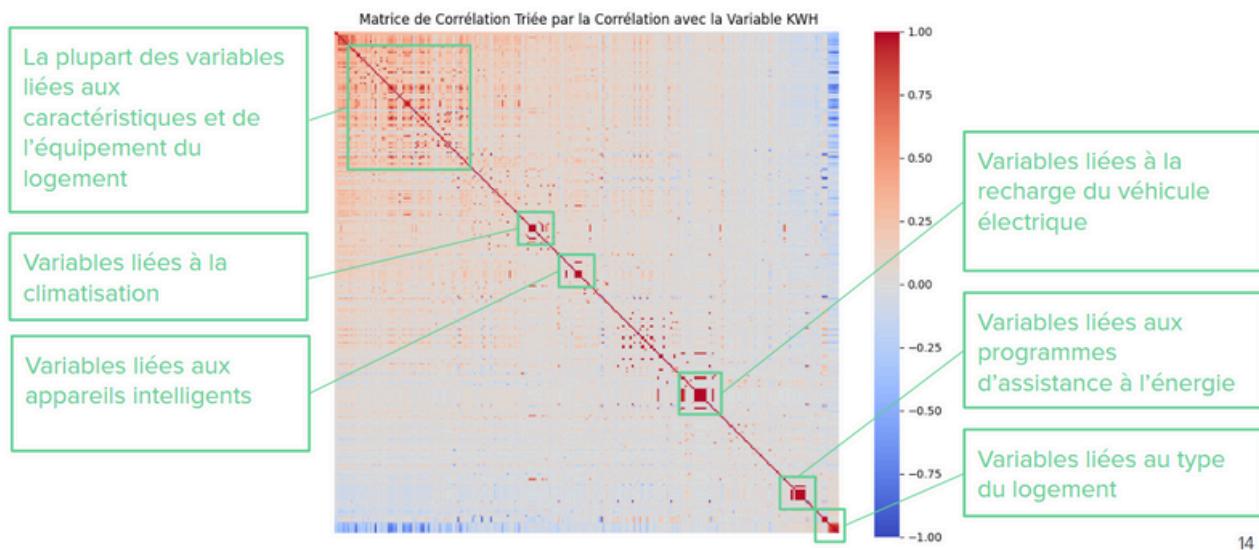


Figure 1 : Matrice de corrélation

Réduction du nombre d'échantillons

Les données isolées, souvent considérées comme des valeurs aberrantes, peuvent biaiser les modèles de machine learning et fausser les résultats des analyses statistiques. Il est donc essentiel de les identifier et de les supprimer pour obtenir des analyses plus précises et plus fiables, permettant ainsi une meilleure compréhension des données. Dans notre étude, nous avons utilisé l'algorithme Isolation Forest pour identifier ces données isolées. Isolation Forest isole les anomalies en construisant des arbres d'isolement. Chaque arbre partitionne aléatoirement les données jusqu'à ce que chaque point soit isolé dans une feuille. Les anomalies, étant rares et différentes, sont isolées plus rapidement, ce qui signifie qu'elles nécessitent moins de fractionnements pour être séparées des autres points. Chaque point de données se voit attribuer un score d'anomalie basé sur la profondeur moyenne à laquelle il est isolé dans les arbres. Plus la profondeur est faible, plus le point est susceptible d'être une anomalie. Cette méthode a détecté 894 échantillons considérés comme des valeurs aberrantes, représentant environ 5% de l'ensemble de données. La suppression de ces points a réduit le nombre total de ménages dans notre ensemble de données de 18206 à 17312. En supprimant ces points de données isolés, nous avons considérablement amélioré la qualité de l'ensemble de données, permettant des analyses plus précises et plus robustes.

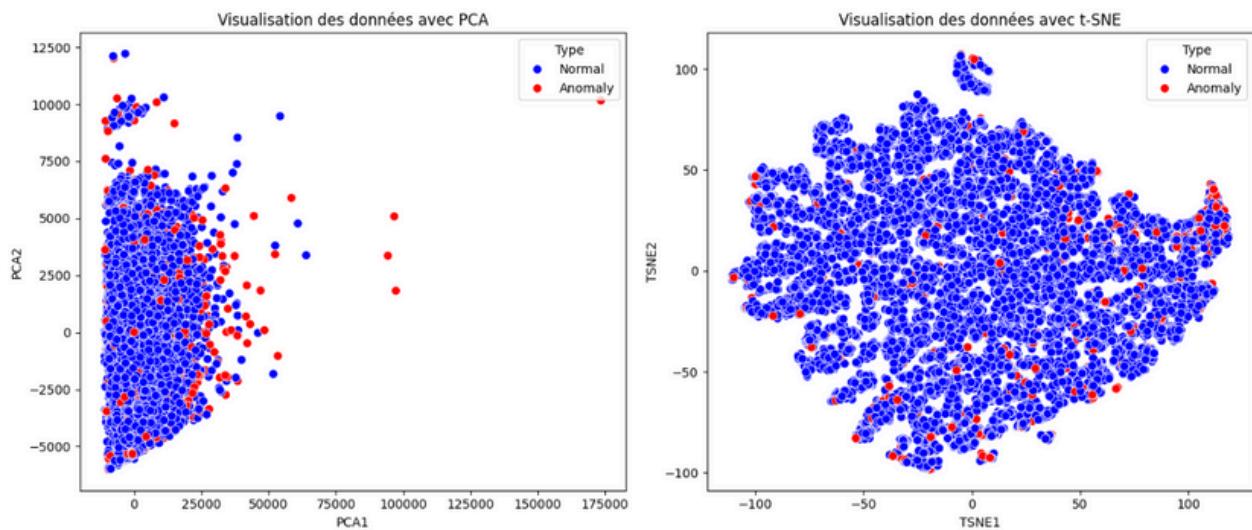


Figure 2 : Visualisations des données aberrantes

Conclusion du pré-traitement

À l'issue des étapes rigoureuses de prétraitement, nous avons considérablement affiné notre dataset pour le rendre plus pertinent et exploitable pour notre analyse.

Le nombre d'échantillons a été réduit de 18 496 à 17 312, suite à la suppression de 894 points aberrants identifiés comme isolés et de 290 échantillons contenant au moins une valeur manquante. Cette réduction a permis de maintenir une base de données plus cohérente et représentative.

En ce qui concerne les variables, nous avons simplifié notre dataset en réduisant le nombre total de variables de 799 à 208. Cette diminution a été réalisée en éliminant les variables relatives à d'autres types d'énergie, les variables indicatrices d'imputation et de calibration, ainsi que celles fortement corrélées. Cette démarche a permis de réduire les redondances et de concentrer notre analyse sur les facteurs les plus pertinents.

Ces étapes de prétraitement ont non seulement optimisé la qualité des données, mais ont également facilité des analyses plus précises et significatives. En fin de compte, ces ajustements permettent d'obtenir des résultats plus fiables et d'approfondir notre compréhension des comportements de consommation électrique.

Sélection des variables

La sélection des variables est une étape cruciale dans le prétraitement des données, visant à améliorer la qualité des modèles prédictifs en identifiant les variables les plus pertinentes. Dans notre étude, nous avons appliqué plusieurs méthodes pour affiner notre choix de variables.

Méthode 1 : Corrélation sur l'ensemble des données

La première méthode de sélection des variables repose sur l'analyse de la corrélation entre toutes les variables du dataset et les variables cibles.

Nous avons d'abord calculé la corrélation entre chaque variable et la variable cible KWH pour identifier celles ayant une relation significative avec la consommation électrique. Les variables dont la corrélation avec KWH était supérieure à la moyenne des corrélations calculées ont été sélectionnées.

De manière similaire, nous avons évalué la corrélation entre chaque variable et la deuxième variable cible, DOLLAREL, pour identifier les variables influençant le coût de la consommation d'électricité. Les variables dont la corrélation avec DOLLAREL était également supérieure à la moyenne des corrélations calculées ont été retenues.

Enfin, les variables pertinentes pour chacune des deux cibles ont été croisées pour obtenir leur intersection, ce qui a permis de sélectionner 76 variables importantes pour les deux variables cibles, garantissant leur pertinence dans les deux contextes d'analyse.

Méthode 2 : Corrélation sur les données catégorisées

La deuxième méthode de sélection des variables repose sur la catégorisation des variables du dataset et l'analyse de la corrélation au sein de ces catégories. Cette approche permet de structurer le processus de sélection en traitant les variables selon leurs catégories spécifiques.

L'EIA a partitionné les variables en 14 catégories distinctes. Nous avons ajouté une catégorie supplémentaire spécifiquement dédiée aux variables liées au comportement, comprenant 88 variables identifiées manuellement. Cette approche structurée a permis de sélectionner 71 variables en calculant la corrélation au sein de chaque catégorie et en choisissant celles ayant une corrélation significative avec les cibles.

Méthode 3 : Gain d'information sur l'ensemble des données

Le DecisionTreeRegressor est un algorithme de régression qui construit un arbre de décision pour prédire des valeurs continues, telles que la consommation d'énergie (KWH) ou le coût (DOLLAREL).

L'arbre de décision divise les données en posant des questions successives sur les caractéristiques des données d'entraînement. Pour chaque nouvelle observation, l'arbre suit les branches correspondant aux caractéristiques de l'observation et prédit une valeur basée sur la moyenne des valeurs des données qui arrivent à la feuille terminale.

L'algorithme évalue l'importance de chaque caractéristique en mesurant la réduction de l'impureté, quantifiée par l'erreur quadratique apportée par chaque division de l'arbre. À chaque nœud, il sélectionne la caractéristique qui minimise cette impureté, optimisant ainsi les prédictions.

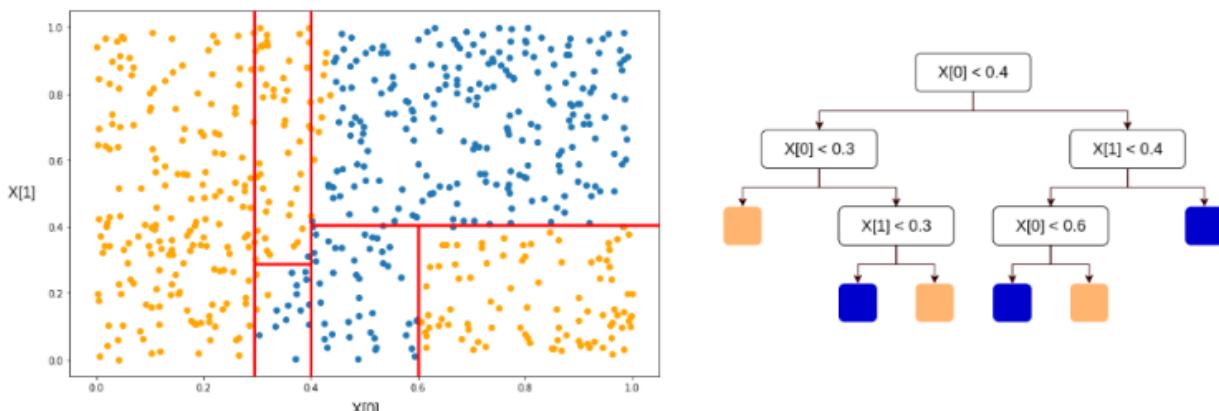


Figure 3 : Illustration du fonctionnement de DecisionTreeRegressor

Après avoir déterminé l'importance de chaque caractéristique, nous classons ces caractéristiques en fonction de leur contribution à la réduction de l'erreur quadratique globale. Une courbe cumulative est ensuite tracée pour visualiser la contribution de chaque caractéristique. En identifiant le point où la courbe se stabilise, nous pouvons déterminer le nombre optimal de caractéristiques à retenir, équilibrant ainsi précision et simplicité du modèle. Cette approche permet de sélectionner les variables les plus pertinentes et de construire un modèle de régression efficace, capable de faire des prédictions précises tout en évitant le surapprentissage. Appliquée à l'ensemble des données prétraitées, cette méthode a permis de sélectionner 63 variables.

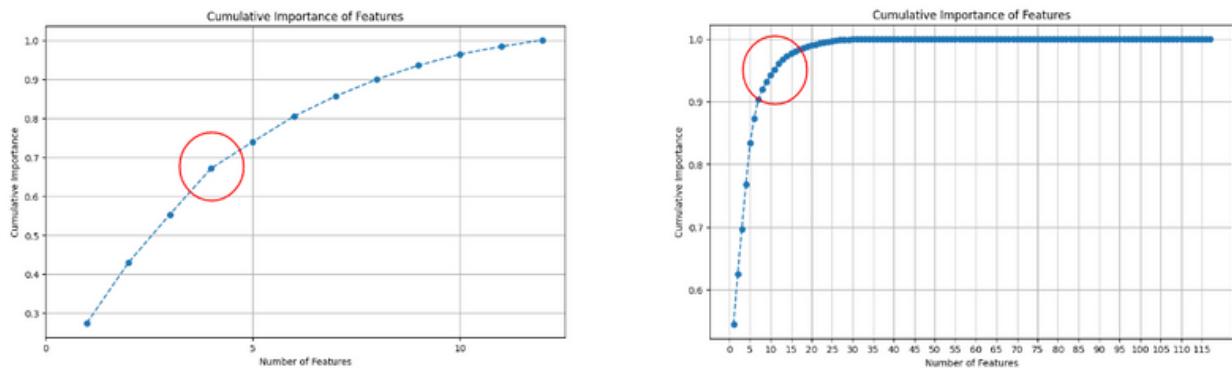


Figure 4 : Exemples d'application de la méthode du coude

Méthode 4 : Gain d'information sur les données catégorisées

Comme mentionné précédemment, les données d'origine sont organisées en différentes catégories. Pour chaque catégorie, nous avons appliqué la méthode du coude afin d'identifier le nombre optimal de variables à conserver. Cette approche consiste à tracer la réduction cumulative de l'impureté en fonction du nombre de variables retenues et à déterminer le seuil au-delà duquel l'ajout de variables supplémentaires n'améliore que marginalement la réduction d'impureté.

En appliquant cette méthode de manière individuelle à chaque catégorie, nous avons pu sélectionner 83 variables pertinentes. Cette approche permet d'assurer que les variables choisies contribuent de manière significative à la prédiction tout en maintenant la simplicité et l'efficacité du modèle.

Intersection des 4 méthodes

En croisant les résultats obtenus par les quatre méthodes de sélection, cela nous a permis d'identifier 17 variables parmi les 208 initiales ayant le plus grand impact sur la consommation d'énergie. Cette sélection conjointe met en lumière les variables les plus significatives en combinant les perspectives des différentes méthodes. Parmi ces 17 variables, 9 proviennent des 88 variables comportementales, soulignant leur importance particulière dans l'analyse de la consommation d'énergie.

Analyse des variables sélectionnées

Après le prétraitement, nous avons identifié 17 variables présentant la plus forte influence sur les variables cibles. Ces variables ont été analysées à l'aide de méthodes de clustering non supervisées, étant donné l'absence d'étiquettes dans les données.

Détermination du nombre optimal de clusters

Pour déterminer le nombre optimal de clusters dans notre analyse, nous avons utilisé quatre méthodes complémentaires : la méthode du coude avec l'inertie, le silhouette score, le score de Davies-Bouldin, et le critère de Calinski-Harabasz. Chaque méthode fournit une perspective différente sur la qualité du clustering et aide à identifier le nombre de clusters le plus approprié pour représenter au mieux la structure des données.

- **Méthode du coude avec l'inertie :** La méthode du coude est une approche classique pour déterminer le nombre optimal de clusters. Elle repose sur l'inertie, également appelée somme des carrés intra-cluster. L'inertie mesure la dispersion des points de données au sein de chaque cluster : plus l'inertie est faible, plus les points sont proches du centre du cluster. En traçant l'inertie en fonction du nombre de clusters, nous obtenons une courbe qui montre généralement une réduction rapide au début, suivie d'une stabilisation. Le "coude" de cette courbe, où la réduction de l'inertie commence à diminuer de manière significative, indique le nombre optimal de clusters. Ce point représente un équilibre entre la compacité des clusters et la complexité du modèle.
- **Silhouette score :** Le silhouette score évalue la qualité du clustering en mesurant la similarité des points au sein de leur propre cluster par rapport à la similarité avec les points des autres clusters. Le score varie de -1 à 1 : un score proche de 1 indique que les points sont bien regroupés dans leur propre cluster et éloignés des autres clusters, tandis qu'un score proche de -1 suggère que les points pourraient être mal classifiés. Pour différents nombres de clusters, nous avons calculé le silhouette score moyen. Le nombre de clusters qui maximise ce score est considéré comme optimal, car il indique la meilleure séparation et la plus grande cohésion entre les clusters.
- **Score de Davies-Bouldin :** Le score de Davies-Bouldin est une autre mesure de la qualité du clustering, qui évalue la compacité et la séparation des clusters. Il est calculé en mesurant la moyenne des ratios de similarité entre chaque cluster et le cluster le plus similaire. Un score plus bas indique une meilleure séparation entre les clusters, suggérant que les clusters sont bien distincts et compacts. En comparant les scores pour différents nombres de clusters, le nombre de clusters qui minimise le score de Davies-Bouldin est choisi comme optimal, car il reflète la meilleure séparation des groupes.
- **Critère de Calinski-Harabasz :** Le critère de Calinski-Harabasz, également connu sous le nom de "variance ratio criterion", évalue la qualité du clustering en comparant la variance intra-cluster à la variance inter-cluster. Plus ce ratio est élevé, meilleure est la séparation entre les clusters et plus les clusters sont compacts. Nous avons calculé ce critère pour différents nombres de clusters et sélectionné le nombre qui maximise le critère de Calinski-Harabasz. Cela indique que les clusters sont bien définis avec une séparation claire et une cohésion interne élevée.

L'application des quatre méthodes de sélection du nombre optimal de clusters a conduit à une conclusion convergente : toutes les méthodes suggèrent que le nombre optimal de clusters est 2. L'accord entre ces différentes méthodes renforce la conclusion selon laquelle deux clusters sont le choix le plus approprié pour capturer la structure sous-jacente des données. Ce consensus multi-méthodes valide la robustesse et la fiabilité de cette solution, garantissant ainsi une représentation adéquate des groupes au sein de notre analyse.

K-means

Le K-means est une méthode de clustering non supervisée qui vise à diviser un ensemble de données en un nombre fixe de groupes, appelés clusters. L'objectif est de minimiser la variance au sein des clusters et de maximiser la variance entre les clusters.

L'algorithme K-means fonctionne en plusieurs étapes. L'algorithme commence par l'initialisation des centres des clusters. Ces centres peuvent être déterminés aléatoirement ou en utilisant des techniques spécifiques comme K-means++ pour améliorer la sélection initiale des centres. Chaque point de données est assigné au cluster dont le centre est le plus proche. Cette proximité est généralement mesurée par la distance euclidienne entre les points de données et les centres des clusters. Une fois que tous les points ont été assignés à des clusters, les centres des clusters sont recalculés en prenant la moyenne des points appartenant à chaque cluster. Les étapes d'assignation des points et de mise à jour des centres sont répétées jusqu'à ce que les centres des clusters convergent, c'est-à-dire que les changements deviennent négligeables, ou jusqu'à ce qu'un critère d'arrêt soit atteint (tel qu'un nombre maximal d'itérations). L'algorithme s'arrête donc lorsque les centres des clusters ne varient plus de manière significative, ce qui indique que les clusters ont atteint une forme de stabilité.

Dans le cadre de cette recherche, l'algorithme K-means a été appliqué en utilisant le nombre optimal de clusters déterminé par les méthodes précédentes, établi à deux clusters.

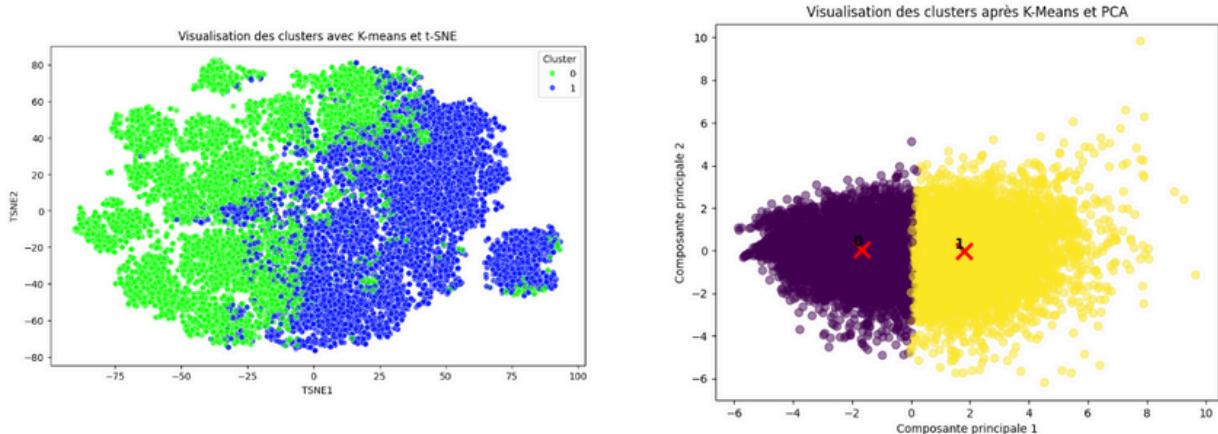


Figure 5 : Visualisation du clustering des variables sélectionnées avec K-Means

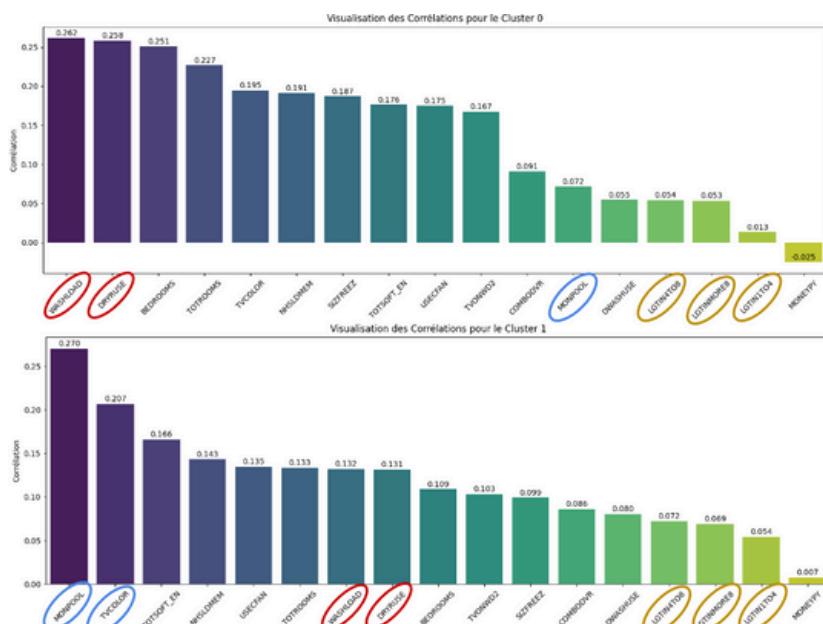


Figure 6 : Corrélation des variables pour les deux clusters identifiés par K-Means

Kernel K-means

Kernel K-means est une extension de l'algorithme K-means classique qui utilise une fonction noyau pour mapper les données dans un espace de dimension plus élevée, permettant ainsi de capturer des structures non linéaires dans les données. En transformant les données à l'aide d'une fonction noyau, Kernel K-means peut identifier des clusters qui ne seraient pas détectables par K-means classique, car ils se trouvent dans un espace où les frontières entre les clusters ne sont pas linéaires.

L'algorithme suit des étapes similaires à K-means : initialisation des centres de clusters, assignation des points aux clusters en fonction des distances calculées avec la matrice noyau, mise à jour des centres de clusters, et itération jusqu'à convergence. Cependant, au lieu d'utiliser la distance euclidienne traditionnelle, Kernel K-means utilise les distances calculées dans l'espace transformé par la fonction noyau. Cette approche permet de mieux capturer des relations complexes entre les points de données.

Dans notre étude, après avoir appliqué K-means, nous avons utilisé Kernel K-means pour vérifier si des structures plus complexes existaient dans les données. Les résultats ont confirmé que les clusters identifiés étaient robustes et nous ont permis de mieux comprendre les comportements de consommation énergétique des ménages.

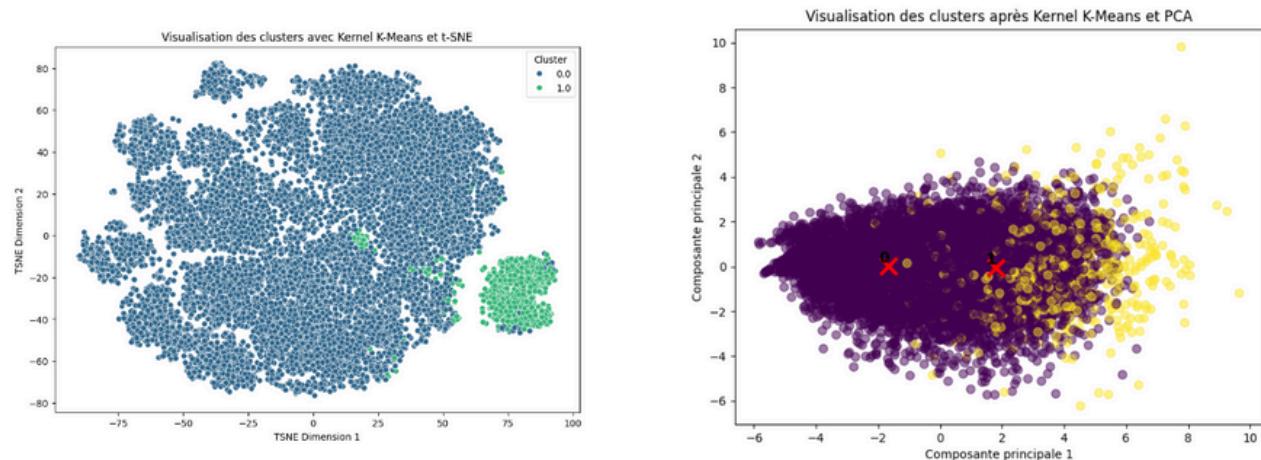


Figure 7 : Visualisation du clustering des variables sélectionnées avec Kernel K-Means

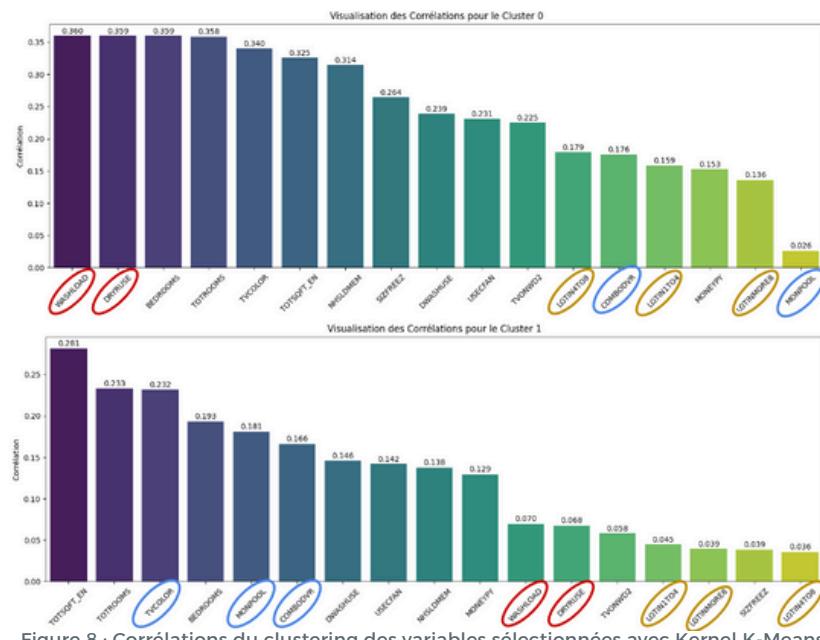


Figure 8 : Corrélations du clustering des variables sélectionnées avec Kernel K-Means

DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est un algorithme de clustering non supervisé largement utilisé pour identifier des clusters dans des ensembles de données en se basant sur la densité des points de données. Contrairement à d'autres méthodes de clustering telles que K-Means, qui nécessitent de définir le nombre de clusters à l'avance, DBSCAN peut identifier automatiquement le nombre de clusters en fonction de la densité locale des points. L'algorithme repose sur deux principaux paramètres : ε (epsilon), qui représente la distance maximale entre deux points pour qu'ils soient considérés comme voisins, et minPts, le nombre minimum de points dans le voisinage ε pour qu'un point soit considéré comme un point central (core point).

DBSCAN commence par sélectionner un point non visité dans l'ensemble de données et identifie tous les points dans son voisinage ε . Si le nombre de points voisins est supérieur ou égal à minPts, le point est considéré comme un point central et un nouveau cluster est créé. Si le nombre de points voisins est inférieur à minPts, le point est marqué comme du bruit, bien qu'il puisse être inclus plus tard dans un cluster s'il est un voisin d'un point central. Pour chaque point central, l'algorithme explore tous les points dans son voisinage et, si ces points sont également des points centraux, leur voisinage est également exploré. Ce processus continue jusqu'à ce que le cluster ne puisse plus être étendu. L'algorithme répète ces étapes jusqu'à ce que tous les points soient visités.

DBSCAN présente plusieurs avantages. Il peut identifier des clusters de formes arbitraires, contrairement à K-Means qui suppose des clusters sphériques, et traite explicitement les points de bruit, ce qui améliore la qualité des clusters formés. De plus, il détermine automatiquement le nombre de clusters en fonction de la densité des points, ce qui est avantageux lorsque le nombre de clusters n'est pas connu à l'avance. Il est également efficace pour les bases de données contenant de grands volumes de données et fonctionne bien avec les bases de données de densité variable.

L'application de l'algorithme DBSCAN à notre ensemble de données, avec des paramètres ε définis à 2 et minPts fixé à 19, a révélé la formation de quatre clusters distincts. Cependant, il est important de noter que ces clusters sont relativement petits en taille. La majorité des points de données ont été classés comme des anomalies, reflétant la nature stricte de l'algorithme dans la définition des points de cluster en fonction de la densité locale. Cette caractéristique de DBSCAN souligne sa capacité à identifier et isoler les points aberrants dans les ensembles de données, mais elle peut également indiquer que les paramètres ε et minPts pourraient nécessiter un ajustement pour mieux capturer les structures de clusters potentielles présentes dans les données.

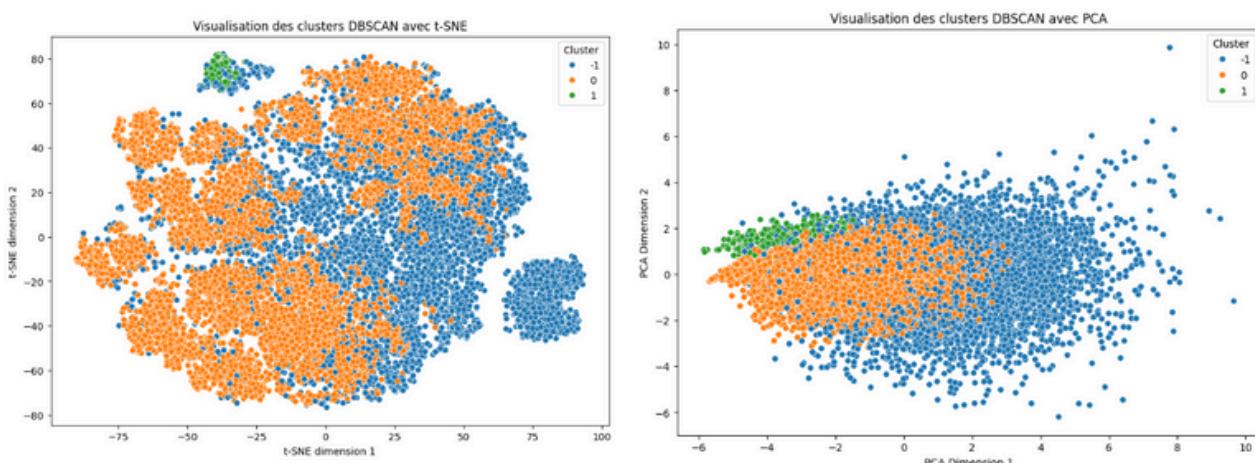


Figure 9 : Visualisation du clustering des variables sélectionnées avec DBSCAN

Analyse des clusters

Les deux méthodes de clustering, K-means et Kernel K-means, ont segmenté les ménages en deux groupes distincts, chacun mettant en évidence des caractéristiques spécifiques de consommation énergétique.

Les ménages appartenant au premier cluster se caractérisent par des logements plus petits, un revenu plus faible, et une possession et utilisation moindre des équipements. Les principaux facteurs de consommation d'énergie dans ce cluster sont liés aux besoins essentiels d'un logement, principalement l'utilisation d'appareils électroménagers pour des tâches quotidiennes comme la cuisson, la réfrigération et le lavage.

Les ménages du second cluster, en revanche, se distinguent par des logements plus grands, un revenu élevé, et une utilisation beaucoup plus intensive des appareils électroménagers et électroniques. Dans ce cluster, les principaux facteurs de consommation d'énergie incluent des équipements liés aux loisirs, tels que la piscine, qui sont largement responsables de la consommation élevée d'électricité.

K-Means a identifié deux clusters distincts. Le premier cluster comprend des ménages avec un revenu moyen compris entre 30 000 et 35 000 dollars par an, une superficie de logement moyen de 125 mètres carrés, et des habitudes de consommation principalement axées sur les besoins essentiels. Le second cluster, en revanche, inclut une large gamme de ménages à revenu élevé, avec un revenu annuel moyen compris entre 60 000 et 75 000 dollars, une superficie de logement moyen de 230 mètres carrés, une utilisation quotidienne d'électroménagers tels que le sèche-linge, et une consommation de lumière deux fois plus importante que celle du premier cluster. Les conclusions montrent ainsi une distinction claire entre les ménages à faible revenu avec des besoins essentiels et ceux à revenu élevé avec des équipements de loisir.

Kernel K-Means a identifié un sous-groupe spécifique de ménages très riches dans le second cluster, réduisant ainsi le nombre total d'échantillons dans ce cluster comparé à K-Means. Ce sous-groupe se distingue par des ménages possédant un logement encore plus grand, avec une superficie moyenne de 260 mètres carrés, et une utilisation accrue de tous les équipements. Cela signifie que Kernel K-Means a pu identifier un groupe de ménages extrêmement aisés, dont les habitudes de consommation énergétique sont encore plus marquées par des équipements de loisir et des comportements de consommation luxueux.

Les différences entre les deux clusters mettent en lumière les divers aspects de la consommation énergétique des ménages. Pour le premier cluster, la consommation d'énergie est principalement due à des besoins essentiels, reflétant des ménages vivant dans des logements modestes avec des revenus plus faibles. Les stratégies de réduction de la consommation d'énergie dans ce cluster pourraient se concentrer sur l'efficacité énergétique des appareils électroménagers et des pratiques quotidiennes. Pour le second cluster, la consommation d'énergie élevée est majoritairement attribuée à des équipements de loisir tels que la piscine ou la box satellite, caractéristiques de ménages à revenu élevé vivant dans de grands logements. Les efforts pour réduire la consommation d'énergie dans ce groupe pourraient inclure des mesures pour rendre ces équipements plus efficaces ou promouvoir des alternatives moins énergivores.

Pour compléter cette analyse, nous avons également appliqué l'algorithme DBSCAN, qui a également identifié deux clusters distincts. Cependant, il convient de noter que la majorité des points de données ont été classés comme des anomalies. Contrairement aux résultats des algorithmes précédents, les deux clusters détectés par DBSCAN regroupent principalement des ménages avec des revenus plus faibles. Bien que ces ménages présentent des différences en termes de revenus et de taille de logement, ils montrent des comportements similaires en termes de faible utilisation des équipements et appareils domestiques. Une distinction notable est que l'un de ces clusters ne possède pas de télévision du tout.

En somme, cette analyse révèle que les comportements de consommation énergétique diffèrent grandement selon le revenu et la taille du logement, influençant directement les stratégies potentielles d'économie d'énergie à adopter pour chaque groupe. Ces résultats mettent en lumière l'importance de cibler des mesures spécifiques d'efficacité énergétique adaptées aux caractéristiques et aux besoins des différents segments de la population, afin de maximiser les économies d'énergie et d'améliorer la durabilité environnementale.

Analyse des variables comportementales sélectionnées

Nous avons maintenant concentré notre attention sur les variables comportementales parmi celles qui ont été sélectionnées. Rappelons que l'objectif initial de notre étude était d'analyser l'influence du comportement sur la consommation énergétique. Parmi les 17 variables résultant de la sélection des variables, 9 sont spécifiquement liées aux comportements des ménages. En isolant ces variables comportementales, nous pouvons examiner plus en détail comment les habitudes et pratiques des ménages impactent leur consommation énergétique.

Détermination du nombre optimal de clusters

Pour déterminer le nombre optimal de clusters en analysant les variables comportementales, nous avons utilisé de la même façon les méthodes d'évaluation utilisées précédemment. La méthode du coude et le critère de Calinski-Harabasz ont tous deux indiqué que le nombre optimal de clusters est 2. En revanche, le silhouette score a suggéré 7 clusters, tandis que le score Davies-Bouldin a recommandé 13 clusters. Ces divergences montrent l'importance de considérer plusieurs critères pour une évaluation robuste du nombre de clusters, chaque méthode offrant une perspective différente sur la structure des données.

Analyse des variables comportementales

Par la suite, j'ai appliqué les algorithmes de K-means, Kernel K-means et DBSCAN pour segmenter les données selon ces différents nombres de clusters. Pour chaque segmentation, j'ai mené une analyse approfondie des résultats en examinant la moyenne des variables dans chaque cluster, les corrélations entre les comportements et la consommation d'énergie, ainsi que la distribution des données au sein des clusters. Cette approche m'a permis d'identifier les principaux facteurs influençant la consommation énergétique, tels que l'utilisation des piscines, des appareils électroménagers, et d'autres habitudes spécifiques.

Un exemple des résultats de ces segmentations est présenté ci-dessous. Les autres visualisations, représentant les regroupements en 3, 10 clusters et DBSCAN, sont disponibles en annexe. Ces images permettent de mieux comprendre comment les groupes se forment et se distinguent en fonction des variables comportementales et de consommation énergétique.

En comparant les résultats obtenus avec ces différentes méthodes de clustering, j'ai pu confirmer la robustesse des conclusions et révéler des variations marquées dans les comportements énergétiques selon les profils socio-économiques et géographiques des ménages. Cette mission m'a permis de développer des compétences en analyse de données, en interprétation des résultats, et en application de méthodes de machine learning, tout en contribuant à une meilleure compréhension des comportements énergétiques dans le cadre du projet.

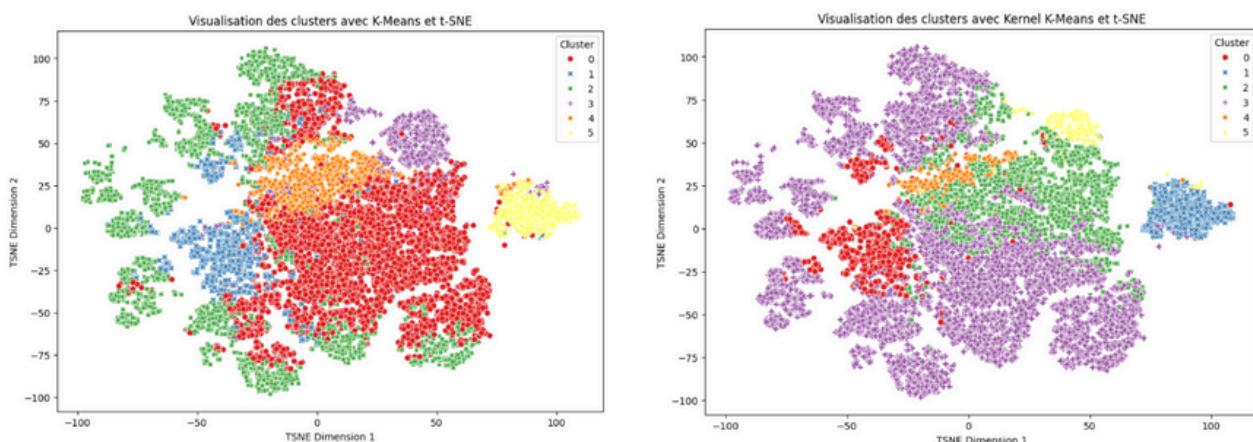


Figure 10 : Visualisation du clustering des variables comportementales en 6 clusters

Résultats de l'analyse

Dans cette étude, nous avons identifié plusieurs comportements clés influençant la consommation énergétique résidentielle. Cette section met en évidence les principaux résultats en relation avec les données et les tendances pertinentes.

Utilisation de la piscine

L'analyse a révélé que l'utilisation de la piscine est le facteur le plus énergivore dans la consommation énergétique résidentielle. Même dans les clusters où l'utilisation de la piscine est moins fréquente, cette variable reste fortement corrélée à la consommation énergétique. Cela indique que, quelle que soit la fréquence d'utilisation des piscines, elles ont un impact significatif sur les comportements énergétiques des ménages. De plus, une utilisation élevée de la piscine est plus énergivore que l'utilisation intensive de toute autre variable analysée, soulignant le rôle crucial des piscines dans les analyses de consommation énergétique.

Cette observation est renforcée par les données de marché, qui montrent une augmentation de 101% des ventes de pompes à chaleur pour piscines entre 2019 et 2021. Aux États-Unis, le marché des piscines résidentielles est important, avec 10,7 millions de piscines, dont 10,4 millions sont résidentielles, et près de 59% sont enterrées. En effet, les États-Unis sont le pays qui compte le plus de piscines privées au monde. Les ménages aisés, dont beaucoup sont situés en Floride, ont tendance à utiliser leur piscine pendant près de la moitié de l'année, ce qui entraîne une consommation énergétique élevée. Ces ménages ont souvent des maisons plus grandes, ce qui peut expliquer leur utilisation plus élevée des piscines et de l'éclairage. Par ailleurs, ils peuvent être moins sensibles à l'impact énergétique de ces équipements.

Les appareils électroménagers

Les familles avec enfants consomment beaucoup d'énergie en raison de leur utilisation intensive des appareils électroménagers et de cuisine. Ces familles utilisent souvent la machine à laver et le sèche-linge presque deux fois par jour pour chaque appareil. En revanche, d'autres types de ménages ont tendance à avoir des habitudes d'utilisation plus standard pour ces appareils. Cette utilisation accrue des appareils ménagers contribue de manière significative à leur consommation énergétique globale, ce qui illustre comment la structure familiale peut influencer les comportements énergétiques.

Eclairage

Une distinction notable concerne l'utilisation des lumières. L'analyse montre que la consommation d'énergie associée à l'éclairage dépend moins du nombre d'heures d'utilisation des lumières que du type d'ampoules employées. Environ 37 % des ménages utilisent peu ou pas d'ampoules LED pendant la journée. De plus, environ 14 % des personnes qui laissent la lumière allumée toute la nuit n'utilisent pas d'ampoules LED, ce qui est significatif compte tenu de la forte consommation d'électricité des ampoules à incandescence et halogènes.

Télévision

Bien que l'impact de la télévision sur la consommation d'énergie puisse être relativement faible lorsqu'elle est utilisée avec modération, son effet cumulatif peut devenir important lorsque son utilisation est généralisée. Cela est particulièrement évident chez les retraités, qui sont les plus grands consommateurs de télévision, avec une utilisation supérieure à 10 heures par jour. Ce comportement reflète leur tendance à passer une bonne partie de la journée devant l'écran. Les familles avec enfants contribuent également à une forte consommation d'énergie par l'utilisation de la télévision. Elles possèdent souvent plusieurs téléviseurs, dont l'un fréquemment utilisé pour les jeux vidéo des enfants, ce qui augmente encore la consommation d'énergie.

Utilisation des appareils de cuisine

Bien que les appareils électroménagers ne soient pas spécifiquement abordés dans cette étude, ils jouent néanmoins un rôle dans la consommation énergétique des ménages. L'impact relativement faible de la cuisson sur la consommation globale d'électricité des ménages américains est en grande partie dû à leurs habitudes culinaires. En moyenne, les Américains passent environ 1 heure à table, contre plus de 2 heures pour les Français. Comme indiqué, « les Américains ont de très belles cuisines, mais ils ne cuisinent pas ». La préférence pour la malbouffe et la restauration rapide, en particulier dans les familles à faibles revenus, limite l'utilisation des appareils électroménagers, car les produits frais sont souvent plus chers et moins accessibles. Cependant, les familles avec enfants utilisent généralement les appareils de cuisine plus fréquemment, ce qui augmente leur consommation globale d'énergie dans ce domaine.

Les appareils intelligents

Bien que les appareils intelligents n'aient pas été largement utilisés dans cette étude en raison de leur faible consommation individuelle et de leur faible utilisation dans les ménages américains, ils sont présents dans l'ensemble de données d'origine, notamment pour contrôler la température, la télévision, l'éclairage et la sécurité. 62% des ménages étudiés ne disposent pas d'appareils connectés, ce qui limite considérablement leur capacité à surveiller et optimiser leur consommation énergétique. De plus, bien que 26% de l'échantillon disposent d'un compteur connecté, seuls 8% vérifient leur consommation à intervalles réguliers. Ces chiffres relativement faibles mettent en évidence un potentiel sous-exploité.

Recommendations

Sur la base des résultats de cette étude, nous avons décidé de formuler des recommandations à la fois pour les particuliers et pour les autorités. Ces recommandations visent à optimiser la consommation énergétique en tenant compte des comportements identifiés comme les plus influents.

Pour les particuliers

Pour les particuliers, une approche multidimensionnelle peut améliorer considérablement la consommation d'énergie et contribuer à un mode de vie plus sain.

La promotion des activités physiques en plein air réduit la dépendance aux appareils énergivores tout en ayant un impact positif sur le bien-être général. Limiter le temps passé devant un écran et explorer d'autres activités de loisirs peuvent encore réduire la consommation d'énergie.

L'intégration de technologies intelligentes, telles que des thermostats programmables et des commandes d'éclairage automatisées, permet une gestion plus efficace de la consommation d'énergie en garantissant que les appareils ne fonctionnent que lorsque cela est nécessaire. De plus, l'adoption de pratiques écoénergétiques telles que l'utilisation d'ampoules LED, qui consomment jusqu'à 10 fois moins d'énergie que les ampoules à incandescence et 6 à 8 fois moins que les ampoules halogènes, peut réduire considérablement le gaspillage. L'éclairage prolongé, y compris les lumières extérieures laissées allumées la nuit, représente 10 à 15 % de la consommation électrique résidentielle moyenne, donc le passage aux ampoules LED et la mise en œuvre de commandes d'éclairage automatisées peuvent entraîner des économies importantes.

Pour les ménages disposant d'une piscine, des stratégies telles que l'optimisation des systèmes de chauffage ou l'utilisation de couvertures thermiques peuvent réduire considérablement la consommation d'énergie et l'évaporation de l'eau. Ces mesures peuvent également contribuer à réduire la consommation globale d'eau. En cuisine, l'adoption de pratiques de cuisson plus économies en énergie, comme l'utilisation de fours à convection ou de plaques de cuisson à induction, et l'optimisation de leur efficacité, peuvent encore réduire la consommation d'énergie.

Les appareils électroménagers jouent également un rôle important dans la consommation globale d'énergie. Les appareils modernes à haut rendement, certifiés par des labels comme Energy Star, peuvent réduire la consommation d'énergie jusqu'à 15 % par rapport aux modèles plus anciens. Par conséquent, investir dans ces appareils et utiliser leurs modes d'économie d'énergie, comme le mode éco sur les lave-vaisselle et les machines à laver, peut contribuer à des économies substantielles à long terme.

En procédant à ces ajustements, les individus peuvent jouer un rôle crucial dans la réduction de la consommation d'énergie des ménages tout en bénéficiant d'un meilleur bien-être.

Pour les pouvoirs publics

Pour les autorités, favoriser l'efficacité énergétique et la durabilité implique une approche à multiples facettes.

L'une des stratégies clés consiste à se concentrer sur les infrastructures publiques. Développer des piscines publiques, plutôt que d'encourager les piscines privées, peut réduire à la fois la consommation d'eau et d'énergie associée aux piscines individuelles des ménages. De plus, augmenter le prix d'achat des piscines pourrait potentiellement réduire le nombre de nouvelles installations de piscines, ce qui entraînerait de nouvelles réductions de la consommation d'énergie et d'eau, à condition que cette approche soit politiquement acceptable.

Améliorer l'isolation des maisons en améliorant l'efficacité thermique des murs, des fenêtres et des toits peut réduire considérablement les besoins de chauffage et de climatisation, ce qui entraîne d'importantes économies d'énergie.

La promotion des énergies renouvelables est un autre domaine essentiel. Investir dans des panneaux solaires pour les maisons peut fournir une source d'énergie durable et réduire la dépendance aux ressources non renouvelables. Soutenir les normes de construction économes en énergie pour les nouvelles constructions et les rénovations, comme la promotion des bâtiments passifs ou à faible consommation d'énergie, peut également contribuer à réduire la consommation d'énergie et la durabilité à long terme.

Encourager l'adoption d'appareils économes en énergie par le biais d'incitations telles que des rabais ou des subventions peut favoriser une utilisation plus large des technologies qui réduisent considérablement la consommation d'énergie. La sensibilisation et l'éducation du public jouent également un rôle crucial. La mise en œuvre de campagnes et d'ateliers éducatifs sur les économies d'énergie et les avantages des technologies intelligentes peut inciter les individus à adopter des pratiques d'économie d'énergie. La promotion d'alternatives à la télévision, telles que les activités sociales ou physiques, pourrait non seulement réduire la consommation d'énergie, mais aussi être bénéfique pour la santé et le bien-être. Encourager ces comportements peut ainsi contribuer aux économies d'énergie tout en améliorant la qualité de vie des retraités. Le développement d'infrastructures de recharge pour véhicules électriques et l'investissement dans l'expansion des réseaux de transports publics peuvent réduire la consommation d'énergie liée aux véhicules personnels.

En intégrant ces stratégies, les autorités peuvent créer un environnement favorable qui favorise les efforts individuels et collectifs en faveur d'une plus grande efficacité énergétique et d'une plus grande durabilité, ce qui profitera en fin de compte à la fois à l'économie et à la planète.

Conclusion

Cette étude a réalisé une analyse approfondie de la consommation énergétique résidentielle en utilisant des méthodes de prétraitement, de sélection de caractéristiques et de clustering pour comprendre l'influence des comportements des ménages et identifier des groupes distincts de consommation énergétique.

Nous avons commencé par un prétraitement rigoureux des données, réduisant le nombre d'échantillons de 18 496 à 17 312 et les variables de 799 à 208. Quatre méthodes de sélection des caractéristiques ont permis de réduire le nombre de variables pertinentes à 17, dont 9 étaient comportementales.

L'application des algorithmes de clustering, tels que K-Means et Kernel K-Means, a révélé des variations significatives dans les comportements énergétiques des ménages. Les analyses ont permis d'identifier plusieurs comportements clés en fonction de leur impact énergétique. Parmi ces comportements, l'utilisation des piscines se distingue comme le facteur le plus énergivore. Les appareils électroménagers ont également un impact significatif, tandis que l'éclairage, en particulier lorsqu'il utilise des ampoules non-LED, contribue de manière notable à la consommation d'énergie. L'utilisation des appareils de cuisine et le visionnage de la télévision, bien que moins énergivores, contribuent également à la consommation globale.

Cette répartition souligne l'importance de mettre en place des stratégies adaptées pour gérer la consommation d'énergie de manière efficace. Les équipements de loisirs, en particulier les piscines, nécessitent des interventions ciblées pour améliorer leur efficacité énergétique. Les appareils électroménagers doivent être optimisés pour minimiser leur impact, tandis que des mesures telles que l'adoption d'ampoules LED et l'instauration de systèmes de contrôle automatisés peuvent offrir des économies notables en matière d'éclairage.

À l'avenir, les technologies numériques, telles que le streaming vidéo et le cloud computing, qui sont très gourmands en ressources, pourraient accroître significativement la consommation énergétique résidentielle. En 2023, le secteur informatique représente environ 10 % de la consommation mondiale d'électricité, en hausse par rapport aux 7 % en 2016, et cette proportion pourrait atteindre 20 % d'ici 2030. Le streaming vidéo, qui constitue environ 80 % du trafic web mondial, contribue largement à cette tendance. L'impact croissant de ces technologies sur la consommation d'électricité nécessite une attention particulière pour élaborer des stratégies efficaces visant à atténuer leur effet sur la demande énergétique.

D'un autre côté, il est pertinent de noter que la prédition de la consommation électrique pourrait bénéficier de l'intégration des technologies numériques émergentes. Ces technologies jouent un rôle de plus en plus crucial dans la gestion et la réduction de la consommation d'énergie résidentielle. L'Internet des objets (IoT) et les objets connectés permettent une collecte de données en temps réel sur la consommation énergétique, offrant des opportunités pour des analyses plus fines et des prévisions plus précises. Les systèmes de gestion de l'énergie domestique (HEMS) peuvent surveiller et optimiser l'utilisation de l'énergie, tandis que les réseaux intelligents (smart grids) permettent une distribution plus équilibrée et efficace de l'électricité. L'Internet du comportement (IoB) pourrait également jouer un rôle crucial en permettant une compréhension approfondie des habitudes et des comportements des utilisateurs. Malheureusement, pour cette recherche, nous n'avons pas pu accéder à des données aussi détaillées, ce qui limite notre capacité à tirer pleinement parti de ces technologies émergentes. Néanmoins, elles pourraient transformer la manière dont nous anticipons les besoins énergétiques et optimisons la gestion de la consommation, en offrant des outils puissants pour améliorer l'efficacité énergétique et réduire l'empreinte carbone des ménages à l'avenir.

Aujourd'hui, alors que nous atteignons le Jour du Dépassement de cette année 2024, cette étude souligne l'impact environnemental crucial de nos comportements énergétiques. Le Jour du Dépassement marque la date à laquelle l'humanité a consommé toutes les ressources que la Terre peut renouveler en une année. Adopter des stratégies ciblées pour optimiser la consommation d'énergie au niveau des ménages est essentiel pour favoriser une prise de conscience accrue et des pratiques plus durables, contribuant ainsi à réduire notre empreinte écologique et à préserver les ressources de la planète.

Présentation des résultats

Au cours du stage

Tout au long de mon stage, j'ai régulièrement présenté les résultats de mes analyses lors de réunions avec mon tuteur, tant en présentiel qu'en ligne via Zoom. Ces réunions avaient pour objectif de partager les progrès réalisés, de discuter des résultats obtenus, et de recueillir des retours pour affiner les approches méthodologiques. Pour chaque réunion, j'élaborais un nouveau PowerPoint afin de structurer mes exposés et de faciliter la communication des données, des méthodes utilisées, et des conclusions tirées.

Durant la phase de l'état de l'art, les réunions se tenaient toutes les deux semaines, ce qui m'a permis de m'assurer que mes recherches étaient alignées avec les attentes et les besoins du projet. Ensuite, pendant la phase d'expérimentation, les présentations se faisaient chaque semaine. Chaque réunion était l'occasion de réévaluer les hypothèses initiales, de confronter les résultats intermédiaires aux attentes, et d'explorer de nouvelles pistes pour améliorer la qualité des analyses. Les discussions portaient sur des points critiques comme la validité des résultats et les améliorations méthodologiques possibles. Cette démarche itérative a été essentielle pour peaufiner les analyses, améliorer la précision des résultats, et assurer une rigueur méthodologique tout au long du projet.

Rédaction d'un Article Scientifique

À la fin de mon stage, j'ai entamé la rédaction d'un article scientifique pour présenter de manière structurée les méthodes, les analyses, et les résultats du projet. Bien que l'article ne soit pas encore achevé, il est conçu pour fournir un compte rendu détaillé de la méthodologie employée et des résultats obtenus.

L'article commence par une introduction qui définit le contexte de l'étude et les objectifs visés. La section méthodologie décrit en détail les techniques de clustering utilisées, notamment K-means, Kernel K-means et DBSCAN, ainsi que les critères d'évaluation pour déterminer le nombre optimal de clusters. Ensuite, l'article présente les résultats des différentes analyses effectuées, mettant en lumière les principaux facteurs influençant la consommation énergétique des ménages et les variations observées entre les différents clusters.

Une attention particulière est portée sur l'interprétation des résultats, où sont discutées les implications des comportements observés, tels que l'impact de l'utilisation des piscines ou des appareils électroménagers sur la consommation d'énergie. La conclusion de l'article synthétise les découvertes majeures, propose des recommandations pour améliorer l'efficacité énergétique, et suggère des pistes pour des recherches futures.

La rédaction de cet article m'a permis de formaliser les résultats du projet de manière rigoureuse, de documenter les méthodes utilisées, et de préparer le travail pour une publication potentielle. Ce processus a également été une opportunité d'approfondir mes compétences en rédaction scientifique et en communication des résultats de recherche.

Les problèmes rencontrés, les solutions envisagées

5

Les problèmes rencontrés

1

Manque de données

La première étape d'un projet de machine learning commence souvent par la collecte des données. Au début du stage, nous avons donc sollicité des données auprès d'un contact chez EDF, M. Christian OBRECHT, avec qui le laboratoire avait déjà collaboré pour des projets antérieurs. Bien que ce dernier ait confirmé la disponibilité des informations demandées, il nous a informés qu'il ne pouvait pas les transmettre en raison de préoccupations liées à la sécurité et à la confidentialité.

2

Données non temporelles

Les données temporelles sont cruciales pour identifier des tendances, voir l'effet saisonnier, modéliser des comportements futurs et réaliser des prévisions fiables. Sans ces informations, il est impossible de construire des modèles prédictifs robustes, ce qui limite la portée et la précision de nos analyses. Les données disponibles ne comportaient pas d'informations sur les évolutions dans le temps, ce qui a empêché notre capacité à effectuer des analyses prédictives.

3

Diversité et précision des données

Les comportements énergétiques des ménages sont complexes et influencés par de nombreux facteurs, mais les données collectées ne couvraient qu'un éventail restreint de ces variables. En particulier, le nombre de ménages utilisant des objets connectés dans notre échantillon était faible, ce qui a limité notre capacité à évaluer l'impact potentiel de ces dispositifs sur la réduction de la consommation d'énergie.

Les solutions envisagées

Pour surmonter les défis rencontrés durant mon stage, il est vite apparu que la clé résidait dans l'identification d'un dataset approprié. Face au manque de données, aux limitations liées à l'absence de données temporelles, et au manque de diversité et de précision des informations disponibles, nous avons cherché des solutions pour obtenir un ensemble de données plus complet. Comme mentionné précédemment, nous avons tenté de collaborer avec un contact chez EDF, M. Christian Obrecht, mais malgré son accord sur la disponibilité des données, des préoccupations liées à la sécurité, à la confidentialité et à des contraintes juridiques ont empêché leur transmission.

Après avoir exploré plusieurs options sans succès, nous avons finalement opté pour le dataset RECS (Residential Energy Consumption Survey). Ce dernier s'est avéré être la solution la plus adéquate, offrant une couverture suffisamment complète pour répondre aux besoins de notre analyse, malgré les limitations initiales.

Bilan du stage

6

Les réalisations

Pendant mon stage, j'ai eu l'opportunité de rédiger un article scientifique basé sur les analyses réalisées. Bien que je sois conscient que ma contribution n'est pas particulièrement révolutionnaire ni significative dans le cadre global de la recherche, la possibilité que cet article soit publié reste une source de satisfaction.

Contribuer, même modestement, à un projet de recherche est une expérience précieuse. L'écriture de cet article m'a permis de me familiariser avec le processus de recherche académique et de comprendre l'importance de chaque petite étape dans l'avancement des connaissances. Même si mon travail n'est qu'une petite pierre à l'édifice, savoir que cela pourrait être utile à d'autres chercheurs est gratifiant.

Les acquis du stage

Mon stage m'a offert une immersion précieuse dans le monde de la recherche, où j'ai pu acquérir une compréhension approfondie de la méthodologie et des pratiques rigoureuses qui la caractérisent. L'une des leçons majeures que j'ai tirées de cette expérience est l'importance cruciale du prétraitement des données. Cette étape, souvent sous-estimée, s'est révélée être le fondement sur lequel repose la validité de toute analyse. Une erreur, même minime, dans le traitement initial des données peut entraîner la remise en cause de l'ensemble du processus, obligeant à repartir de zéro. Cette prise de conscience m'a appris à aborder chaque étape avec une attention accrue, garantissant ainsi l'intégrité des résultats obtenus.

Par ailleurs, j'ai eu l'opportunité d'explorer et de mettre en pratique divers algorithmes de clustering. Bien que tous n'aient pas été explicitement mentionnés dans ce rapport, cette exploration a élargi mon spectre de compétences techniques et m'a permis de mieux comprendre les spécificités et les applications de ces méthodes dans l'analyse des comportements énergétiques.

Enfin, ce stage m'a permis de m'immerger dans l'état actuel des recherches sur les méthodes de prédiction et d'influence de la consommation électrique dans les ménages. Cette vue d'ensemble m'a non seulement permis de saisir les enjeux actuels du secteur, mais aussi d'identifier les directions futures que pourrait prendre la recherche dans ce domaine en constante évolution.

En conclusion, cette expérience a été formatrice à plusieurs niveaux. Elle m'a non seulement permis de développer des compétences techniques spécifiques, mais elle m'a également inculqué une rigueur méthodologique essentielle pour toute démarche de recherche scientifique.

Les perspectives

Ce stage a constitué une excellente introduction au secteur de l'énergie, un domaine qui s'avère à la fois crucial et en constante évolution. À travers l'analyse des données énergétiques et l'exploration des comportements de consommation, j'ai pu développer des compétences clés en méthodologie de recherche et en analyse de données, qui me seront précieuses pour la suite de mon parcours professionnel.

L'expérience acquise m'a non seulement permis de comprendre les enjeux énergétiques actuels, mais elle a également éveillé en moi un intérêt pour ce secteur. Poursuivre dans cette voie pourrait être particulièrement enrichissant, notamment en intégrant une entreprise qui dispose de données plus précises et complètes. Cela me permettrait de contribuer à des projets plus ambitieux, avec des analyses plus poussées et des résultats ayant un impact significatif.

L'idée de travailler directement avec des données de terrain, au sein d'une entreprise engagée dans la transition énergétique, est une perspective intéressante. Une telle opportunité me permettrait d'appliquer et d'étendre les connaissances acquises, tout en participant activement à la recherche de solutions pour une consommation énergétique plus efficace et durable.

En somme, ce stage m'a ouvert des perspectives intéressantes pour ma carrière, en renforçant mon intérêt pour le secteur de l'énergie et en me donnant envie d'explorer davantage ce domaine, que ce soit par le biais d'une entreprise spécialisée ou dans le cadre de projets de recherche plus approfondis.

Conclusion

Mon ressenti à l'issue de ce stage est plutôt mitigé. D'un côté, j'ai eu l'opportunité de m'immerger dans la méthodologie de la recherche, ce qui a été l'un des aspects les plus enrichissants de mon expérience. Réaliser un état de l'art, bien que fastidieux, m'a permis d'explorer une variété de techniques actuellement utilisées dans le monde, et j'ai trouvé cela particulièrement intéressant. Cela m'a offert une vision plus globale des avancées dans le domaine, même si cela demandait beaucoup de lecture et d'efforts pour comprendre les subtilités des techniques étudiées. D'un autre côté, j'ai réalisé que je devais adopter une approche plus rigoureuse dans mon travail. Souvent, j'allais trop vite, ce qui m'a montré l'importance de la minutie et de la patience dans ce type de projet. Cette expérience m'a rappelé qu'une précipitation excessive peut nuire à la qualité du travail, et que chaque étape mérite une attention particulière pour éviter des erreurs coûteuses par la suite.

Techniquement, ce que j'ai réalisé durant ce stage ne m'a pas vraiment apporté de nouveautés. Les tâches qui m'ont été confiées étaient similaires à ce que j'avais déjà rencontré et réalisé dans des projets antérieurs, ce qui a quelque peu limité mon apprentissage sur ce plan. Cependant, cela m'a permis de consolider certaines compétences, même si j'aurais souhaité être davantage confronté à des défis techniques inédits.

Sur un autre plan, ce stage a été une bonne introduction au secteur de l'énergie. Même si je ne sais pas encore si je travaillerai dans ce secteur à l'avenir, j'ai acquis une base solide et une compréhension générale du domaine. Cette expérience m'a permis d'aborder les questions liées à l'énergie avec plus de confiance et d'avoir une perspective plus informée sur les défis et les opportunités présents dans ce secteur.

Références bibliographiques

1. Page d'accueil de l'Université Claude Bernard Lyon 1, <https://www.univ-lyon1.fr/>
2. Page d'accueil du LIRIS, <https://liris.cnrs.fr>
3. Page d'accueil de Université Lyon, <https://www.universite-lyon.fr/version-francaise/>
4. Blent. Arbres de décision en Machine Learning : tout comprendre. <https://blent.ai/blog/a/arbres-de-decision-en-machine-learning>
5. U.S. Energy Information Administration. 2020 Residential Energy Consumption Survey. <https://www.eia.gov/consumption/residential/data/2020/>
6. Statista. Évolution du nombre d'habitants sur Terre entre 1950 et 2024 et projections jusqu'en 2100. <https://fr.statista.com/statistiques/564933/population-mondiale-jusqu-en-2080/>
7. D. Chareyron, H. Horsin-Molinaro, B. Multon. Concepts et chiffres de l'énergie : la consommation de l'électricité domestique en France. Culture Sciences Physiques, ENS Lyon. <https://culturesciencesphysique.ens-lyon.fr/ressource/chiffres-energie-electricite-domestique.xml>
8. H. M. Sani, S. O. Tehrani, B. Behkamal, H. Amintossi : Extracting Effective Features for Descriptive Analysis of Household Energy Consumption Using Smart Home Data. 2019.
9. M. Heinrich, M. Ruellan, L. Oukhellou, A. Samé, J.-P Lévy : From energy behaviors to lifestyles: Contribution of behavioral archetypes to the description of energy consumption patterns in the residential sector. 2022.
10. T. F. Sanquist, H. Orr, B. Shui, A. C. Bittner : Lifestyle factors in U.S. residential electricity consumption. 2012.
11. Aquark. Analyse du marché : Quelle est la taille du marché mondial de la piscine ?. <https://www.aquark.com/fr/analyse-du-marche-mondial-de-la-piscine/>
12. U.S. Energy Information Administration. A call to action on efficient and smart appliances. <https://www.iea.org/articles/a-call-to-action-on-efficient-and-smart-appliances>
13. Engie. LED, halogène, connectée... quelle est la consommation d'une ampoule ?. <https://particuliers.engie.fr/economies-energie/conseils-economies-energie/conseils-calcul-consommation/consommation-ampoule.html>
14. Élisabeth Chesnais. Ampoules LED De belles économies. UFC Que Choisir. <https://www.quechoisir.org/guide-d-achat-ampoules-basse-consommation-led-n11547/>
15. M. Labiad. Methodology for construction of adaptive models for the simulation of energy consumption in buildings. 2022.
16. H. Elayan, M. Aloqaily, F. Karray, M. Guizani. Decentralized IoB for Influencing IoT-based Systems Behavior. 2024.
17. H. Elayan, M. Aloqaily, F. Karray, M. Guizani. Internet of Behavior and Explainable AI Systems for Influencing IoT Behavior. 2023.
18. J. Wang, X. Chen, F. Zhang, F. Chen, Y. Xin. Building Load Forecasting Using Deep Neural Network with Efficient Feature Fusion. 2021.
19. J. W. Chan, C. K. Yeo. Electrical Power Consumption Forecasting with Transformers. 2022.
20. D. Hadjout, J.F. Torres, A. Troncoso, A. Sebaa, F. Martínez-Alvarez. Electricity consumption forecasting based on ensemble deep learning. 2021.
21. D. Hadjout, A. Sebaa, J. F. Torres, F. Martínez-Álvarez. Electricity consumption forecasting with outliers handling based on clustering and deep learning with application to the Algerian market. 2023.

- 22.** D. Syed, H. Abu-Rub, A. Ghrayeb, S. S. Refaat. Household-Level Energy Forecasting in Smart Buildings Using a Novel Hybrid Deep Learning Model, 2021.
- 23.** M. Alhussein , K. Aurangzeb, S. I. Haider. Hybrid CNN-LSTM Model for Short-Term Individual Household Load Forecasting. 2020.
- 24.** X. Li, Y. Zhong, W. Shang, X. Zhang, B. Shan, X. Wang. Total electricity consumption forecasting based on Transformer time series models. 2022.
- 25.** Y. Liu, D. Zhang, H. B. Gooi. Optimization Strategy Based on Deep Reinforcement Learning for Home Energy Management. 2020.
- 26.** S.-H. Kim, C. Lee, C.-H. Youn. An Accelerated Edge Cloud System for Energy Data Stream Processing Based on Adaptive Incremental Deep Learning Scheme. 2020.
- 27.** EuropUSA. Etats-Unis : les habitudes alimentaires. <https://www.europusa.com/vivre-aux-etats-unis/vie-personnelle-aux-usa/vie-quotidienne/etats-unis-les-habitudes-alimentaires/>
- 28.** SudOuest. Les Français passent deux fois plus de temps à table que les Américains. <https://www.sudouest.fr/premium/art-de-vivre/les-francais-passent-deux-fois-plus-de-temps-a-table-que-les-americains-3136883.php>
- 29.** Christina Gierse. Les Français et la nourriture aux Etats-Unis : un sujet « touchy ». Studyrama. <https://www.studyrama.com/pro/destination/les-francais-et-la-nourriture-aux-etats-unis-un-sujet-touchy-21891.html>
- 30.** Guide Piscine. https://www.guide-piscine.fr/pro/marche-de-la-piscine/piscine-eco-responsable/l-impact-d-une-couverture-sur-la-consommation-en-eau-et-en-electricite-d-une-piscine-5620_A
- 31.** Christophe Magdelaine. Streaming : quelles émissions de CO2 et consommation d'énergie ?, 07/03/2024. <https://www.notre-planete.info/actualites/247-streaming-Internet-electricite-CO2>
- 32.** Institut National d'Etudes Démographiques, 2024 : les Nations unies publient de nouvelles projections de population mondiale, <https://www.ined.fr/fr/tout-savoir-population/memos-demo/focus/2024-les-nations-unies-publient-de-nouvelles-projections-de-population-mondiale>
- 33.** Nations Unies, Population, <https://www.un.org/fr/global-issues/population>
- 34.** Pablo-Romero, M. del P., Pozo-Barajas, R., Yñiguez, R., Global changes in residential energy consumption, Energy Policy (2016), <https://doi.org/10.1016/j.enpol.2016.10.032>
- 35.** Taconet, C., Objets connectés : 50 milliards d'émetteurs de CO2 ?, Télécom SudParis, <https://www.telecom-sudparis.eu/actualite/objets-connectes-50-milliards-demetteurs-de-co2/>
- 36.** Sun, Jiayi, Gan, Wensheng, Chao, Han-Chieh, Philip, S Yu, Ding, Weiping. Internet of behaviors: A survey. IEEE Internet of Things, vol. 10, no. 13, pp. 11117 – 11134, 2023.
- 37.** Zhao, Q., Li, G., Cai, J., Zhou, M., and Feng, L. A Tutorial on Internet of Behaviors: Concept, Architecture, Technology, Applications, and Challenges. IEEE Communications Surveys & Tutorials, vol. 25, no. 2, pp. 1227 –1260, 2023.
- 38.** Ziani, L., Khanouche, M. E., Belaid A.: Internet of behaviors: A literature reviewof an emerging technology. In 1st Int. Conf. on Big Data, IoT, Web Intelligence andApplications, pp. 42-47, Sidi Bel Abbes, Algeria, 2022.
- 39.** Javaid, M., Haleem, A., Singh, R.P., Khan, S., Suman, R.: An extensive studyon internet of behavior (iob) enabled healthcare-systems: Features, facilitators, andchallenges. BenchCouncil Transactions on Benchmarks, Standards and Evaluations,100085, 2023.

- 40.** Mezair, Tinhinane, Djenouri, Youcef, Belhadi, Asma, Srivastava, Gautam, Lin, Jerry Chun-Wei. Towards an Advanced Deep Learning for the Internet of Behaviors: Application to Connected Vehicles. ACM Transactions on Sensor Networks, vol. 19, no. 2, pp. 1-18, 2022.
- 41.** Song, Qun, Tan, Rui, Wang, Jianping. Towards Efficient Personalized Driver Behavior Modeling with Machine Unlearning. In Proceedings of Cyber-Physical Systems and Internet of Things, pp. 31-36, 2023.
- 42.** Embarak, O.H.: Internet of behaviour (IoB)-based AI models for personalized smart education systems. Procedia Computer Science 203, 103-110, 2022.
- 43.** Embarak, Ossama. An adaptive paradigm for smart education systems in smart cities using the internet of behaviour (IoB) and explainable artificial intelligence (XAI). 8th Int. Conference on Information Technology Trends, pp. 74-79, 2022.

Annexe

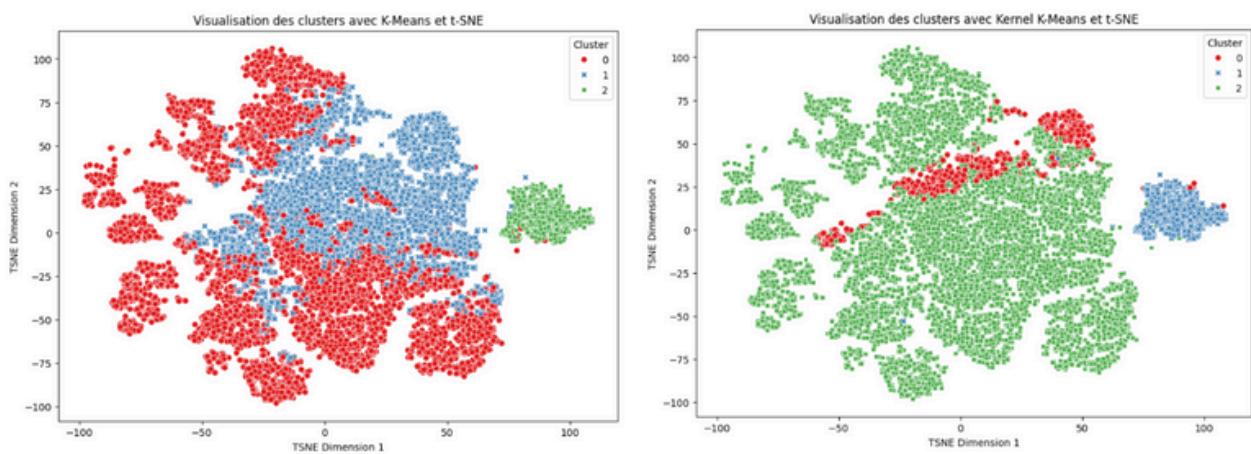


Figure 1 : Visualisation du clustering des variables comportementales en 3 clusters

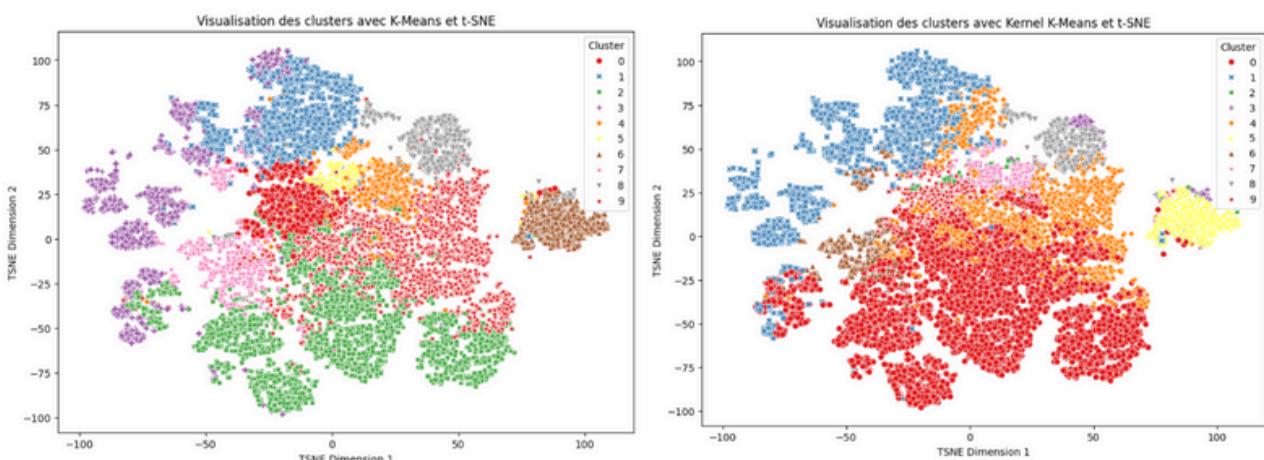


Figure 2 : Visualisation du clustering des variables comportementales en 10 clusters

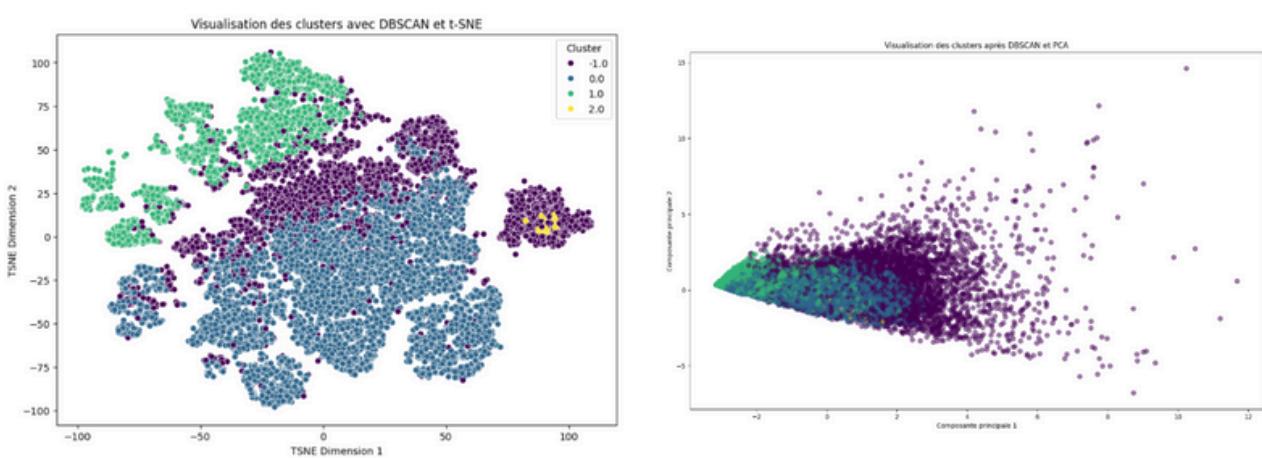


Figure 3 : Visualisation du clustering des variables comportementales avec DBSCAN