UQAC
Université du Québec
à Chicoutimi

PREDICTION POUR L'OPTIMISATION DE LA CONSOMMATION ÉLECTRIQUE RESIDENTIELLE À L'AIDE DE L'APPRENTISSAGE AUTOMATIQUE SUR DES DONNEES ELECTRIQUES ET METEOROLOGIQUES

PAR ESVADO PRUDENCIO M. HOUNTON

Mémoire présenté à l'Université du Québec à Chicoutimi en vue de l'obtention du grade de Maîtrise ès sciences (M. Sc.) en informatique

QUEBEC, CANADA

RÉSUMÉ

L’optimisation de la consommation électrique résidentielle représente un défi majeur à l’ère de la transition énergétique. Pour y répondre, la capacité de prédire cette consommation s’impose comme un outil presque indispensable, car elle permet d'anticiper et d'assurer la gestion de l’énergie d’une manière plus efficace, durable et économique. C’est dans ce cadre que le présent projet de maîtrise explore les solutions offertes par l’apprentissage automatique pour la prédiction à partir de données de consommation et météorologiques.

La solution développée est axée sur l’utilisation de divers modèles d’apprentissage automatique et d’apprentissage profond. Ces modèles ont été entraînés sur des données de consommation réelle issues des données publiques d’Hydro-Québec. L’étude aborde plusieurs phases de prédiction, allant de l’ingénierie des caractéristiques à l’optimisation des hyperparamètres, en passant par l’explicabilité des modèles à l’aide de la technique shapley additive explanations (SHAP). Elle explore également les modèles préentraînés utilisés pour la prédiction de séries temporelles, tels que TimeGPT et TimesFM.

Au cours de ce travail, une évaluation comparative des performances de différents modèles (régression linéaire, XGBoost, CatBoost, RNN, LSTM, TimeGPT, TimesFM) a été effectuée, en utilisant des métriques telles que la racine de l’erreur quadratique moyenne (RMSE), l’erreur absolue moyenne (MAE), l’erreur quadratique moyenne (MSE), l’erreur absolue moyenne en pourcentage (MAPE), et le coefficient de détermination (R^2). Les résultats ont montré que les modèles d'ensemble peuvent être performants pour la prédiction de la consommation électrique. Mieux encore, les modèles préentraînés ont démontré une capacité à produire des prédictions fiables sans nécessiter d'entraînement local.

Ce projet de recherche vise ainsi à démontrer comment l’intelligence artificielle peut contribuer à une meilleure anticipation de la consommation et à l’optimisation énergétique. Il propose également une méthodologie applicable à d’autres contextes de prédiction énergétique, avec un intérêt particulier pour la performance, l’explicabilité et l’applicabilité des modèles.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

IA	Intelligence Artificielle
DL	Deep Learning
ML	Machine Learning
ACF	AutoCorrelation Function
PACF	Partial AutoCorrelation Function
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
ARIMA	AutoRegressive Integrated Moving Average
GBRT	Gradient Boosting Regression Trees
SVR	Support Vector Regression
XGBoost	Extreme Gradient Boosting
CatBoost	Categorical Boosting
RF	Random Forest
LGBM	Light Gradient Boosting Machine
HGB	Histogram-Based Gradient Boosting
TimeGPT	Time Generative Pre-trained Transformer
TimesFM	Time Series Foundation Model
SHAP	SHapley Additive exPlanations
MAE	Mean Absolute Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
R^2	Coefficient de détermination
MAPE	Mean Absolute Percentage Error
PCA	Principal Component Analysis
ReLU	Rectified Linear Unit
ELU	Exponential Linear Unit
API	Application Programming Interface
kWh	Kilowattheure
AMS	Systèmes de mesure avancés

CPU	Unité centrale de traitement
GES	Gaz à effet de serre

DÉDICACE

À toute ma famille, et à ma petite famille

Votre amour, vos sacrifices et votre foi en moi ont été le socle de mon cheminement.

Merci de m'avoir appris la persévérance, l'humilité et la valeur du travail bien fait.

Ce mémoire est le fruit d'un long parcours, parfois semé d'embûches, mais toujours guidé par votre présence et vos encouragements. Avec toute ma reconnaissance.

REMERCIEMENTS

Je tiens à remercier sincèrement toutes les personnes qui m'ont soutenu tout au long de la réalisation de ma maîtrise et durant la rédaction de ce mémoire.

Je remercie particulièrement monsieur Kevin Bouchard, mon directeur de mémoire, pour sa disponibilité, ses conseils avisés, sa rigueur scientifique et son accompagnement constant. Ses orientations ont été précieuses à chaque étape de ce projet.

Je remercie chaleureusement l'ensemble du personnel enseignant et administratif du département d'informatique et de mathématique de l'Université du Québec à Chicoutimi, pour la qualité de leur enseignement, leur disponibilité et leur bienveillance, qui ont grandement enrichi mon parcours.

Enfin un grand merci à ma famille, pour ses encouragements, sa patience, son amour et sa foi inébranlable en moi. Vos sacrifices, vos prières et votre soutien, même à distance, ont été la source de ma persévérance.

Ce mémoire vous est dédié.

CHAPITRE 1

INTRODUCTION

1.1 CONTEXTE DE RECHERCHE

L'énergie est un moteur clé du développement économique et technologique. Elle contribue à la création d'emplois et à l'innovation technologique, apportant un soutien indispensable à l'atteinte des objectifs du développement durable. Cependant, sa consommation est également une grande préoccupation à l'échelle mondiale, car elle a des conséquences néfastes sur notre environnement(*Emissions Gap Report*, 2021).

En effet, la production et la consommation abusive de l'électricité engendrent une bonne quantité des émissions de gaz à effet de serre, contribuant activement au réchauffement climatique et à la dégradation de l'environnement. Selon un rapport du ministère de la transition énergétique français, en 2022, la production d'électricité reste le principal secteur émetteur de gaz à effet de serre (GES). Dans le monde, elle représente 39 % des émissions totales dues à la combustion d'énergie (*Panorama mondial des émissions de GES*, 2024).

Parallèlement, la demande croissante pour l'énergie électrique pose d'importants défis économiques, notamment en ce qui concerne la gestion des ressources, l'utilisation inefficace et les coûts supportés par les ménages et les entreprises. Cette situation est confirmée par les données indiquant que, entre 2015 et 2021, la population mondiale qui utilise l'électricité a augmenté de 87 % à 91 % (World Energy Outlook., 2023.). Pourtant, 675 millions de personnes dans le monde continuent de vivre sans électricité (Al Kez et al., 2024). Les utilisateurs ayant déjà accès à l'électricité l'utilisent de manière incontrôlée. Cela engendre

une consommation énergétique excessive impliquant des factures d'électricité très élevées. Au vu de tout ceci, on remarque que l'énergie joue un rôle primordial dans le développement économique et technologique, mais sa consommation incontrôlée représente un défi majeur sur le plan environnemental et sanitaire. L'augmentation de la demande, associée aux émissions de gaz à effet de serre, met en évidence la nécessité d'adopter des stratégies durables pour une production et une utilisation plus responsable et efficiente.

1.1.1 PLAN DE TRANSITION ÉNERGÉTIQUE

En réponse à cette crise, l'objectif sept du Développement Durable (ODD) définis par les Nations Unies dans le cadre de l'Agenda 2030, vise à assurer l'accès à tous, à une énergie fiable, durable, moderne et à un coût abordable. Pour concrétiser cette vision à travers le monde, plusieurs entités nationales et internationales ont développé des plans d'action ajustés à leurs besoins et ressources propres.

Dans le cadre de sa vision de la transition énergétique vers une économie verte à l'horizon 2030, le ministère de l'Énergie et des Ressources naturelles du Québec présente une stratégie sur l'hydrogène vert et les bioénergies, tout en adoptant une approche centrée sur les comportements des utilisateurs (*Transition énergétique - Québec*, 2022). De même, la commission européenne mise sur l'efficacité énergétique. Elle propose une solution axée sur la bonne gestion de la demande de l'énergie. Cette approche vise à consommer l'énergie de manière plus consciente et intelligente pour permettre de diminuer les factures énergétiques. Quant à l'agence internationale de l'énergie (AIE), dans sa perspective énergétique mondiales 2023, elle souligne des points clés pour relever les défis d'une transition énergétique sécurisée. Elle propose l'utilisation de l'énergies renouvelables tout en

renforçant l'efficacité énergétique (World Energy Outlook, 2023). Toutes ces mesures convergent vers une transition énergétique qui concerne tout les secteurs consommateurs de l'énergie, que ce soit l'industriel, le résidentiel, le commercial ou le transport.

1.1.2 EFFICACITÉ ÉNERGÉTIQUE RÉSIDENTIELLE

Cependant, le secteur résidentiel se distingue particulièrement parmi les grands consommateurs d'énergie pour assurer une transition efficace, et cela, pour plusieurs raisons. Tout d'abord, la croissance des nouvelles constructions et l'amélioration des conditions de vie entraînent une augmentation significative de la demande énergétique (Güneralp et al., 2017). Ensuite, les coûts de l'électricité continuent de grimper, il devient urgent d'optimiser la consommation énergétique au sein des foyers afin d'aider les ménages à alléger leurs factures (Premkumar et al., 2025). L'impact de ces facteurs pourraient être améliorés grâce à des sources d'énergie propre, constituant ainsi une alternative transitoire efficace. L'autre alternative est aussi d'assurer une bonne gestion de la consommation, pour réduire les excès d'utilisation tant qu'avec une source propre ou non.

Selon les résultats de plusieurs études, ce secteur offre de nombreuses opportunités d'optimisation énergétique, notamment grâce à l'intégration de technologies innovantes (Bibri & Krogstie, 2020). Il représente également un levier clé pour une transition énergétique majeure. Il a le potentiel de passer du statut de grand consommateur d'énergie à celui d'un modèle optimisé, plus efficace et durable à l'échelle mondiale (J.-L. Liu et al., 2019).

Compte tenu de ces constats, l'optimisation de la consommation apparaît comme une solution indispensable pour améliorer l'efficacité énergétique résidentielle et mieux maîtriser

l’usage de cette énergie. Selon (Barth et al., 2021), l'optimisation de l’efficacité énergétique résidentielle est le processus d'ajustement automatique des états des récepteurs électriques pour maintenir la consommation d'énergie dans des limites optimales. Il permet de réduire les pointes de consommation d'électricité, notamment celles causées par le chauffage électrique et l’usage d’électroménagers. Cela signifie que l’optimisation peut permettre non seulement de réduire le coût des factures, mais aussi de préserver les appareils des risques de surtension et autres dommages. Plusieurs solutions existent déjà pour y parvenir. Parmi elles, l'installation de capteurs de mouvement et de présence qui permet d'éteindre automatiquement les lumières et d'ajuster la température lorsque les pièces ne sont pas occupées. Cette avancée a ouvert la voie aux bâtiments intelligents, équipés de systèmes avancés de surveillance et de gestion de la consommation énergétique. Les systèmes de gestion de l'énergie domestique, appelés en anglais home energy management system (HEMS), quant à eux, optimisent en temps réel l’utilisation de l’énergie en fonction de la demande, contribuant ainsi à réduire le gaspillage (Minoli et al., 2017). Une autre solution largement utilisée est l’installation de système solaires. Elle permet de produire de l’énergie propre à partir du soleil, réduisant la dépendance aux réseaux électriques conventionnels (Abd El-Aziz, 2022). Ces solutions d’optimisation de la consommation énergétique résidentielle offrent des perspectives prometteuses. Toutefois, elles présentent certaines limites. D’une part, les systèmes solaires restent dépendants des conditions météorologiques variables, avec une durée de vie et une capacité de stockage limitées. D’autre part, les systèmes de gestion de l’énergie domestique (HEMS) sont confrontés à des défis tels que les coûts élevés d’installation, les problèmes de connectivité des capteurs intelligents et la complexité d’adaptation à une architecture existante. C’est dans ce contexte que la prédiction de la consommation énergétique prend tout son sens.

Elle permet d'anticiper la consommation future d'énergie en se basant sur les données de consommation passée et des modèles mathématiques (Pham et al., 2020). Les travaux de ce mémoire sont axés sur l'exploitation des données électriques résidentielles disponibles pour prédire la consommation électrique. Cette prédiction permet d'anticiper les besoins énergétiques, d'identifier les facteurs influents et est réalisée à l'aide de divers algorithmes afin de déterminer les mesures possibles d'optimisation pour une gestion plus efficace de l'énergie.

1.2 PROBLEMATIQUE

Au vu de tous ces aspects précédents, la prédiction de la consommation électrique résidentielle pourrait jouer un rôle essentiel dans le but d'optimiser l'efficacité énergétique. De plus, elle pourrait également permettre de planifier l'achat, la vente et le stockage de l'électricité pour une distribution sans pertes (Matos et al., 2024). C'est pourquoi il serait important de prédire la consommation des ménages. Les avantages de la prédiction de consommation électrique sont multiples et interviennent aussi dans les systèmes automatiques existants pour leur amélioration. Selon (Pham et al., 2020), cette prédiction facilite le développement de systèmes intelligents plus performants. Tous ces arguments amènent à constater que l'intégration de la prédiction ne se contenterait pas de compléter les solutions existantes en matière d'efficacité énergétique, en les rendant encore plus adaptatives, mais qu'elle permettrait également d'optimiser davantage l'utilisation de l'énergie.

Ainsi, la recherche menée dans le cadre de ce mémoire a pour finalité d'apporter une réponse à la problématique suivante :

Comment prédire efficacement la consommation électrique en exploitant les données de consommation et météorologiques, pour optimiser l'efficacité énergétique résidentielle ?

Pour apporter une réponse à cette problématique, notre recherche s'est concentrée sur les approches de l'apprentissage automatique, qui seront développées dans les chapitres suivants. On se concentre sur des prédictions globales et contextuelles, prenant en compte les facteurs influençant la consommation énergétique des bâtiments résidentiels, notamment les données météorologiques. Notre approche consiste à examiner les capacités techniques des modèles de prédiction, en particulier leur aptitude à analyser les données électriques domestiques et à évaluer l'impact des conditions météorologiques. Ensuite, une étude comparative sera menée afin de sélectionner les modèles les plus adaptés à notre problématique. Afin de mieux appréhender ces approches, il est important de comprendre les techniques fondamentales de l'apprentissage automatique.

L'apprentissage automatique est une branche de l'intelligence artificielle qui permet aux ordinateurs de comprendre les relations entre données sans être explicitement programmés. Selon (Amasyali & El-Gohary, 2018), l'apprentissage automatique permet d'analyser les données et apprendre à partir de celle-ci afin d'effectuer des prédictions avec la plus de précision possible. Dans le cadre de ce travail, son application favorise l'optimisation de la gestion énergétique des bâtiments grâce à sa capacité à identifier les tendances et à ajuster en temps réel les paramètres de consommation (Dinmohammadi et al., 2023).

Dans cette perspective, tirer parti de ces avancées technologiques devient indispensable pour optimiser la gestion énergétique des bâtiments résidentiels. C'est avec cette ambition

que notre recherche se fixe quelques objectifs pertinents afin de maximiser l'efficacité énergétique.

1.3 OBJECTIFS

Les objectifs de cette recherche serviront de guide tout au long de notre approche méthodologique et permettront de suivre l'évolution de notre recherche dans le but d'aboutir aux meilleurs résultats possibles. L'objectif général est de concevoir un modèle de prédiction performant de la consommation électrique, basé sur l'apprentissage automatique, capable de fournir des recommandations optimales pour améliorer l'efficacité énergétique des bâtiments résidentiels. Pour y parvenir, les objectifs spécifiques suivants sont définis :

- Développer et évaluer des modèles d'apprentissage automatique pour la prédiction de la consommation électrique résidentielle ;
- Optimiser et identifier le modèle offrant la meilleure précision pour les prédictions futures ;
- Expliquer la prédiction pour connaitre les approches d'optimisation énergétique possible ;
- Proposer des stratégies pour optimiser la consommation électrique résidentielle.

En atteignant ces objectifs, ce mémoire propose une solution exploitable pour la prédiction de la consommation énergétique résidentielle et constitue également un appui aux prises de décision en faveur d'une optimisation durable de l'énergie.

1.4 MÉTHODOLOGIE

La méthodologie adoptée dans ce mémoire constitue une structure clé facilitant l'atteinte des objectifs spécifiques définis. Elle vise à garantir que la solution proposée réponde efficacement à la problématique étudiée.

Dans un premier temps, le contexte a été établi et les méthodes existantes ont été analysées, afin d'évaluer leurs performances ainsi que leurs limites. Cette analyse a permis d'identifier les faiblesses des approches actuelles et d'affiner les techniques mises en œuvre en vue de concevoir une solution plus efficace. Les modèles prédictifs appliqués à la consommation électrique résidentielle ont ensuite été examinés, ainsi que les différentes stratégies d'optimisation disponibles.

Dans un second temps, des données de consommation électrique résidentielle, ainsi que des données météorologiques couvrant une période définie, ont été collectées, analysées et prétraitées. Cette étape a permis d'identifier la structure des données ainsi que les principaux facteurs influençant la consommation énergétique. Une fois ces éléments clarifiés, les données ont été divisées en ensembles d'entraînement et de validation, puis les modèles d'apprentissage automatique les plus adaptés ont été sélectionnés et entraînés. Ces modèles ont ensuite été optimisés par un affinage des caractéristiques des données et un ajustement des paramètres. L'évaluation finale a permis de valider les performances obtenues et de retenir le modèle offrant la meilleure précision pour les prédictions futures.

Dans un dernier temps, le travail a été orienté vers l'explicabilité des prédictions, afin de mieux comprendre les facteurs influant sur l'efficacité énergétique. Une fois le meilleur modèle sélectionné, des techniques adaptées ont été mobilisées pour interpréter les résultats

et en extraire des informations clés sur la consommation. Après identification des principaux déterminants de la consommation énergétique, des recommandations et des stratégies ont été formulées en vue d'un contrôle automatique ou manuel de l'énergie.

1.5 STRUCTURE DU MÉMOIRE

Ce mémoire est structuré comme suit :

Le premier chapitre introduit le projet de recherche consacré à l'optimisation de la consommation d'énergie électrique résidentielle à l'aide de la prédiction de la consommation électrique. Il présente le contexte général du projet, la transition vers une meilleure efficacité énergétique dans les bâtiments résidentiels, ainsi que les objectifs et les impacts attendus pour une consommation optimisée de l'énergie électrique.

Le deuxième chapitre est consacré à un état de l'art. Il rassemble les travaux de recherche pertinents sur les méthodes de prédiction de la consommation énergétique. Il décrit les types de données exploitées, les différentes approches des modélisations ainsi que les méthodes d'explicabilité appliquées, en mettant en avant les contextes d'utilisation.

Le troisième chapitre traite de l'implémentation de notre solution. Il détaille la collecte des données secondaires, les sources exploitées, ainsi que les étapes de prétraitement et de structuration visant à garantir la qualité des informations utilisées dans les modèles prédictifs. Une visualisation des données est également réalisée pour identifier les tendances. Ensuite, l'entraînement des modèles est réalisé en appliquant des techniques d'apprentissage automatique dédiées à la prédiction de la consommation électrique domestique, suivi de l'évaluation des résultats expérimentaux. Les performances sont analysées à l'aide de

plusieurs métriques, et les améliorations apportées par les techniques d'optimisation sont mises en évidence. Afin de justifier le choix du meilleur modèle, une analyse d'explicabilité est conduite pour interpréter les prédictions et identifier les facteurs déterminants. Enfin, ce chapitre met l'accent sur les modèles de fondation, également appelés modèles de pré-entraînement, en explorant leur potentiel dans le cadre de la modélisation énergétique.

Enfin, le quatrième chapitre propose une conclusion générale des travaux, en récapitulant les principales contributions de cette étude, les limites rencontrées et les pistes envisageables pour des recherches futures, notamment en explorant des approches hybrides ou plus avancées pour améliorer la gestion de la consommation énergétique.

CHAPITRE 2

ÉTAT DE L'ART

PRÉDICTION POUR L'OPTIMISATION DE L'ÉNERGIE RÉSIDENTIELLE

Ce second chapitre se focalise sur les concepts existants et les méthodes scientifiques employées pour atteindre l'objectif de la prédiction et de l'optimisation de la consommation d'énergie électrique. Pour cela, dans un premier temps, les dimensions et les types de données exploités ont été détaillés. Ensuite, les principales approches de prédiction couramment utilisées dans le cadre de la consommation électrique ont été présentées. Enfin, les modèles d'apprentissage automatique fréquemment mobilisés pour ce type de prédiction ont été étudiés, en mettant en évidence leurs caractéristiques et leurs domaines d'application. Enfin, le volet explicabilité des prédictions est examiné, avec une attention spéciale aux techniques telles que SHAP et LIME, pour assurer une interprétation claire des résultats des prédictions.

2.1 DESCRIPTION DES DONNÉES

Commençons cette partie de ce mémoire par un élément fondamental pour la prédiction, qui est la donnée. En général, une donnée est une information connue, sur laquelle on peut fonder un raisonnement. Elle peut être collectée, mesurée, analysée pour être utilisée à diverses fins. En intelligence artificielle, elle est définie comme un ensemble d'informations structurées ou non contenant des mesures ou des observations qui sont utilisées pour prendre des décisions (Mathumitha et al., 2024).

Dans le domaine de la prédiction de la consommation électrique des bâtiments résidentiels, plusieurs ensembles de données sont couramment exploités. Parmi ceux-ci

figurent les relevés de consommation électrique, les informations météorologiques extérieures, les données temporelles, les caractéristiques physiques des bâtiments, ainsi que les données relatives à l'occupation des logements. Ces sources permettent de modéliser avec précision les comportements énergétiques en tenant compte des facteurs environnementaux, structurels et humains (Y. Sun et al., 2020). Chaque ensemble de données joue un rôle spécifique et contribue à la qualité de la prédiction, selon ses caractéristiques et sa pertinence. Certaines données sont plus informatives que d'autres, et il peut être difficile d'identifier les plus significatives sans une analyse approfondie. Par exemple, les données de consommation électrique comportent une variable cible, représentant la valeur qu'on choisit de prédire. Elles peuvent donc être qualifiées d'ensemble de données avec étiquettes. Les autres ensembles viennent en complément pour identifier les facteurs explicatifs et faciliter la détection des relations de dépendance, ce qui améliore la précision des prédictions. Une étude récente montre l'importance des données de consommation électrique pour entraîner les modèles de prévision et confirme que l'intégration des paramètres météorologiques améliore la précision des prédictions (Bai, 2024). En effet, la combinaison de ces différentes sources de données permet de capturer une relation claire de la consommation énergétique. Selon (Z. Wang & Srinivasan, 2017), les données météorologiques et d'occupation des bâtiments utilisées pour la prédiction de la consommation électrique dans les articles scientifiques sont respectivement de 60 % et 29 %. Ce qui implique que les données météorologiques sont plus utilisées et participent davantage à la performance des prédictions. Les données de consommation et météorologiques sont donc populaires pour une prédiction efficace de la consommation électrique résidentielle.

Les données de consommation électrique résidentielle présentent plusieurs variables importantes pour les tâches de prédiction. Notamment, la puissance consommée à un instant (t) donné qui s'exprime en watt/heure, la quantité d'énergie utilisée sur une période donnée, consommée par les appareils ou par ligne de phase souvent en kWh, et la consommation totale liée aux différentes charges électriques des équipements (Bai, 2024). Ces variables représentent les profils de consommation sous forme numérique sur différentes échelles temporelles (Mariano-Hernández et al., 2020). Elles sont capturée grâce à des capteurs électroniques selon une fréquence définie. Ces capteurs sont gérés par des compteurs intelligents installés dans le cadre des systèmes de mesure avancés (AMS) pour transmettre les données en temps réel (Lien & Rajasekharan, 2024). La technologie (AMS) est aussi une infrastructure des réseaux intelligents qui permet de mesurer en temps réel la consommation énergétique des ménages (Kim et al., 2023). Diverses méthodes sont donc utilisées pour collecter ces données de façon périodique soit sur une durée de quinze minutes, d'une heure, ou de vingt-quatre heures selon l'application (Y. Sun et al., 2020). Cette fréquence des enregistrements est un facteur clé dans les études de prédiction énergétique. Les données collectées par ces systèmes de mesure avancés présentent un caractère séquentiel dans le temps, formant ainsi des séries temporelles. En effet, la collecte régulière et continue de la puissance consommée et de l'énergie utilisée à des intervalles prédéfinis génère des séquences de données chronologiques. D'autres ensembles de données, notamment ceux liés à la météorologie, sont combinés à celui-ci afin de faciliter la prédiction.

Les données météorologiques sont des informations recueillies sur les conditions atmosphériques et climatiques à un endroit et pendant une période donnée. Elles sont plutôt faciles à obtenir grâce aux stations météorologiques. Elles facilitent l'étude de l'impact des

conditions climatiques sur les comportements de consommation d'énergie, notamment en période de températures extrêmes ou non (Amin & Mourshed, 2024). Selon une étude, elles sont souvent collectées via l'API OpenWeather qui vient compléter les systèmes de mesure avancés pour une collecte plus ou moins complète (Aguirre-Fraire et al., 2024). Elles comprennent des variables clés comme la température, l'humidité, la vitesse du vent et d'autre conditions climatiques générales (Aguirre-Fraire et al., 2024). Selon (Berardi & Jafarpur, 2020) d'ici 2070, des températures hivernales devraient réduire les besoins en chauffage de 18 % à 33 %, tandis que la demande en climatisation augmentera de 15 % à 126 %, en fonction des scénarios climatiques. Ces prédictions météo permettent d'adapter la prédiction de la consommation électrique à l'évolution des conditions climatiques, en particulier pour l'utilisation du chauffage et la de climatisation. La prédiction de la consommation électrique, avec les données météorologiques telles que la température moyenne mensuelle, la vitesse du vent et la pression atmosphérique, a démontré une grande efficacité dans la prédiction (Olu-Ajayi et al., 2022). Cela montre que les données météorologiques créent des relations moins complexes pour une prédiction efficacité par modèles d'intelligence artificielle. Ces données sont donc considérées comme des variables indépendantes, tandis que les données de consommation énergétique, collectées pour la prédiction, représentent les variables dépendantes.

En somme, les données utilisées pour prédire la consommation électrique résidentielle sont variées et complémentaires. Parmi elles, les données de consommation occupent une place principale, car elles sont directement liées à ce qu'on cherche à prédire. Étant collectées à intervalles réguliers, elles constituent des séries temporelles, ce qui rend possible l'analyse

de leur évolution dans le temps. Ce type de données sera exploré plus en détail dans la section suivante.

2.2 SERIES TEMPORELLES

Les séries temporelles, sont un ensemble d'observations successives ordonnées selon le temps ; c'est une chronologie prise à des moments différents. Elles modélisent les relations temporelles et prédisent des tendances, des évolutions futures (Z. Han et al., 2021). Dans le contexte de la consommation énergétique, l'étude des séries temporelles favorise l'analyse, la compréhension et l'anticipation des variations de la consommation au fil du temps, qu'il s'agisse de données horaires, quotidiennes ou mensuelles (Huuki et al., 2024). Selon (Gellert et al., 2022), les séries temporelles peuvent révéler des motifs complexes, tels que des variations selon les moments de la journée ou des saisons, offrant des pistes pour adapter la gestion énergétique en fonction des prévisions météorologiques. Par exemple, elles peuvent révéler des motifs réguliers tels que des pics de consommation en soirée ou des baisses pendant la nuit. Pareil que des observations de consommation différentes pendant l'hiver et l'été. Pour la prédiction des séries temporelles, la méthode traditionnelle recommande de vérifier si elles sont stationnaires ou non. Car la stationnarité permet aux approches statistiques comme ARIMA de repérer facilement la relation entre les données, ce qui conduit à des estimations plus fiables. Pour connaitre cet état de la série temporelle, l'une des méthodes est d'identifier ses caractéristiques. Une fois les caractéristiques connues, il suffit de remarquer l'absence des saisonnalités ou des autocorrélations pour juger que la série est stationnaire. En présence de ces caractéristiques, la série est dite non stationnaire.

Box, Jenkins et Reinsel, dans leur livre (Box et al., 2015), expliquent comment identifier les caractéristiques importantes des séries temporelles, telles que la saisonnalité, les tendances et les autocorrélations. Cela permet d'obtenir des estimations fiables, et donc des prédictions précises. L'identification de ces caractéristiques aide aussi dans le choix des paramètres du modèle de prédiction. Selon, (Sim et al., 2019) la fonction d'autocorrélation (ACF) et la fonction d'autocorrélation partielle (PACF) sont des outils d'analyse visuelle utilisés à cet effets, pour identifier la présence de saisonnalité et pour examiner la stationnarité. Ces fonctions sont aussi utilisées pour visualiser les tendances et les autocorrélations des série temporelle.

L'autre méthode utilisée pour cette tâche de vérification de stationnarité est le test statistique comme le Dickey-Fuller augmenté (Dil, Aakash Ramchand, 2025). Par exemple le test de Dickey-Fuller est utilisé pour détecter la stationnarité d'une série temporelle en testant l'hypothèse nulle selon laquelle la série possède une racine unitaire (Darne & Diebolt, 2007). Une série temporelle possède une racine unitaire si elle est non stationnaire, c'est-à-dire que ses valeurs sont influencées par les tendances dans le temps. Cela signifie qu'il y a une dépendance évolutive dans les données. Grâce à ces deux méthodes, on peut comprendre les tendances, les saisonnalités, de même que les effets réguliers ou non, et conclure si une série est stationnaire ou pas. Dans le cas où la série est stationnaire, on peut passer à la modélisation pour la prédiction avec les modèles statistiques. Dans le cas contraire, les séries non stationnaires sont rendues stationnaires par différentiation, qui est une méthode de transformation où l'on équilibre la moyenne ou la variance de la série.

Voici quelques illustrations typiques de la visualisation des séries temporelles stationnaires et non stationnaires :

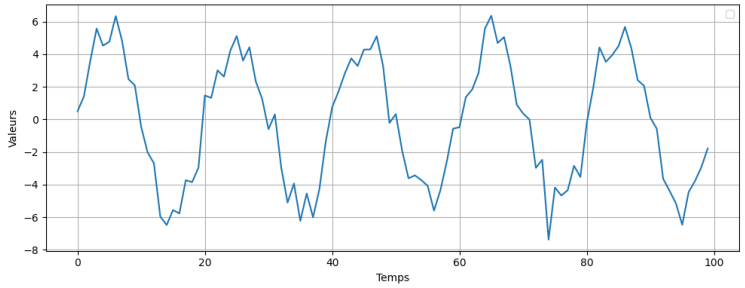

Figure 2.1 : Visualisation de série non stationnaire présentant des saisonnalités.

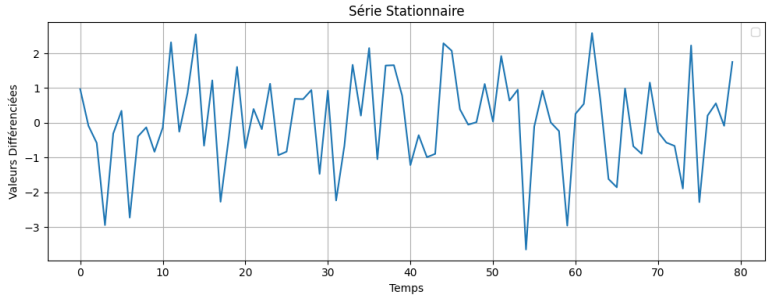

Figure 2.2 : Visualisation de la même série rendu stationnaire par différentiation.

La Figure 2.1 présente des grandes tendances à la hausse et à la baisse, et plus on remarque une variance non constante, c'est-à-dire qui n'est pas répétée, plus cela conduit à dire qu'il n'est pas stationnaire. Quant à la Figure 2.2, elle a été transformée à partir de la première pour un rendu stationnaire, et on remarque des tendances haute et basse à court terme en plus : la variation est presque stable. Comme mentionné en haut, cette conclusion peut être confirmée par des tests mathématiques.

Contrairement aux méthodes statistiques expliquées précédemment, l'intelligence artificielle se montre plus efficace pour prédire les séries temporelles sans avoir à vérifier si la série est stationnaire ou pas. Selon (Kelany et al., 2020), les algorithmes de forêts aléatoires ou le LSTM n'ont pas besoin de tests de stationnarité car elles peuvent apprendre directement à partir des données brutes. Ils prennent en compte les dépendances temporelles sans avoir besoin de rendre les séries stationnaires. Cela permet de gagner du temps sur l'analyse et

également une meilleure prédiction face au méthodes statistique. Il faut noter que les données météorologiques telles que la température, l'humidité ou encore les conditions climatiques sont également associées au temps. Elles peuvent introduire des variations significatives dans la consommation énergétique résidentielle. Ce qui fait que l'analyse associée de ces deux types de données qui sont des séries temporelles peut renforcer la structure des données.

Les modèles classiques de séries temporelles, tels que ARIMA, sont efficaces pour capturer les relations linéaires et saisonnières des données, mais ils présentent des limites lorsqu'il s'agit de modéliser des motifs non linéaires (Chujai et al., 2013). Alors que les techniques d'intelligence artificielle peuvent de dépasser cette limite pour comprendre les schémas des données complexes. Il arrive qu'on combine ces deux types de modèles pour former des approches de prédiction hybrides. Par exemple, un modèle ensembliste pourrait associer la structure d'un ARIMA avec la capacité d'apprentissage des réseaux neuronaux, pour aboutir à une prédiction beaucoup plus robuste et précise. Les approches de prédiction pourraient donc également contribuer à améliorer la performance des modèles.

2.3 APPROCHES DE PREDICTION

Comme défini plus haut, la prédiction demande l'application de certaines techniques, qui sont utilisées de manière différente et appelées modèles. Le modèle de prédiction axé sur les données (black-box) est connu comme la méthode souvent utilisée récemment dans le domaine de la prédiction de l'énergie du bâtiment (Banik et al., 2021). Il peut utiliser uniquement les données pour effectuer des prédictions rapides et précises, sans avoir besoin d'informations supplémentaires (Wei et al., 2018). Dans le but d'exploiter pleinement ces modèles, les approches de prédiction offrent des structures permettant de personnaliser ou de

combiner les modèles. Ces modèles peuvent être complémentaires et proposent des solutions adaptées en fonction des objectifs et de la nature des données disponibles (Amasyali & El-Gohary, 2018). Chaque modèle repose sur des principes de fonctionnement et des fonctions de calcul spécifiques, ce qui fait que certains modèles peuvent présenter des avantages là où d'autres montrent des limites. Cette logique a conduit au développement de méthodes de prédiction classées en trois catégories, notamment la méthode unique, la méthode ensembliste et la méthode avancée, encore appelée apprentissage profond (Y. Sun et al., 2020).

2.3.1 MÉTHODE UNIQUE

La méthode unique est une approche de prédiction qui utilise un seul modèle traditionnel ou d'intelligence artificielle pour effectuer des tâches. Il se base sur une technique statistique ou un algorithme d'apprentissage automatique (Y. Sun et al., 2020). Cette approche se concentre sur une seule variable cible pour la prédiction. Parmi les modèles fréquemment utilisés, on retrouve l'ARIMA (Sim et al., 2019), la régression linéaire (Fumo & Rafe Biswas, 2015), les arbres de décision (B. Han et al., 2022), les machines à vecteurs de support (SVM) (Y. Chen et al., 2017), ainsi que d'autres qui sont décrit plus bas. Selon (Z. Wang & Srinivasan, 2017), Les avantages de cette approche sont la fiabilité, la facilité d'implémentation et la rapidité de calcul. Ces approches sont bien adaptées dans les cas où les ressources sont limitées alors qu'on veut faire la prédiction. En revanche, elle présente parfois une précision limitée et nécessite un meilleur choix de l'algorithme. Cependant, pour corriger les limites de ce dernier, l'étude de la méthode ensembliste est faite pour voir si elle démontre de meilleures performances en termes de précision.

2.3.2 LES MÉTHODES ENSEMBLISTES

La méthode ensembliste, comme son nom l'indique, combine plusieurs modèles uniques de prédiction basés sur l'intelligence artificielle pour améliorer la précision des prédictions. Elles additionnent les avantages de différents modèles individuels pour obtenir de meilleures performances globales. Selon une étude, il est possible d'utiliser des algorithmes de base similaires appelés intégration homogène ou des algorithmes différents nommés intégration hétérogène pour construire nos modèles ensemblistes (R. Wang et al., 2020). Pour l'appliquer, des techniques de combinaison parallèle **(Bagging)** (Nagauri, 2020) ou de combinaisons séquentielles **(Boosting)** (T. Chen & Guestrin, 2016) ou encore d'empilement **(Stacking)** (Mohammed et al., 2021) des algorithmes de prédiction sont utilisées afin d'éviter des biais et le surapprentissage des modèles.

BAGGING

Le Bagging est une technique ensembliste qui est caractérisée par l'entraînement de plusieurs modèles indépendants sur des sous-ensembles aléatoires des données, créés grâce au Bootstrap. C'est-à-dire que le Bootstrap génère des échantillons au hasard de données à partir des originales, cela permet d'obtenir plusieurs ensembles d'apprentissage, un peu différents mais similaires. Chaque modèle est ensuite testé sur l'un de ces sous -échantillons de données, et les prédictions finales sont obtenues en combinant les résultats de tous les modèles. Dans notre cas d'étude, les prédictions finales des modèles sont combinées via une moyenne. C'est pourquoi il est dit que le Bagging, visent à tirer parti des avantages de chaque modèle individuel en combinant leurs prédictions afin de réduire la variance et d'augmenter la robustesse (Z. Wang & Srinivasan, 2017). L'une des implémentations les plus populaires

du bagging dans le contexte de la consommation d'énergie est la forêt aléatoire (RF) (Biau & Scornet, 2016), qui utilise des arbres de décision pour capter la relation entre les différentes variables, comme les conditions météorologiques et la consommation électrique (Dostmohammadi et al., 2024). Une autre étude, (Pham et al., 2020) confirme que le modèle de forêt aléatoire combine plusieurs arbres de décision via le bootstrap et une sélection aléatoire de variables météorologiques et de consommation énergétique, pour améliorer la précision globale des prédictions. Bien que le Bagging permette de réduire la variance et d'améliorer la robustesse du modèle, il peut être difficile de combiner et de raffiner efficacement les différents modèles pour assurer leur compatibilité et leur complémentarité. La Figure 2.3 illustre bien son architecture.

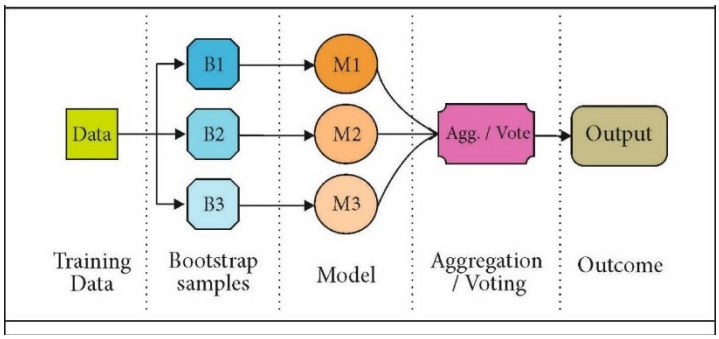

Figure 2.3 : Schéma illustratif du Bagging (Nagauri, 2020).

BOOSTING

Après avoir exploré le Bagging, il est essentiel de s'intéresser au Boosting. Le Boosting est aussi une méthode ensembliste qui construit des modèles de plus en plus performants en se concentrant sur les erreurs des modèles précédents. À chaque étape, les échantillons qui ont été mal prédits reçoivent une priorité de prédiction, ce qui oblige le modèle suivant à se concentrer davantage sur ces erreurs. Ce qui veut dire que les erreurs des modèles précédents

sont corrigées progressivement au même moment, plusieurs modèles sont entraînés de manière séquentielle, et leurs prédictions sont ensuite combinées à chaque étape. Son utilisation réduit les erreurs de prédiction et optimise les performances par rapport aux autres méthodes ensembliste (Abd El-Aziz, 2022). Parmi les algorithmes populaires de boosting, on trouve les Gradient Boosting Machines (GBM) (Sivakumar et al., 2024) et l'Extreme Gradient Boosting (XGBoost) (Vu et al., 2023), qui sont utilisés pour ajuster des modèles faibles, comme des arbres de décision. Ces techniques sont particulièrement efficaces dans des contextes où les données sont non linéaires, comme celles liées à la consommation d'énergie, notamment lors de périodes de forte demande énergétique (Dostmohammadi et al., 2024). L'ensemble du modèle de moyenne mobile autorégressive intégrée et de l'arbre de régression par gradient boosting (ARIMA-GBRT) fait aussi l'objet de plusieurs études pour améliorer les performances de prévision des séries temporelles dans la consommation électrique (Lu et al., 2025). Il combine les forces de l'ARIMA pour les données linéaires et du GBRT qui construit un ensemble d'arbres de décision successifs pour les données non linéaires (Nie et al., 2021). Ce type d'approche améliore considérablement la précision des prédictions dans des contextes tels que la consommation énergétique. Les principales limites du boosting sont le risque élevé de surapprentissage, le coût computationnel important, la complexité du réglage des hyperparamètres, la sensibilité aux données aberrantes, et la difficulté d'interprétation des modèles résultants. La Figure 2.4 illustre bien l'architecture du boosting.

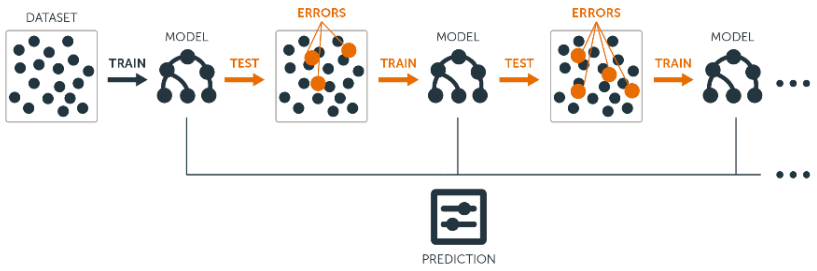

Figure 2.4 : Schéma illustratif du fonctionnement du Boosting (Kumar, 2020).

STACKING

Enfin, le Stacking, consiste à empiler plusieurs modèles de base de types différents, tels que l'arbres de classification et de régression, la forêt aléatoire, le modèle d'arbre M5 (Akgündoğdu et al., 2019) et XGBoost, comme illustré dans une étude (Mohammed et al., 2021). Dans cette étude, les modèles de base servent d'entrée pour les prédictions, et un méta-modèle est chargé d'apprendre à combiner de manière optimale les prédictions des modèles de base afin d'améliorer la performance globale, comme l'illustre la Figure 2.5. Grâce au méta modèle, on peut faire une combinaison de modèles complexes et de modèles simples ou linéaires pour obtenir une prédiction de plus grande précision ce qui fait la force du stacking. Cette efficacité du stacking est également confirmée par (Ali et al., 2024), dans une étude où le stacking combine les modèles XGBoost, LGBM et HGB pour atteindre un RMSE le plus bas démontrant une réduction considérable des erreurs de prédiction par rapport à d'autres méthodes avec une précision de 91 %. Sa force se trouve aussi dans la bonne sélection des modèles de base, chaque modèle doit apporter sa spécialité. Par exemple, un modèle peut bien gérer les tendances linéaires, tandis qu'un autre gère des relations polynomiales pour obtenir une solution plus robuste et précise. Il peut également présenter plusieurs limites, telles que la complexité du processus d'implémentation, la nécessité de

données supplémentaires pour entraîner le méta-modèle, une sensibilité à la qualité et la complémentarité des modèles de base, ainsi qu'un risque potentiel de rajustement si les modèles ne sont pas bien ajustés.

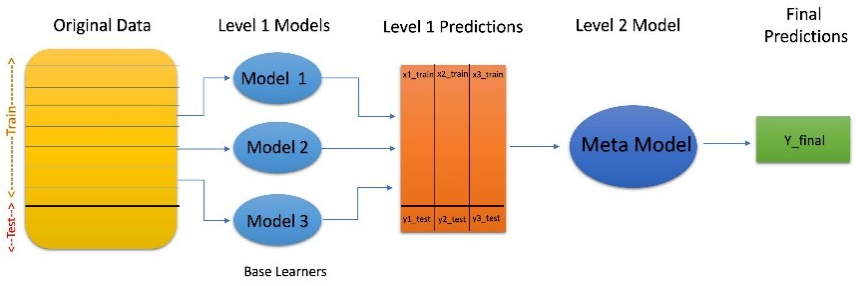

Figure 2.5 : Schéma illustratif de l'architecture du Stacking (*Stacking in Machine Learning*, 2021).

2.3.3 MÉTHODE D'APPRENTISSAGE PROFOND

Abordons à présent la méthode avancée qui, comparativement aux précédentes pourrait renforcer d'avantage la performance des prédictions. Plusieurs approches de prédiction en apprentissage automatique peuvent être identifiées comme avancées, mais dans notre cas, l'attention est davantage portée sur la méthode de l'apprentissage profond. Cet apprentissage se base sur les réseaux de neurones pour faire la prédiction, c'est une structure qui est composée de plusieurs neurones superposé formant des couches. Par défaut le réseau est constitué de trois couches comme le montre la Figure 2.6, notamment la couche d'entrée des données, la couche de traitement et la couche de sortir de prédiction. La couche de traitement est la principale, elle est cachée et c'est elle qui s'occupe d'apprendre à partir données. Dans

le bloc de traitement, le nombre de couche peut augmenter en fonction des tâches et objectifs visés.

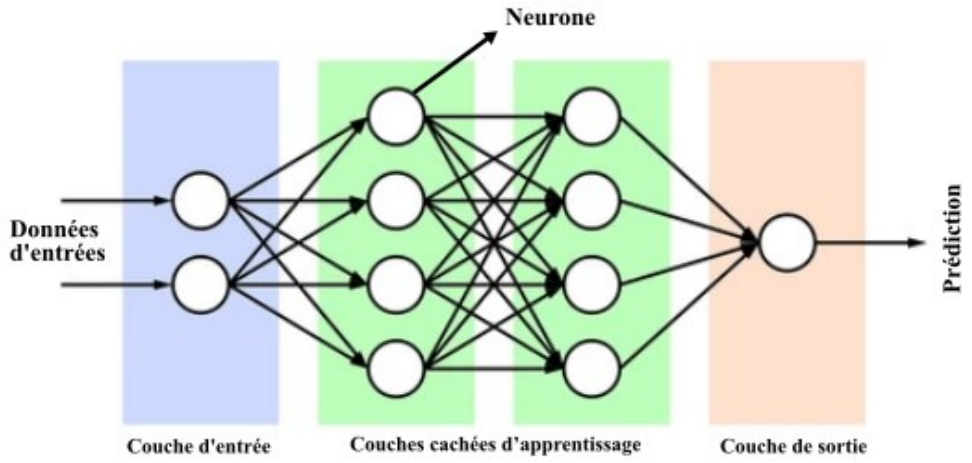

Figure 2.6 : Schéma illustratif d'un exemple d'architecture d'apprentissage profond.

Chacune des couches analyse les données reçues et envoie la nouvelle version à la couche suivante, ainsi de suite jusqu'à parcourir toutes les couches prédéfinies. Un neurone qui reçoit une information fait une opération avant d'envoyer cette dernière à l'entrée des neurones suivants. Dès qu'il reçoit une information, il applique un poids à celle-ci pour montrer son niveau d'importance comme le montre la Figure 2.7. Ensuite il ajoute un biais qui est une valeur ajustant de la somme pondérée pour définis quand es ce que la fonction d'activation sera actif ou non.

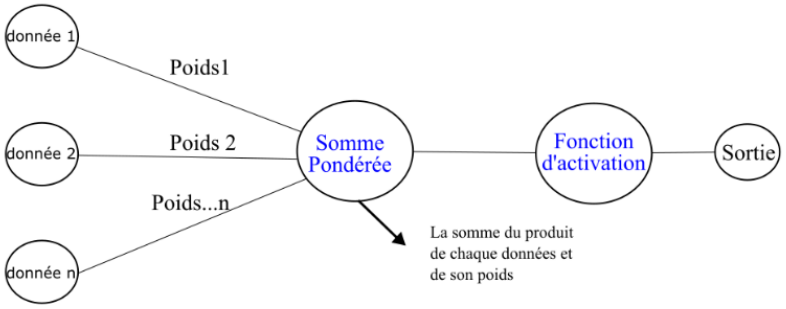

Figure 2.7 : Schéma illustratif du traitement dans un neurone.

C'est une méthode itérative qui propage l'information à travers des couches successives de neurones, en apprenant automatiquement les caractéristiques pertinentes (Chassagnon et al., 2020). C'est une approche qui ajuste ses paramètres en minimisant une fonction de perte, en utilisant des techniques comme la descente de gradient ou la rétropropagation. Parmi ces modèles, les réseaux de neurones convolutifs (CNN) (Ullah et al., 2020) et les réseaux de neurones récurrents (RNN) (Shachee et al., 2022), possèdent des architectures avancé pour traiter des problèmes complexes et exigent généralement un volume de données d'entraînement plus important (Hsu et al., 2025). Les données de consommation électrique étant des séries temporelles, on s'est intéressé plus au RNN qui est développé plus bas parce qu'il est bien adapté aux séries temporelles.

2.4 APPRENTISSAGE AUTOMATIQUE

L'apprentissage automatique est une discipline de l'intelligence artificielle basée sur des données qui exploite des algorithmes où des modèles supervisés et non supervisés sont appliqués pour traiter de grandes quantités de données historiques et contextuelles afin de

prédire (Mathumitha et al., 2024). L'apprentissage automatique appliqué aux données des réseaux électriques, permet de faire des prévisions énergétiques futures et aussi de comprendre la relation entre les données, afin de réduire les coûts de consommation et de production. Ils permettent de résoudre les incertitudes liées à la gestion des coûts et à l'efficacité énergétique, notamment pour la planification de la production et de la distribution énergétique pour minimiser les pertes (Banik et al., 2021). Une autre étude souligne qu'il permet de détecter des anomalies dans les systèmes intelligents, ainsi que la réduction des émissions de gaz avec des techniques de contrôle basées sur les prévisions (Mariano-Hernández et al., 2020).

Deux différents types apprentissages sont généralement utilisés, dont l'apprentissage supervisé et l'apprentissage non supervisé. L'apprentissage supervisé apprend à partir des données d'entraînement étiquetées pour établir des relations entre les entrées et les sorties, il compare la sortie prédite à la sortie réelle après un entraînement (Bourhnane et al., 2020). Alors que selon (Mathumitha et al., 2024), la méthode non supervisée est constituée d'algorithmes de clustering tels que le clustering par k-means (Chévez et al., 2017) et le clustering flou (AbuBaker, 2021), qui sont utilisés pour analyser des données non étiquetées et identifier des motifs ou des structures cachés dans les données. Dans l'apprentissage supervisé, une sortie correspondante *(y)* est liée pour chaque donnée d'entrée *(y)*. La relation entre x et y est représentée par $y = f(x)$, où *(y)* est l'étiquette ou la variable cible (Radhoush et al., 2023). Étant donné que les données historiques de consommation électrique sont étiquetées et sont aussi des valeurs continues, l'apprentissage supervisé est priorisé dans notre cas d'étude avec une attention particulière au modèle de régression. Certaines études récentes montrent que l'apprentissage supervisé est bien adapté pour prédire la consommation

d'énergie en raison de sa capacité à identifier des relations entre les variables explicatives et les variables cibles (Klyuev et al., 2022). Selon (Yuan et al., 2018), l'apprentissage supervisé est particulièrement recommandé pour la prédiction dans le domaine de l'énergie électrique, en raison de sa précision élevée et de sa capacité à analyser et traiter des données. Dans le même sens, (Albahli et al., 2020) ont utilisé des modèles supervisés tels que la régression à vecteurs de support (SVR), les forêts aléatoires et XGBoost, appliqués à des ensembles de données historiques, afin de prédire la consommation future et d'estimer les prix des factures d'électricité. Cette technique de prédiction est appliquée à la fois pour les bâtiments individuels et pour des ensembles urbains, ce qui montre leur capacité à s'adapter à des échelles et des complexités variées dans la prédiction de la consommation énergétique (Fathi et al., 2020). Toutes ces études et beaucoup d'autres ont montré l'efficacité des algorithmes supervisés dans l'exécution des tâches de prédiction de la consommation électrique.

La pratique commune de ces études est l'entraînement de plusieurs algorithmes de régression de façon séparée sur un même ensemble de données pour enfin comparer leur performance. Cela permet de mieux comprendre le comportement des données et de choisir le modèle qui donne la meilleure performance ou, s'il le faut, de passer à un modèle hybride. Cette méthode est appliquée toujours dans le but de faire une prédiction efficace et robuste. Ce document se concentre alors sur quelques algorithmes d'apprentissage supervisé, qu'ils soient uniques ou d'ensembles, pour assurer une prédiction efficace de la consommation électrique domestique.

2.4.1 FORÊT ALÉATOIRE

Selon une étude de (Sayed et al., 2023) qui est axée sur la prévision de la consommation d'énergie électrique, l'utilisation d'ARIMA a permis d'avoir une prédiction avec 93 % de précision. Ensuite, sa combinaison avec la forêt aléatoire (RF), qui a traité les relations non linéaires dans les données, a aussi permis d'atteindre 97 % de précision générale. Les forêts aléatoires sont basées sur les arbres de décision comme l'arbre de classification et de régression ; elles utilisent des techniques de bagging et de sélection aléatoire des variables. Les arbres sont entraînés sur des sous-échantillons créés par Bootstrap. À chaque nœud, un sous-ensemble aléatoire de variables est sélectionné pour déterminer la meilleure scission (Lauzon & Gloaguen, 2024). Cela diminue la variance et renforce la robustesse des prédictions en combinant les résultats de plusieurs arbres de décision indépendants. Il est important de contrôler la taille des arbres pour éviter un rajustement. Plus y a d'arbres et moins y a de risque de surapprentissage, mais cela augmente le temps de calcul (Lauzon & Gloaguen, 2024). Une étude (Biau & Scornet, 2016) présente une procédure basique pour développer cet algorithme. La forêt aléatoire implique des étapes dans lesquelles les partitions des données dans les arbres ne dépendent pas de l'ensemble d'apprentissage. Pour comprendre les propriétés théoriques, des modèles de forêts purement aléatoires ont été étudiés, où les données sont normalisées dans un espace $x = [0, 1]d$. Tout d'abord, dans cet espace, toutes les données sont utilisées directement, sans rééchantillonnage. Ensuite, à chaque nœud d'un arbre, une coordonnée est choisie aléatoirement parmi tous les ensembles de dimensions. Enfin, une coupure est effectuée au centre de l'intervalle courant pour la coordonnée sélectionnée. Ce processus est répété k fois, où k est un paramètre fixe, jusqu'à ce que chaque arbre atteigne k niveaux, formant un arbre binaire complet comportant 2^k

feuilles. Pour effectuer une prédiction, on identifie la cellule correspondant au point x dans chaque arbre, puis on calcule la moyenne des valeurs xi associées aux données situées dans cette cellule. Ce mécanisme assure une partition régulière et aléatoire de l'espace, les estimations étant obtenues par une simple moyenne dans les régions définies par les feuilles. La formulation générale, proposée par (Khalil et al., 2022) est représenté par l'équation (2.1):

$$f = \frac{1}{K}\sum_{k=1}^{K} f_{k(X')} \qquad \textbf{(2.1)}$$

K est le nombre d'arbres dans la forêt,

$f_{k(X')}$ est la prédiction individuelle réalisée par le k-ième arbre du modèle pour les données d'entrée X',

f est la prédiction finale obtenue.

Plusieurs hyperparamètres clés influencent sa performance notamment le nombre d'arbres, le nombre de variables à chaque nœud, la profondeur maximale de chaque arbre, le nombre minimal des échantillons dans une feuille. L'ajustement de ces paramètres permet d'optimiser le modèle en améliorant sa précision, en réduisant le temps de calcul (Khalil et al., 2022).

2.4.2 BOOSTING CATEGORIEL

CatBoost (Categorical Boosting) est un modèle d'apprentissage automatique développé par de Yandex (L. Zhang et al., 2023). Il peut fonctionner sur différents formats de données, ce qui accroît son efficacité par rapport à d'autres modèles de machine learning pour traiter des problèmes de régression et de classification. Il se distingue des autres modèles

de régression par sa capacité à appréhender et à mieux représenter la relation entre différents types d'ensembles de données, ce qui en fait un algorithme particulièrement adapté à la prédiction de la consommation électrique, une variable de nature continue.

Selon (Li et al., 2024) le CatBoost utilise la méthode du gradient boosting où des arbres de décision sont combinés pour créer un modèle prédictif efficace. Cette méthode lui permet de construire une série d'arbres de décision de manière progressive. L'idée principale est de créer chaque nouvel arbre en se basant sur les erreurs commises par les arbres précédents. Au départ, le modèle crée un premier arbre de décision pour faire une prédiction initiale. Ensuite, à chaque nouvelle itération, un nouvel arbre est ajouté, mais cette fois-ci, il se concentre sur les données mal prédites par les arbres précédents. Par exemple, si certains points de données ont été mal évalués, le nouvel arbre accorde plus d'importance à ces points pour essayer de corriger les erreurs. C'est ce processus d'ajout d'arbres pour améliorer progressivement les prédictions qui est appelé le Gradient Boosting. Le modèle continue ainsi jusqu'à atteindre un nombre d'arbres défini ou jusqu'à ce que les améliorations deviennent minimes. Dans le cas où le modèle ne s'améliore plus suffisamment après un certain nombre d'itérations, il arrête de construire de nouveaux arbres.

Par exemple, pour prédire la consommation électrique avec des données horaires, de température extérieure, des consommations passées et du jour de la semaine, l'algorithme commence par créer un premier arbre de décision (Uddin et al., 2024). Puis estime la consommation initiale en fonction de ces variables. Ensuite, à chaque itération, un nouvel arbre est ajouté pour corriger les erreurs des prédictions précédentes. Par exemple, si la consommation a été mal prédite lors d'une journée froide, le modèle accordera plus

d'importance à ces erreurs et ajustera sa prédiction en fonction de cette variable. Ce processus continue jusqu'à ce que l'amélioration donne une valeur prédictive proche de la réelle.

Une autre étude de (F. Zhang et al., 2022) explique que CatBoost gère efficacement les variables catégorielles et numériques. Un de ces points forts est sa capacité à gérer les données catégorielles sans transformation complexe. Il utilise une méthode unique basée sur la permutation aléatoire pour attribuer des valeurs numériques aux catégories. Ce qui lui permet d'éviter les prétraitements lourds souvent nécessaires avec d'autres modèles.

Le CatBoost est un modèle puissant et flexible pour traiter des problèmes complexes de régression, notamment ceux liés à la prédiction de la consommation énergétique. Grâce à sa capacité à gérer à la fois des variables numériques et catégorielles sans nécessiter de prétraitements complexes, il offre une solution robuste pour traiter des ensembles de données variés, comme les informations temporelles et météorologiques pour la consommation électrique. De plus, l'approche du Gradient Boosting permet d'améliorer progressivement les prédictions en ciblant les erreurs des arbres précédents, assurant ainsi une précision optimale. Cette combinaison de caractéristiques fait du CatBoost un choix pertinent pour des tâches de prédiction de valeurs continues, telles que la consommation d'électricité. En somme, le CatBoost offre une solution efficace et pragmatique pour aborder des problèmes de prévision dans des domaines variés, y compris l'optimisation de la gestion de l'énergie.

2.5 RESEAU DE NEURONES RECURRENT (RNN)

Les réseaux neurones récurrents sont une extension des réseaux de neurones traditionnels capables d'avoir un comportement temporel dynamique pour permettre l'utilisation d'états cachés et de sorties précédentes comme entrées (Yazdan et al., 2022). Ce

qui signifie qu'ils peuvent utiliser les états cachés pour conserver en mémoire les données précédentes. Cette mémoire interne permet au réseau d'utiliser non seulement les entrées actuelles, mais aussi de prendre en compte les sorties précédentes pour faire des prédictions futures. Ces informations passées conservées en mémoire, sont mise à jour à chaque nouvelle entrée, pour permettre de les réutiliser avec les nouvelles comme entrées pour les étapes suivantes.

En pratique, cela se traduit par la capacité du modèle à analyser une séquence chronologique, comme la consommation d'énergie quotidienne. Supposons qu'on veuille prédire la consommation d'un bâtiment pour un jour suivant en fonction de la consommation des jours précédents. Le modèle ne se limite pas à la consommation du jour précédent la prédiction, mais prend également en compte plusieurs jours antérieurs présentant des conditions similaires afin d'améliorer l'exactitude de la prédiction. En d'autres termes, le RNN utilise les données de consommation passées pour mieux comprendre comment la consommation change au fil du temps et des conditions afin de faire des prédictions plus précises pour l'avenir. C'est pourquoi on dit que les RNN ont des connexions récurrentes, ce qui leur permet de traiter des informations à travers le temps et en fonction des facteurs. Cette fonction récurrente fait leur particularité pour prédire des séries temporelles stationnaire ou non, où chaque nouvelle prédiction s'appuie sur des données passées pour offrir des résultats plus cohérents.

Toujours selon (Yazdan et al., 2022), leur structure inclut un état interne qui leur permet de traiter des séquences d'entrées de différentes longueurs, en reliant les sorties de tous les neurones à leurs entrées. Cet état interne est également mis à jour en fonction de l'entrée actuelle et de l'état précédent, créant une boucle récurrente. Cela permet au réseau de capturer

des dépendances temporelles et de se souvenir des informations passées, tout en envoyant les sorties à d'autres neurones des couches suivantes pour une prédiction précise. Chaque neurone reçoit une entrée, qu'il traite à l'aide des paramètres avant de produire une sortie. Chaque entrée est multipliée par une valeur de connexion appelé le poids, déterminant l'importance de cette entrée, puis un ajustement aussi appelé le biais qui est ajouté pour affiner la sortie. Les résultats sont ensuite combinés dans une somme pondérée, qui passe à travers une fonction d'activation, permettant au neurone de moduler sa sortie. Cette fonction d'activation introduit une non-linéarité qui permet au réseau neurone d'apprendre des modèles complexes à partir des données disponible (Mienye et al., 2024). Ces fonctions sont utilisées dépendamment des types de tâche à exécuter et des natures des données, on peut citer la fonction sigmoïde, la fonction tanh, ReLU et ELU. La même étude de (Mienye et al., 2024) note les cas d'utilisation de chaque type de fonction d'activation. Pour avoir des sorties probabilistes, elle conseille l'utilisation de la fonction sigmoïde car elle transforme une valeur en un nombre compris entre 0 et 1. Quant à tanh (hyperbolic tangent), elle transforme les entrées en valeurs comprises entre -1 et 1, ce qui la rend adaptée aux séquences comportant des valeurs positives et négatives. Le ReLU (Rectified Linear Unit) pour sa part, renvoie l'entrée si elle est positive et zéro sinon, pour aider à atténuer le problème du gradient évanescent. Pour améliorer la rapidité d'apprentissage, la ELU (Exponential Linear Unit) est utilisé et elle accepte les valeurs négatives contrôlées pour stabiliser le réseau.

Au vu de l'ensemble des résultats, il est possible d'affirmer que le RNN présente une robustesse notable et permet de prendre en compte un grand nombre de détails liés à l'analyse et à la prédiction. Mais il faut également prendre en compte ses faiblesses, notamment le problème de vanishing gradient souvent rencontré et leur capacité à gérer les dépendances à

long terme qui est quand même importante pour estimer avec précision la consommation d'énergie sur des périodes étendues (Kolluru et al., 2024).

2.6 INGÉNIERIE DES CARACTERISTIQUES

L'ingénierie des caractéristiques est un processus essentiel dans l'optimisation des modèles prédictifs. Plusieurs techniques peuvent être employées selon le type de données et l'objectif de prédiction visé. On utilise, d'une part, l'analyse en composantes principales (ACP) pour transformer les caractéristiques existantes, et d'autre part, une approche manuelle visant à créer de nouvelles variables, afin d'évaluer leur capacité à améliorer les résultats obtenus précédemment. Les performances des différentes approches sont ensuite comparées.

Analyse en composantes principales

Selon une étude de (Verdonck et al., 2024), l'ACP (Analyse en Composantes Principales) est une technique de réduction de dimension appliquée aux variables, qui permet de transformer pour simplifier des données en conservant les informations importantes. Elle fonctionne en créant de nouvelles variables qui sont des combinaisons linéaires des variables explicatives, appelées composantes principales. Pour assurer son application, une évaluation de lien entre les variables est faite, pour choisir celles qui fournissent le plus de renseignements et de les standardiser en appliquant la formule (2.2). Cela garantit une intégration optimale dans l'analyse.

$$z = \frac{x-\mu}{\sigma} \qquad \textbf{(2.2)}$$

$\boldsymbol{x}$: la valeur de la variable,

$\boldsymbol{\mu}$: la moyenne de la variable et

$\boldsymbol{\sigma}$: l'écart-type de la variable.

Ainsi, l'Analyse en Composantes Principales s'avère être une méthode incontournable pour explorer et interpréter les données complexes, en mettant en évidence les relations structurelles entre les variables et en facilitant la visualisation et l'extraction d'informations pertinentes.

Création manuelle de nouvelles variables

Une autre approche couramment utilisée repose sur la construction manuelle de nouvelles variables dérivées des données initiales. Selon (Čistý et al., 2024), la méthode de construction algébriques et physique des variables d'entrée influence grandement l'efficacité des modèles utilisés ultérieurement et est considérée comme une méthodologie pertinente.

Les variables temporelles créées incluent, par exemple, l'identification des jours de semaine et des fins de semaine, le numéro de semaine afin de capter d'éventuels effets saisonniers, et aussi l'encodage des saisons (hiver, printemps, été) sous forme de variables catégorielles à l'aide d'un encodage one-hot.

Les variables physiques incluent la température ressentie qui est exprimé en degrés Celsius, obtenue par la relation entre température extérieure et l'humidité en pourcentage, comme le montre la formule 2.3. Cela permet de traduire la sensation thermique perçue par les occupants, qui est facteur clé dans l'utilisation du chauffage.

$$Température\ ressentie = Temperature\ exterieure - 0.7 \times humidité \qquad \textbf{(2.3)}$$

De même, l'écart à la température de consigne représente la différence entre la température intérieure mesurée et celle souhaitée par les usagers.

Les variables statistiques tel que la moyenne mobile permettent de lisser les variations et de faire ressortir les tendances générales, tandis que l'écart-type mobile sur la même période mesure la variabilité de la consommation. Ces variables offrent une mémoire contextuelle sur le comportement énergétique à court terme aux modèles.

Ces différentes approches d'ingénierie des caractéristiques constituent un facteur clé dans l'amélioration des modèles de prédiction appliqués à la consommation énergétique.

2.7 AJUSTEMENT DES HYPERPARAMETRES

L'ajustement des hyperparamètres est une étape cruciale dans l'optimisation des modèles prédictifs. Les hyperparamètres contrôlent le comportement de l'apprentissage et influencent directement la précision et la robustesse des modèles. Les techniques comme la recherche en grille (grid search), l'hyperbande (hyperband) et l'optimisation bayésienne (bayesian optimization) sont utilisées pour automatiser ce réglage (Yang & Shami, 2020).

Recherche en grille

La recherche en grille consiste à tester exhaustivement toutes les combinaisons possibles des valeurs d'hyperparamètres définies dans une grille (Bergstra & Bengio, 2012). Elle permet d'identifier la combinaison optimale selon une métrique de performance (par exemple RMSE), mais peut être très coûteuse en temps de calcul lorsque le nombre d'hyperparamètres ou la taille de la grille est élevé.

Optimisation bayésienne

Selon une étude de (Hertel et al., 2020), l'optimisation bayésienne est une recherche d'hyperparamètres qui utilise un modèle qui, pour chaque itération, sélectionne le paramètre le plus prometteur en fonction des résultats antérieurs. Cette étude décrit aussi que, lors de son application, une mise à jour bayésienne est effectuée pour ajuster cette estimation, afin d'atteindre le maximum de la fonction avec peu d'essais en peu de temps. Le processus de minimisation comprend souvent trois composantes principales notamment un modèle gaussien pour la fonction objectif, un processus bayésien de mise à jour qui modifie le modèle gaussien après chaque nouvelle évaluation de la fonction objectif, et une fonction d'acquisition (Injadat et al., 2018).

Hyperbande

L'hyperbande une technique d'optimisation dont le fonctionnement est d'allouer davantage de ressources aux configurations d'hyperparamètres les plus prometteuses, puis élimine progressivement celles qui performent le moins bien (J. Wang et al., 2018). Les résultats d'une étude ont démontré que l'hyperbande réduit considérablement le temps d'entrainement des modèles d'apprentissage profond (Falkner et al., 2018) .

2.8 EXPLICABILITE DES MODELES

Après une prédiction, l'explicabilité est importante pour comprendre la décision du modèle. L'explicabilité permet non seulement de faire comprendre la prédiction, mais aussi de donner des détails sur les facteurs importants qui influencent celle-ci. Dans notre cas, l'explicabilité aide à mieux comprendre les facteurs qui influencent la consommation

d’énergie. Elle permet d'expliquer les raisons derrière une prévision de consommation et d'adopter des comportements adaptés pour assurer l’efficacité énergétique. Des facteurs comme la température ou le temps d’utilisation d’un appareil spécifique peuvent influencer la consommation. Les identifier permet de prendre des décisions efficaces pour réduire la consommation. Prenons l'exemple d'un modèle de prédiction utilisé sur les données électriques d'une maison pour estimer la consommation d'énergie au cours des jours suivants. Supposons qu'il prédit une forte consommation d'énergie pour une journée particulièrement froide dans une saison d’hiver. L’analyse de l'impact de ces facteurs peut démontrer que la température froide augmente la demande en chauffage. En plus, l'utilisation prolongée du chauffage pendant la journée augmente excessivement la consommation électrique. Grâce à cette explication, les utilisateurs peuvent prendre des décisions afin de réduire considérablement leur facture.

Dans cette revue (Linardatos et al., 2020), les auteurs décrivent l’explicabilité comme étant la structure logique interne et le mécanisme d’un système d’apprentissage automatique. Autrement dit, l'explicabilité rend transparents les processus techniques et les relations entre les données d'entrées et de sorties du modèle pour permettre de comprendre comment et pourquoi il prend une décision pendant son entraînement. Donc, plus le modèle est explicable, plus il est facile à quiconque de comprendre ses prises de décision. Avec les avancées récentes en intelligence artificielle, les modèles comme les réseaux de neurones profonds sont largement utilisés pour améliorer la précision des prédictions par rapport aux modèles basiques. Toutefois, leur complexité ne permet pas la compréhension de leurs décisions et il est aussi difficile aux techniques d’explicabilité de donner assez de détails. Deux techniques sont couramment utilisées, notamment l'explication additive de Shapley

(SHAP) et l'explication locale interprétable indépendante du modèle (LIME). Elles permettent d'analyser et de donner une explication technique et littéraire des décisions des modèles complexes, même si leur application aux modèles avancés reste un défi.

2.8.1 SHAP

La technique SHAP, est celle la plus utilisée pour l'explicabilité et l'interprétabilité des modèles de ML dans le domaine de la consommation d'énergie. Selon (Dinmohammadi et al., 2023), SHAP est une méthode d'explicabilité qui permet d'interpréter les résultats des modèles en attribuant à chaque variable d'entrée une valeur représentant son impact sur la prédiction. Il permet ainsi d'identifier l'importance de chaque caractéristique dans une tâche de prédiction. Cette valeur attribuée à une variable spécifique représente son niveau d'implication ou sa contribution à la prédiction finale du modèle.

Selon (Linardatos et al., 2020), le SHAP est basé sur une théorie qui utilise les valeurs de Shapley, qui sont une solution équitable pour attribuer une importance à chaque variable dans une prédiction donnée. Quant à sa fonction technique, la formule de base pour calculer les valeurs SHAP repose sur la théorie des valeurs de Shapley. Pour une caractéristique ***i*** dans un ensemble de caractéristiques ***f***, la valeur SHAP $\boldsymbol{\phi i}$ est définie par la formule (2.4) :

$$\boldsymbol{\phi i} = \sum_{S \subseteq ni} \frac{(|\ s\ |\ !|n|-|s|-1)!}{|n|!} [f(s \cup i) - f(s)] \qquad \textbf{(2.4)}$$

n est l'ensemble des caractéristiques,

s est un sous-ensemble de ***n*** sans ***i***,

f(s) est la prédiction du modèle en utilisant uniquement les caractéristiques de ***s***,

$\boldsymbol{\phi i}$ représente la contribution moyenne de $\boldsymbol{i}$ sur toutes les combinaisons possibles de caractéristiques,

$\boldsymbol{f(s)}$ est la prédiction du modèle pour le sous-ensemble S des caractéristiques.

$\boldsymbol{f}(\boldsymbol{s} \cup \boldsymbol{i})$ est la prédiction du modèle quand on ajoute la caractéristique i au sous-ensemble s.

Cette formule repose sur l'idée que la contribution d'une caractéristique {i} est calculée comme la moyenne pondérée des variations des prédictions du modèle lorsque cette caractéristique est incluse ou non, en tenant compte de toutes les combinaisons possibles des autres caractéristiques. Cela signifie que pour savoir à quel point une caractéristique {i} est importante dans la prédiction d'un modèle, on compare les prédictions faites avec et sans cette caractéristique, en testant toutes les configurations possibles des autres caractéristiques.

Ensuite, on calcule la moyenne des écarts entre ces prédictions, et on accorde plus ou moins de poids à chaque écart selon l'importance de la configuration testée. Cela permet d'obtenir une valeur SHAP qui reflète l'effet réel et équitable de la caractéristique *{i}* sur la prédiction, en tenant compte de ses interactions avec les autres. Comme toutes les combinaisons n'ont pas la même importance, la pondération est utilisée pour représenter le nombre de permutations dans lesquelles l'ensemble S des caractéristiques apparaît avant {i}. Cette pondération est essentielle pour s'assurer que chaque caractéristique soit évaluée équitablement, en tenant compte de toutes les situations possibles où elle interagit avec les autres. Prenons l'exemple d'un modèle qui prédit la consommation électrique quotidienne en utilisant les caractéristiques comme la température extérieure, le jour de la semaine et la consommation des appareils en veille. Supposons qu'un jour donné, le modèle prédit une consommation de 25 kWh. Pour comprendre l'importance de chaque caractéristique, on peut analyser l'impact de leur absence sur la prédiction. Sans la variable de température extérieure,

la prédiction passe à 28 kWh, ce qui signifie que la température contribue à réduire la consommation de 3 kWh, probablement parce qu'une température plus basse réduit l'utilisation de la climatisation. Si l'on retire la variable jour de la semaine, la prédiction reste à 25 kWh, indiquant que cette variable n'a pas d'impact particulier sur ce jour précis.

Enfin, sans la variable de consommation des appareils la veille, la prédiction baisse à 23 kWh, montrant que ces appareils augmentent la consommation de 2 kWh. Cependant, la particularité des valeurs SHAP est qu'elles vont au-delà de ces analyses simples. Elles prennent en compte toutes les combinaisons possibles des trois caractéristiques, car l'influence d'une variable peut changer selon la présence ou l'absence des autres.

Après les teste des combinaisons, SHAP attribue une valeur moyenne pondérée, tel que T=la température égale à -3 kWh, le jour de la semaine égale à +0,5 kWh, et l'appareils en veille est égale à +1,5 kWh. Ces valeurs montrent que la température réduit la consommation de 3 kWh en moyenne, les appareils en veille l'augmentent de 1,5 kWh, et le jour de la semaine l'augmente faiblement (+0,5 kWh). Cela permet d'obtenir une estimation plus fiable et équitable de l'importance réelle de chaque facteur sur la consommation électrique. Pour une lecture plus facile, les résultats peuvent être visualisés pour montrer l'effet de chaque variable sur toutes les prédictions ou pour classer les variables par importance.

Tout ceci démontre que SHAP aide à comprendre un modèle et même à l'optimiser. Il peut être utilisé sur un modèle pour comprendre les entrées qui influence afin de reprendre son entrainement avec les variables importantes pour se rapprocher de la réalité. Bien que SHAP soit un outil puissant pour expliquer les modèles du ML, il présente plusieurs limites, notamment une complexité computationnelle élevée qui peut ralentir les calculs, surtout pour

les grands ensembles de données. Il est aussi parfois difficile à appliquer sur des modèles avancés comme les réseaux de neurones. De plus, SHAP peut être sensible à l'échantillonnage, ce qui impacte la précision des résultats. Enfin, il peut avoir des difficultés d'interprétation lorsqu'il y a beaucoup d'interactions entre les variables. Ces limitations doivent être considérées pour s'assurer que l'utilisation de SHAP soit efficace et appropriée dans le contexte donné.

2.8.2 LIME

L'explications Locales Interprétables et Indépendantes d'un Modèle (LIME) est une approche qui permet de comprendre comment un modèle d'apprentissage automatique a généré une prédiction donnée. Contrairement à SHAP, qui repose sur la théorie des valeurs de Shapley pour attribuer une importance globale aux variables, LIME adopte une approche locale en construisant un modèle interprétable autour d'une observation spécifique. Selon (ElShawi et al., 2021), LIME est une technique d'interprétabilité locale qui simplifie le comportement d'un modèle complexe en construisant un modèle explicatif autour d'une prédiction spécifique. Elle explique aussi qu'elle fonctionne en générant des versions légèrement modifiées de l'observation d'origine, en entraînant un modèle simple sur ces nouvelles données, puis en donnant plus de poids aux exemples les plus proches de l'observation initiale.

Pour donner plus de détails sur ce principe de fonctionnement du LIME, l'idée est que, même si le modèle est complexe, on suppose que, tout près de l'exemple à expliquer, son comportement est plus simple. Concrètement, LIME commence par sélectionner une prédiction spécifique à expliquer. Il génère ensuite plusieurs instances légèrement modifiées

de l'exemple initial en altérant certaines valeurs des caractéristiques. Pour chacune de ces instances, il sollicite la prédiction du modèle, ce qui permet d'observer l'impact des variations locales sur le résultat obtenu. Elle donne plus d'importance aux exemples les plus proches de l'instance d'origine, puis construit un petit modèle simple qui imite le comportement du modèle complexe, mais uniquement autour du cas étudié. Ce petit modèle est facile à comprendre et permet de voir quelles variables ont eu le plus d'influence sur la prédiction. Ainsi, il explique la décision du modèle de façon claire et compréhensible, sans dépendre de la complexité du modèle global. Cela aide à expliquer, avec des mots simples ou des facteurs clairs, ce qui a influencé la prédiction pour un exemple particulier. Selon (Ribeiro et al., 2016) également, LIME cherche à minimiser une fonction de perte $\mathrm{l}(\mathrm{f}, \mathrm{g}, \Pi\mathrm{x})$ qui mesure la différence entre le modèle original nommé f et le modèle explicatif nommé g dans la zone de x, tout en maintenant la complexité du modèle explicatif à un niveau acceptable$\big(\Omega(\mathrm{g})\big)$. L'explication $\varepsilon(x)$ produite par LIME est obtenue par l'équation (2.4) :

$$\varepsilon(x) = arg\min_{g \in G} \mathrm{l}(\mathrm{f}, \mathrm{g}, \Pi\mathrm{x}) + \Omega(\mathrm{g}) \quad \textbf{(2.5)}$$

$arg\min_{g \in G}$: le modèle g qui minimise la fonction dans l'ensemble G

LIME, en d'autres termes, cherche à minimiser une fonction de perte qui évalue la différence entre le modèle original qui effectue les prédictions de base et le modèle explicatif simplifié. Elle crée un modèle explicatif qui reproduit les comportements du modèle complexe de manière aussi précise que possible, mais avec une structure plus simple. En même temps, elle veille à ce que cette simplicité n'altère pas trop la capacité du modèle explicatif à capturer les relations essentielles du modèle de base. Ce qui le rend plus adapté au modèle complexe comme les réseaux de neurones, puisqu'elle donne des explications

locales qui montrent comment ces modèles réagissent à des entrées spécifiques, sans avoir à expliquer tout le modèle dans son ensemble.

Prenons l'exemple de la prédiction de la consommation d'énergie pour une journée donnée x, en fonction des conditions météorologiques et de la consommation électrique passée. Le modèle complexe f peut, par exemple, être un réseau de neurones qui estime la consommation d'énergie à partir de plusieurs variables d'entrée. Le modèle explicatif g est un modèle plus simple, comme une régression linéaire, qui cherche à expliquer la prédiction du modèle complexe f d'une manière plus facile. La mesure de proximité est la proximité des observations similaires à x, c'est-à-dire des journées ayant des conditions météorologiques et de consommation proche de celles de x. Ensuite, on mesure la fidélité, qui évalue dans quelle mesure l'explication donnée par g est fidèle à la prédiction du modèle complexe f pour les observations similaires à x. La complexité est une mesure qui évalue à quel point le modèle explicatif g est simple. L'objectif enfin est de minimiser la somme de la fidélité et de la complexité, ce qui revient à trouver un modèle explicatif g qui est à la fois fidèle au modèle complexe f pour cette prédiction spécifique x et suffisamment simple à comprendre. Ce principe de LIME le rend très pratique car il peut expliquer le fonctionnement de n'importe quel modèle d'apprentissage automatique, et ce peu importe sa complexité. Il le rend aussi flexible car il peut fournir des explications sous plusieurs formes textuelles ou graphiques en montrant l'importance des différentes variables.

Cependant, LIME n'est pas aussi parfait. Ses explications peuvent parfois manquer de stabilité. Par exemple, dans l'étude de (Ribeiro et al., 2016), ils démontrent qu'en répétant l'opération, on n'obtient pas toujours exactement les mêmes résultats. De plus, il se concentre uniquement sur des cas locaux, sans donner une vue d'ensemble sur l'influence générale des

variables sur le modèle. Il faut noter aussi que la qualité des explications dépend beaucoup de la manière dont LIME génère ses données d'exemple pour faire ses calculs.

Les limites des deux approches d'explicabilité des modèles d'apprentissage automatique, LIME et SHAP, permettent de conclure qu'en les combinant, on obtient une vision plus complète et fiable des raisons derrière les prédictions d'un modèle, qu'il soit simple ou complexe. En effet, LIME fournit des explications locales spécifiques pour des prédictions données, en cherchant à rendre les décisions du modèle plus compréhensibles dans des contextes précis. D'un autre côté, SHAP offre une vue d'ensemble plus cohérente et stable des importances des caractéristiques à travers l'ensemble du modèle, permettant de mieux comprendre l'impact global de chaque variable. En utilisant ces deux approches ensemble, on bénéficie à la fois de la précision locale de LIME et de la consistance globale de SHAP, ce qui permet de mieux expliquer et comprendre les mécanismes sous-jacents des prédictions d'un modèle, qu'il soit simple ou complexe.

2.9 DISCUSSION

Au vu de tout ce qui précède, on remarque que les données électriques et météorologiques sont des séries temporelles qui sont largement utilisées pour la prédiction de la consommation énergétique domestique. Ensuite, parmi les approches de prédiction utilisées, les approches avancées basées sur l'apprentissage profond semblent corriger certaines limites des approches simple et ensembliste et montrent de meilleurs résultats. En plus de cela, les modèles RNN qui sont de la famille de l'apprentissage profond ont montré une bonne performance dans l'analyse et la prédiction des séries temporelles. Il faut aussi souligner que la plupart des études testent plusieurs modèles et approches pour choisir celui

avec la meilleure performance. Enfin, après l'étape de la prédiction, il est important d'expliquer celle-ci afin de comprendre comment l'utilisation de l'électricité est faite pour prendre les meilleures décisions d'optimisation de consommation. Pour cela, les techniques LIME et SHAP sont exploitées et les avantages du SHAP prouvent qu'il semble être mieux adaptée pour expliquer les modèles de réseaux de neurones. Au vu de tout ceci que l'on pourra comprendre et anticiper au mieux la consommation de l'électricité dans les ménages. L'évaluation des comportements et des tendances de consommation à court et à long terme va permettre de développer des mesures préventives pour divers problèmes liés à l'optimisation (Yazdan et al., 2022). Donc, autant que les mesures et les règles d'optimisation dépendent des résultats et des explicabilités de la prédiction, il est très important de suivre la bonne démarche pour trouver le modèle le plus adapté à notre contexte.

CHAPITRE 3

PRÉDICTION DE LA CONSOMMATION ÉLECTRIQUE RÉSIDENTIELLE

Ce chapitre présente l'implémentation d'une solution prédictive pour la consommation électrique résidentielle, en s'appuyant sur les études précédentes et en utilisant divers modèles d'apprentissage automatique et profond. Le modèle offrant la meilleure performance est retenu, et SHAP est utilisé pour assurer une explicabilité optimale.

Pour ce faire, une collecte de données secondaires sur la consommation électrique des ménages, sur une période déterminée, est réalisée, puis ces données sont traitées et visualisées. Les résultats des modèles entraînés sont analysés, et ceux dont les prédictions sont les plus proches de la réalité sont identifiés avant d'être optimisés afin de renforcer leur précision et de se rapprocher davantage des valeurs réelles.

Enfin, le meilleur modèle optimisé est expliqué afin de mettre en lumière les facteurs et les ajustements possibles pour l'optimisation de la consommation électrique résidentielle. Une simulation de prédiction à partir des modèles préentraînés est également réalisée.

3.1 COLLECTE DE DONNÉES

Dans le cadre de la prédiction de la consommation d'électricité, il est possible d'utiliser des données primaires ou secondaires. Comme indiqué dans la section consacrée à l'état de l'art, la collecte de données primaires exige des ressources et un temps considérable. En raison de ces contraintes, ce mémoire s'appuie sur des données secondaires pour la prédiction de la consommation résidentielle d'électricité.

Les données secondaires correspondent à des informations déjà collectées et publiées, puis réutilisées dans un contexte différent de celui de leur collecte initiale. Dans cette section dédiée à la collecte des données, elles sont recueillies et traitées afin d'analyser leur structure et de déterminer les étapes suivantes de la prédiction. Elles sont ensuite préparées et réparties en ensembles d'entraînement, de validation et de test.

3.1.1 BASE DE DONNÉES

L'ensemble de données collecté provient du catalogue de données d'Hydro-Québec, société d'État, et est intitulé « ***Consommation électrique de la clientèle participant à un programme de gestion locale de la demande de puissance*** » (Hydro-Québec, 2024). Pour avoir des données résidentielles fiables et de qualité, la présente méthode s'est axée sur des bases fiables, d'où l'attention prêtée aux catalogues de données d'Hydro-Québec. Hydro-Québec est un producteur d'électricité pour tous les secteurs consommateurs. Il utilise des compteurs communicants installés dans les ménages pour mesurer et collecter des données de consommation en temps réel. Contrairement aux compteurs traditionnels, les compteurs intelligents transmettent automatiquement les données de consommation au fournisseur d'électricité, sans qu'il soit nécessaire de relever le compteur manuellement (Kemal & Olsen, 2016). Cette technologie appartient à la famille des systèmes avancés de gestion de l'énergie. Elle repose sur l'utilisation de compteurs intelligents qui mesurent la consommation des appareils domestiques à une fréquence prédéfinie. Les données sont ensuite transmises de manière sécurisée vers une base de données, où elles sont analysées afin d'optimiser l'établissement des factures, d'améliorer l'efficacité énergétique et de faciliter la transition vers des sources plus durables.

Les données collectées par les compteurs sont mises à disposition de manière sécurisée pour chaque utilisateur souhaitant consulter sa consommation. De la même manière, les données d'électricité globales, regroupées par groupe d'utilisateurs, par province ou par secteur, sont rendues publiques. C'est le cas de l'ensemble de données utilisé dans cette étude, constitué d'informations anonymisées sur la consommation électrique de la clientèle d'Hydro-Québec participant à un programme nommé Hilo. Dans le cadre de ce programme, trois postes électriques desservent un groupe de clients dans une région de Montréal, et la consommation horaire de chaque client raccordé à un poste est agrégée afin d'obtenir la consommation totale de l'ensemble des clients desservis pour une fréquence et une période donnée, comme illustré dans le Tableau 3.1.

Tableau 3.1 : Présentation incomplet de l'ensemble de données

horodatage_local	clients_connectes	energie_totale_consommee	temperature_consigne_moyenne	temperature_interieure_moyenne	tstats_intelligents_connectes	...	jour_s
2022-12-19T18:00:00-05:00	96	597.881645	18.515519	19.689610	593	...	
2022-12-20T01:00:00-05:00	97	319.050076	17.385694	18.871443	598	...	
2022-12-20T02:00:00-05:00	97	326.860057	17.349507	18.744714	598	...	
2022-12-20T03:00:00-05:00	97	316.814168	17.387108	18.653374	598	...	
2022-12-20T04:00:00-05:00	97	320.871218	17.391231	18.578144	598	...	
...	...	...	...	...	...	...	
2022-10-16T03:00:00-04:00	21	29.326270	13.521076	21.216195	135	...	
2022-10-16T06:00:00-04:00	21	35.732509	13.524052	21.042261	135	...	
2022-10-16T09:00:00-04:00	21	50.634153	13.541909	21.362951	135	...	
2022-10-16T12:00:00-04:00	21	44.023085	13.467351	21.680587	135	...	
2022-10-16T15:00:00-04:00	21	61.497146	13.408720	21.550698	135	...	

Le Tableau 3.1 illustre un extrait représentatif de l'ensemble des données utilisées dans cette étude. Il présente quelques lignes et colonnes sélectionnées, notamment l'horodatage local, le nombre de clients connectés, l'énergie totale consommée (en kWh), ainsi que les températures intérieures et de consigne moyennes. Cet aperçu permet de

constater la granularité temporelle des mesures, la diversité des variables collectées et leur rôle dans la prédiction de la consommation électrique résidentielle.

3.1.2 DESCRIPTION ET STRUCTURE DE DONNÉES

L'ensemble de données secondaires collecté auprès de la clientèle d'Hydro-Québec est constitué de données de consommation d'électricité résidentielle, de données de température intérieure mesurée à l'aide de thermostats intelligents, ainsi que de données météorologiques recueillies toutes les heures grâce à l'API Weatherbit. Ces données sont collectées sur une période de plus de deux ans soit du début janvier 2022 au fin juin 2024 à des fréquences d'une heure. L'ensemble est composé de 64 605 lignes de données horaire des utilisations par ménage combinées par poste électrique et de 30 colonnes de caractéristiques temporelles, de consommations, météorologiques et techniques. La structure de l'ensemble de données se présente comme le montre le Tableau 3.2.

Tableau 3.2 : La structure de l'ensemble de données utilisé pour la prédiction.

Nom de la colonne	Type	Description
Column 1	int64	Identifiant unique (peut être ignoré)
poste	object	Code du poste (ex. : A, B, C)
date	object	Date au format AAAA-MM-JJ
heure_locale	int64	Heure de la mesure (0 à 23)
horodatage_local	datetime	Date et heure complète avec fuseau horaire

clients_connectes	int64	Nombre de clients connectés
energie_totale_consommee	float64	Énergie consommée
temperature_consigne_moyenne	float64	Température de consigne moyenne
temperature_interieure_moyenne	float64	Température intérieure moyenne
tstats_intelligents_connectes	int64	Nombre de thermostats connectés
irradiance_solaire_moyenne	float64	Irradiance solaire moyenne
humidite_relative_moyenne	int64	Humidité relative moyenne
precipitations_neige_moyenne	float64	Précipitations neigeuses moyennes
vitesse_vent_moyenne	float64	Vitesse moyenne du vent
temperature_exterieure_moyenne	float64	Température extérieure moyenne
type_evenement	object	Type d'événement
indicateur_evenement	int64	Indique si un événement est en cours (0 ou 1)
pre_post_indicateur_evenement	int64	1 = avant, 2 = pendant, 3 = après un événement
mois	int64	Mois (1 à 12)
jour	int64	Jour du mois (1 à 31)
jour_semaine	int64	Jour de la semaine (0 = lundi)

mois_cos	float64	Encodage cyclique du mois (cosinus)
mois_sin	float64	Encodage cyclique du mois (sinus)
jour_semaine_cos	float64	Encodage cyclique du jour (cosinus)
jour_semaine_sin	float64	Encodage cyclique du jour (sinus)
indicateur_weekend	bool	Indique si c'est un week-end
indicateur_jour_ferie	bool	Indique si c'est un jour férié
indicateur_weekend_ferie	bool	Week-end ou jour férié
heure_sin	float64	Encodage cyclique de l'heure (sinus)
heure_cos	float64	Encodage cyclique de l'heure (cosinus)

3.2 TRAITEMENT DE DONNÉES

La structure de l'ensemble de données présentée met en évidence la nécessité d'un prétraitement adapté. En particulier, l'objectif de la prédiction est d'estimer la variable cible **énergie totale consommée** à partir des autres variables d'entrée. Cela implique des étapes de nettoyage des données, de gestion des valeurs manquantes, de normalisation et de préparation des variables explicatives afin d'assurer la qualité des prédictions.

La base exploitée pour la collecte des données mentionne que celles-ci sont brutes et sans garantie de qualité (Hydro-Québec). Le traitement des données consiste donc à les transformer en informations propres, compréhensibles et exploitables. Selon (Lei et al., 2021), le prétraitement des données brutes est une étape essentielle avant l'entraînement d'un

modèle, qui garantit sa stabilité et sa performance. Un bon traitement des données contribue ainsi directement à la qualité des prédictions.

Dans cette étude appliquée aux séries temporelles, le traitement comprend principalement le nettoyage et l'analyse des données, ainsi que la sélection des caractéristiques, car l'ensemble contient de nombreuses variables peu pertinentes pour ce cas d'étude. Enfin, une visualisation est réalisée pour de mieux comprendre les relations entre les caractéristiques.

3.2.1 NETTOYAGE ET ANALYSE

Le nettoyage des données vise à identifier et corriger les erreurs et incohérences présentes dans l'ensemble brut afin de le rendre exploitable. Selon (Côté et al., 2024), cette étape essentielle de la préparation des données consiste à détecter et supprimer les erreurs. Notre ensemble de données comporte de nombreuses caractéristiques, comme le montre le Tableau 3.2, dont certaines sont peu pertinentes pour la prédiction de la consommation électrique. Ainsi, cette phase de nettoyage permet d'éliminer les erreurs et de retirer les variables non pertinentes ou redondantes, afin d'obtenir un ensemble de données plus clair et exploitable. Les caractéristiques retenues sont directement liées à la consommation électrique, à la météorologie et au temps, ce qui facilite la visualisation et une prédiction plus efficace.

Informations générales sur les données

La vérification des informations générales et des statistiques de base des colonnes a permis d'identifier les éléments suivants :

- Nombre total de lignes : 64 605
- Nombre total de colonnes : 30
- Aucune **valeur nulle**
- Aucune **valeur manquante** : Toutes les colonnes sont complètes sauf celle de la variable **'type_evenement'** dont la gestion sera faite dans suite après d'autre analyse approfondie.
- Aucun **doublon présent** dans l'ensemble.
- Types de variable : Toutes les variables sont avec des types adaptés sauf, la date qui est de type **'object'** qui a été convertie en type **'datetime'** pour faciliter l'analyse temporelle.

Valeurs aberrantes

Pour mieux analyser l'ensemble afin de vérifier la distribution des données, et voir les valeurs aberrantes et d'autre anomalie, le Tableau 3.3 permet de faire le diagnostic complet. Ce tableau présente un résumé statistique des principales variables du jeu de données, permettant de comprendre leur distribution et leurs caractéristiques générales.

Tableau 3.3 : Présentation partielle des statistiques des caractéristiques.

	Column 1	heure_locale	clients_connectes	energie_totale_consommee	temperature_consigne_moyenne	temperature_interieure_moyenne
count	64605.000000	64605.000000	64605.000000	64605.000000	64605.000000	64605.000000
mean	32302.000000	11.688228	53.158641	150.532166	16.156486	20.988465
std	18650.001408	6.819488	28.379021	191.722162	2.227084	1.985601
min	0.000000	0.000000	9.000000	7.450200	9.315278	16.210841
25%	16151.000000	6.000000	36.000000	61.270458	14.610974	19.412990
50%	32302.000000	12.000000	50.000000	109.811443	16.871208	20.258545
75%	48453.000000	18.000000	68.000000	195.553576	17.971604	22.706080
max	64604.000000	23.000000	104.000000	32240.173270	21.032967	27.084625

Selon le Tableau 3.3, toutes les variables présentent des statistiques descriptives cohérentes, sauf la variable energie_totale_consommee, qui présente un max de 32 240,17 KWh, alors que la médiane (50 %) est égale à 109,81 KWh. Cela indique des valeurs extrêmes à explorer. La méthode Boxplot est très performante en matière de détection et de visualisation des valeurs aberrantes, et améliore considérablement les prévisions de consommation électrique (S. Sun et al., 2017). C'est une méthode de vérification de valeurs aberrantes sur une variable. Elle est utilisée pour décrire la distribution des données numériques avec la valeur minimale (MIN), le quartile inférieur (Q1), la médiane (Q2), le quartile supérieur (Q3) et la valeur maximale (MAX), avec le MIN et le MAX généralement définis comme l'indiquent les formules (3.1) et (3.2) :

$$MIN = Q1 - 1{,}5 * IQR, \quad (3.1)$$

$$MAX = Q3 + 1{,}5 * IQR, \quad (3.2)$$

Où

IQR : l'écart interquartile,

Q1 : le quartile inférieur,

Q2 : la médiane,

Q3 : le quartile supérieur,

soit : $IQR = Q - Q1.$ (3.3)

Toutes valeurs qui se trouvent en dehors de l'intervalle [MIN, MAX] sont alors considérées comme aberrantes à la suite de leur analyse. Cette application donne le résultat de la Figure 3.1. Cette Figure représente un boxplot de la variable energie_totale_consommée, exprimée en kWh. Ce type de graphique est utilisé pour visualiser la distribution statistique des

données, en mettant en évidence la ligne centrale dans la boîte les limites de la boîte et les points situés en dehors des moustaches.

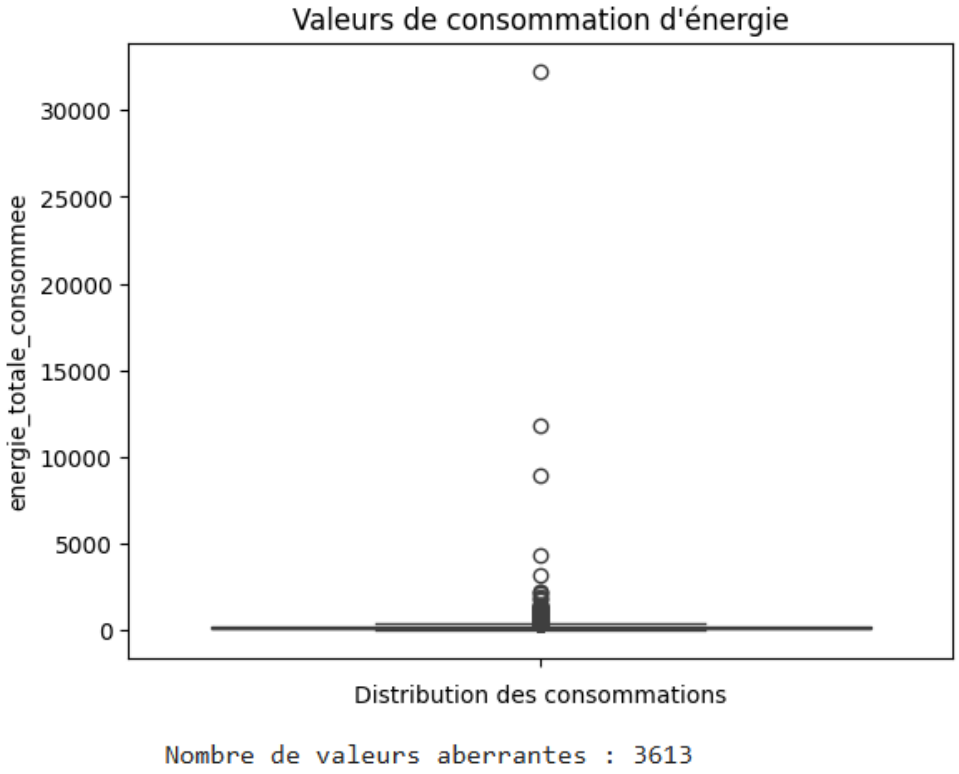

Figure 3.1 : Visualisation des valeurs aberrantes de la consommation de l'énergie.

On observe sur la Figure 3.1 que la plupart des points se concentrent entre 0 et 600 kWh environ et que quelques points sont isolés en haut, jusqu'à 30 000 kWh. Au total, 3613 valeurs aberrantes sur 64 605 observations, soit 5,6 % du total. Ces points sont des valeurs aberrantes trop élevées, elles pourraient être dues à des anomalies d'enregistrement de données ou à de l'utilisation excessive du chauffage en hiver ou encore de la climatisation en été. Cette hypothèse conduit à une vérification de ces consommations abusive afin de connaitre les vraies causes pour savoir comment les gérer. Pour cela, la Figure 3.2 présente deux boxplots comparant la distribution de la variable energie_totale_consommée selon que

l’on soit en période hivernale (True) ou non hivernale (False). L’axe des abscisses distingue ces deux catégories, tandis que l’axe des ordonnées indique les valeurs de consommation d’énergie, exprimées en kWh.

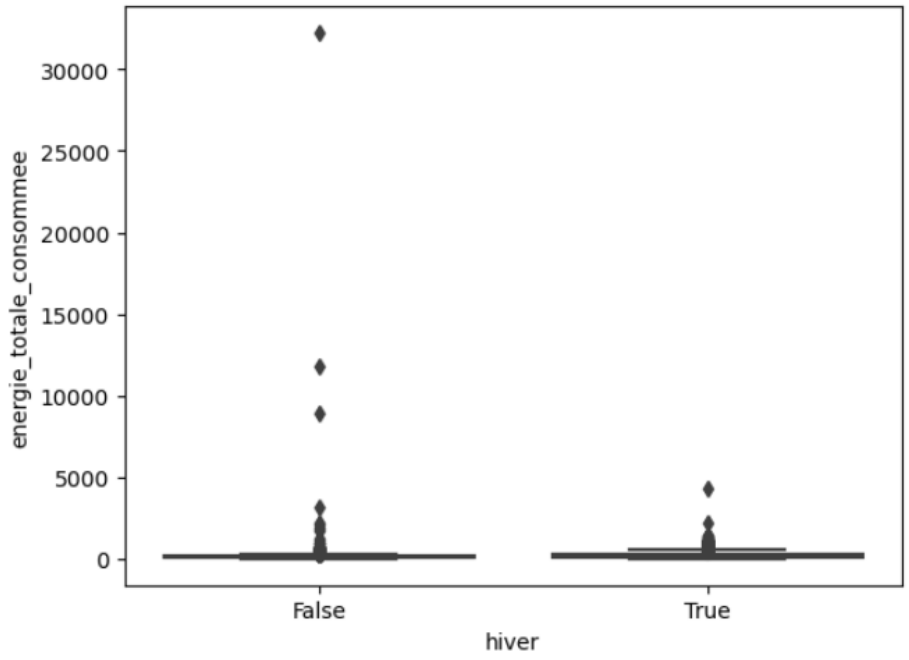

Figure 3.2 : Visualisation des valeurs aberrantes en hiver

La Figure 3.2 montre une distribution des données aberrantes entre les mois d'hiver et les autres mois et montre que les valeurs extrêmes jusqu’à 32 000 kWh ne sont pas dues à l’hiver. Les valeurs aberrantes en hiver sont moins extrêmes et, en plus, dans une limite acceptable. Donc la consommation en hiver n’explique pas les valeurs aberrantes trop extrêmes. La même visualisation est faite pour l’été afin d’éliminer le facteur météorologique de l’hypothèse.

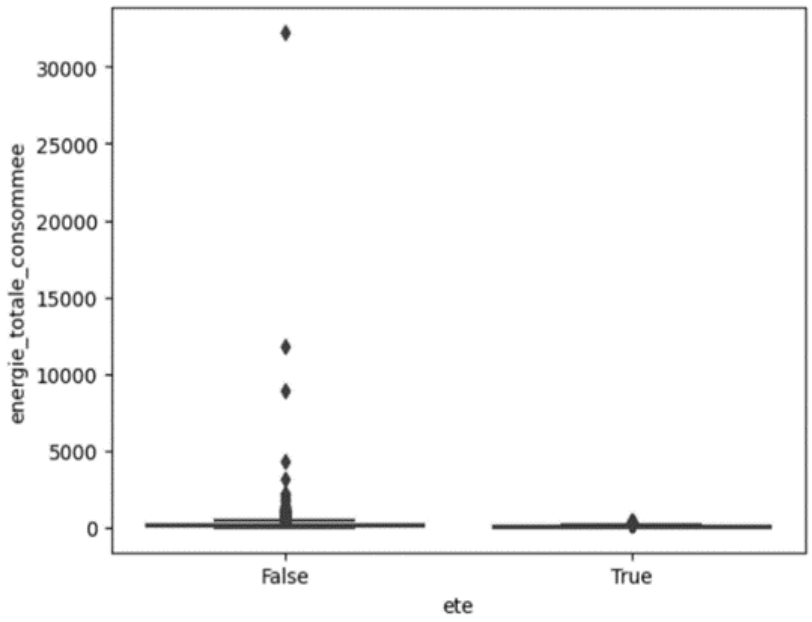

Figure 3.3 : Visualisation des valeurs aberrantes en été.

Comme le montre la Figure 3.3, la période estivale ne contient presque pas de valeurs aberrantes. Elle n'explique donc pas la présence d'anomalies aussi extrêmes. Ces graphiques permettent ainsi d'écarter l'hypothèse selon laquelle les valeurs de consommation très élevées, atteignant jusqu'à 32 000 kWh, seraient liées aux saisons d'hiver ou d'été. La suite de l'analyse s'attache à explorer d'autres facteurs susceptibles d'expliquer une telle anomalie. Par ailleurs, une visualisation de la consommation totale par mois pourrait aider à vérifier si ces valeurs atypiques proviennent plutôt du comportement des utilisateurs. Pour cela, la Figure 3.4 présente des boxplots représentant la distribution de l'énergie totale consommée en kWh pour chaque mois de l'année, de janvier (mois 1) à décembre (mois 12). Chaque boîte illustre les valeurs minimales, maximales, médianes et les quartiles de la consommation mensuelle.

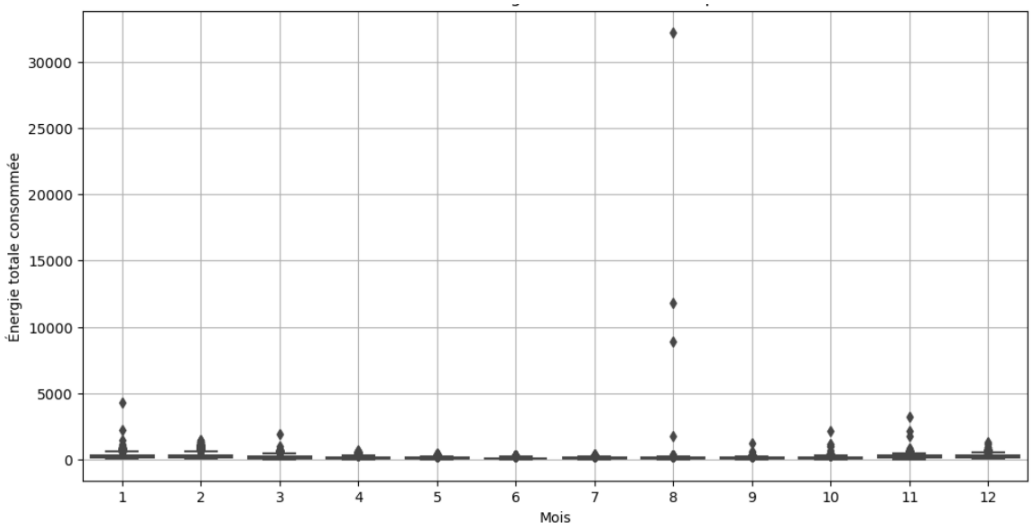

Figure 3.4 : Visualisation des valeurs aberrantes par mois.

Comme le montre la Figure 3.4, tous les mois sont dans la limite de la médiane de la consommation, sauf le mois d'août qui contient les points trop élevés inexpliqués. Ce mois ne démontre aucune particularité causant une telle consommation, sauf si beaucoup de clients se sont connectés durant cette période. La Figure 3.5 qui est un graphique en barres illustre la relation entre le nombre de clients connectés et l'énergie totale consommée en moyenne, exprimée en kWh. L'axe des abscisses présente des intervalles de clients connectés, tandis que l'axe des ordonnées indique la consommation énergétique moyenne correspondante.

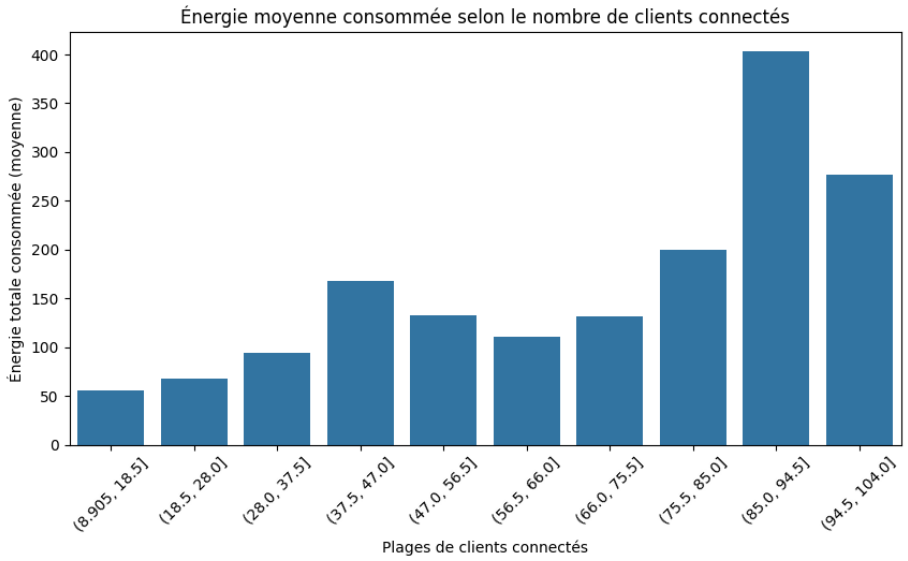

Figure 3.5 : Visualisation de la consommation par rapport aux clients connectés.

La Figure 3.5, montre que la consommation maximale d'énergie atteint 400 kWh, quel que soit le nombre de clients connectés. Cela suggère que ces valeurs aberrantes pourraient résulter d'une erreur de mesure ou d'une défaillance des équipements de mesure. Une simple suppression pourrait aider à les corriger, mais leur quantité étant relativement importante, il est pertinent d'explorer d'autres techniques de gestion sans les éliminer.

Afin de corriger cette anomalie, on applique la winsorisation, qui est une technique statistique pour gérer les valeurs aberrantes. Selon (Nyitrai & Virág, 2019), la winsorisation est une approche courante pour gérer les valeurs aberrantes car elle les remplace par la valeur la plus proche dans une série chronologique donnée. Cette méthode remplace donc les valeurs extrêmes de cet ensemble de données par la plus grande valeur normale le plus enregistrée de la consommation d'énergie.

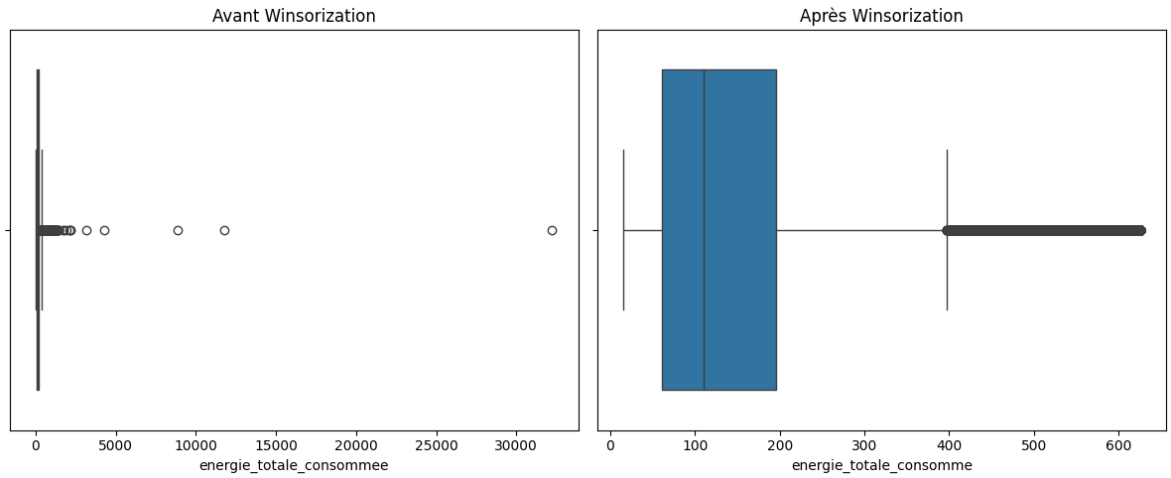

Figure 3.6 : Application de la winsorisation sur les valeurs aberrantes

La Figure 3.6, juxtapose deux boxplots illustrant la distribution de la variable énergie totale consommée en kWh, avant et après l'application de la technique de Winsorization. Sur cette figure, on observe que la winsorization a permis d'ajuster et de stabiliser les valeurs aberrantes dans un intervalle régulé de 0 à 600 kWh, pour limiter leur influence sur la suite de l'analyse. Étant donné que ces valeurs extrêmes peuvent résulter d'un défaut d'enregistrement des données et fausser la prédiction, cette méthode permet de conserver toutes les informations de consommation tout en les ramenant à un niveau normal.

Analyse des corrélations

Toutes les variables explicatives n'ont pas la même influence sur la prédiction de la variable cible, et certaines peuvent s'avérer non pertinentes. Pour cela, l'analyse des corrélations entre les caractéristiques d'entrée et la consommation d'énergie permet d'identifier les variables ayant une relation significative avec cette dernière. Cette technique met en évidence les liens entre les variables et la consommation d'énergie (X. M. Zhang et al., 2018). De plus, ces variables contribuent à construire la meilleure fonction possible pour

l'entraînement des modèles. Dans cette étude, la corrélation est déterminée à l'aide d'une matrice de corrélation et du coefficient correspondant, qui varie de -1 à +1. Celui-ci indique la force et la direction de la relation entre une variable et la consommation d'énergie, permettant d'évaluer son impact. Son application offre une visualisation des connexions entre les variables, comme l'indique la Figure 3.7.

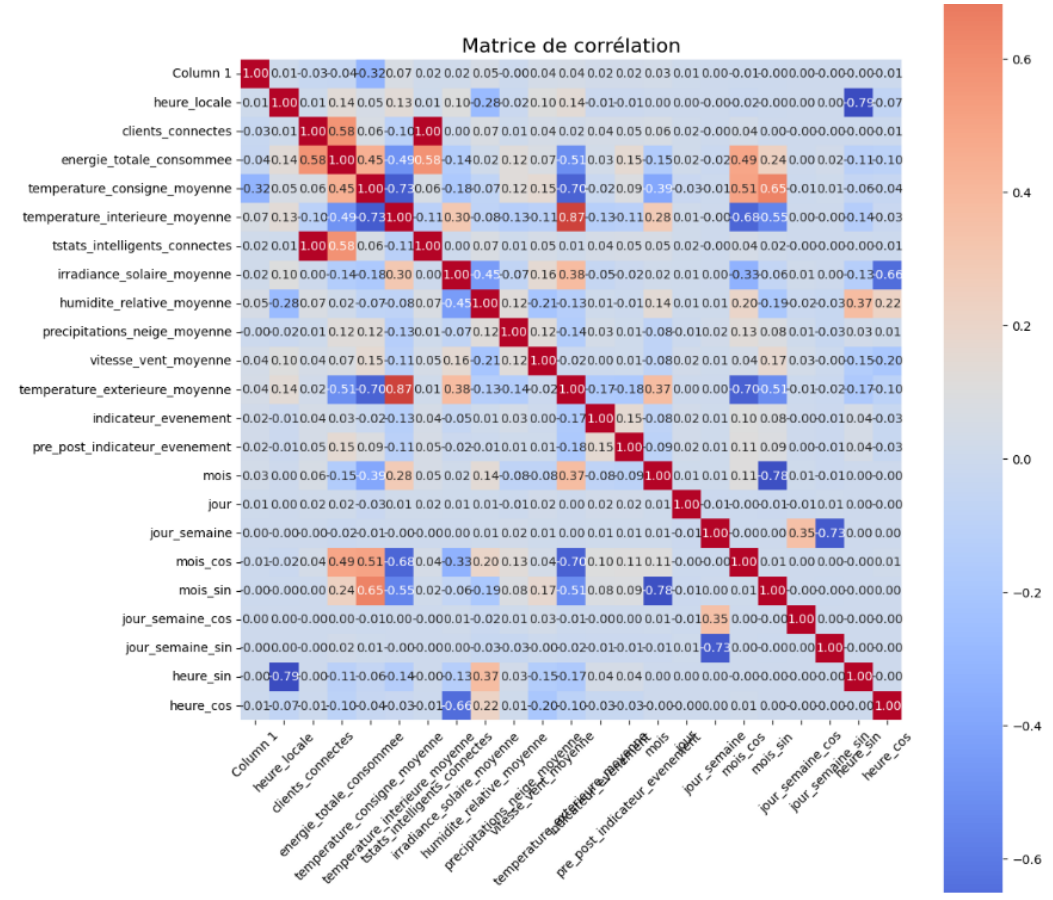

Figure 3.7 : Visualisation les corrélations avec la consommation totale.

Les facteurs qui influencent le plus la consommation totale d'énergie sont principalement le nombre de clients connectés (0,58), suivi par la température enregistrée (0,45) et la température intérieure (0,31), indiquant une importante consommation. Une autre corrélation intéressante concerne la variable "*stats_intelligents_connectées*" (0,56), ce qui

pourrait refléter une utilisation importante de dispositifs intelligents. Les variables temporelles présentent aussi une légère influence. Les autres variables ne montrent pas assez de corrélation. Néanmoins, elles pourraient encore contenir des informations qui pourraient être pertinentes. Une autre analyse qui peut aussi aider dans cette tâche avec plus de fluidité est celle du classement des variables par niveau d'importance.

Pour mesurer l'importance des caractéristiques, on utilise souvent les coefficients de corrélation ou des algorithmes de modélisation prédictive. Parmi ces algorithmes, la forêt aléatoire est l'un des modèles qui permet d'évaluer l'importance des variables de façon automatique (Lovatti et al., 2019). Car elle peut calculer l'impact de chaque variable pour la sélection des meilleures variables. La Figure 3.8 représente un graphique en barres horizontales qui présente l'importance relative des variables. L'axe des abscisses est exprimé en échelle logarithmique de base 5, ce qui permet de mieux visualiser les écarts d'importance entre les variables, même lorsque certaines ont des valeurs très faibles. L''axe des ordonnées présente les différentes variables.

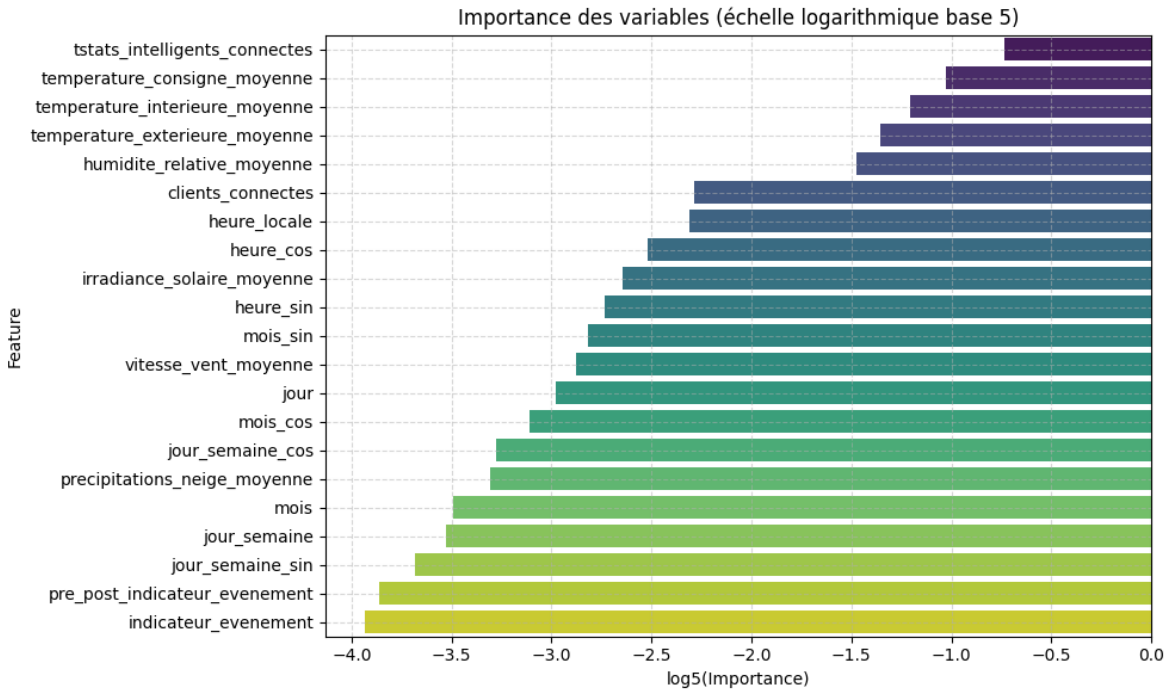

Figure 3.8 : Visualisation des variables selon l'importance.

La visualisation de la Figure 3.8 met en évidence l'influence des variables d'entrée sur celles de sortie. Cette analyse permet d'optimiser le redimensionnement de l'ensemble de données afin d'améliorer la précision des prédictions. Ainsi, les caractéristiques sélectionnées pour l'entraînement des modèles ont été identifiées et classées selon des critères spécifiques, comme présenté dans le Tableau 3.4.

Tableau 3.4 : Variables sélectionnées pour redimensionner l'ensemble de données.

Variables de consommation	Variables météorologiques	Variables temporelles
energie_totale_consommee	temperature_exterieure_moyenne	heure_locale
clients_connectes	irradiance_solaire_moyenne	mois_sin, mois_cos
tstats_intelligents_connectes	humidite_relative_moyenne	heure_sin, heure_cos
temperature_consigne_moyenne	temperature_interieure_moyenne	Jour, mois

L'ensemble de données, désormais nettoyé et préparé, est prêt à répondre aux exigences des prochaines tâches. Son traitement assure une qualité optimale, tant pour les visualisations que pour l'apprentissage des modèles.

3.2.2 VISUALISATION

Une fois les données nettoyées, l'étape suivante consiste à visualiser les schémas de consommation électrique. La visualisation des données aide à mieux comprendre la consommation et à voir comment la puissance électrique évolue selon le temps ou selon les utilisations. Une étude (Herrmann et al., 2018) suggère qu'il est pertinent d'analyser les types d'informations pouvant mieux sensibiliser les consommateurs domestiques, ainsi que les moyens de présentation optimaux afin de maximiser les opportunités d'amélioration et de changement de comportement.

L'objectif principal de cette tâche de visualisation est d'examiner l'évolution temporelle de la consommation électrique à différentes échelles afin de mieux comprendre

les dynamiques temporelles de la consommation énergétique, ce qui permet d'identifier les facteurs influents tels que les saisons, les heures et les jours, et ainsi d'orienter les stratégies d'optimisation. En parallèle, il est essentiel de repérer les périodes critiques de forte consommation afin de mettre en place des mesures adaptées. De plus, associer la consommation à des comportements humains et à des facteurs climatiques permet d'expliquer certaines variations observées. Enfin, détecter des opportunités d'économie d'énergie en ciblant les moments où l'usage peut être optimisé ou réduit constitue une approche efficace pour améliorer l'efficacité énergétique.

Évolution temporelle de la consommation énergétique

La Figure 3.9 est un graphique linéaire représentant l'évolution de la consommation énergétique totale en kWh sur une période allant du début de l'année 2022 jusqu'à la mi-2024. L'axe des abscisses indique les dates, tandis que l'axe des ordonnées mesure la quantité d'énergie consommée. La visualisation de la consommation totale sur toute la période est un graphique de série temporelle qui permet, d'observer les tendances générales, les variations saisonnières ainsi que les anomalies éventuelles, afin de mieux comprendre les dynamiques d'utilisation de l'énergie.

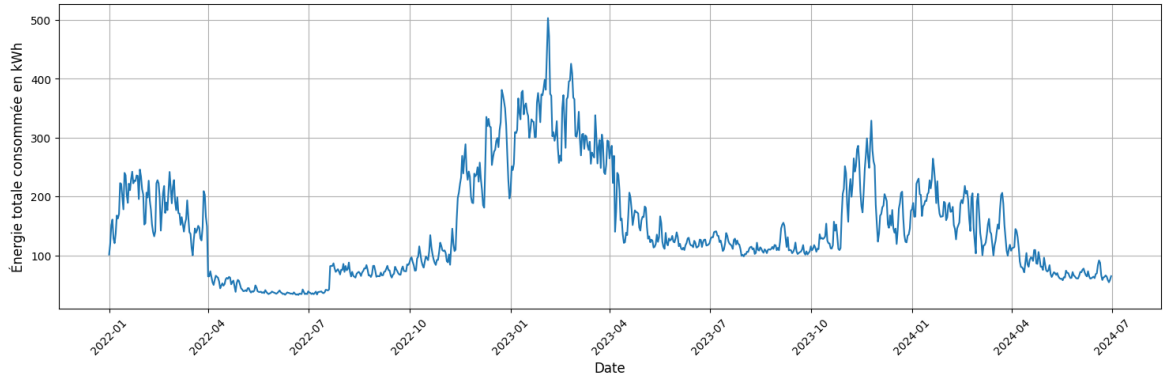

Figure 3.9 : Série temporelle de la consommation totale.

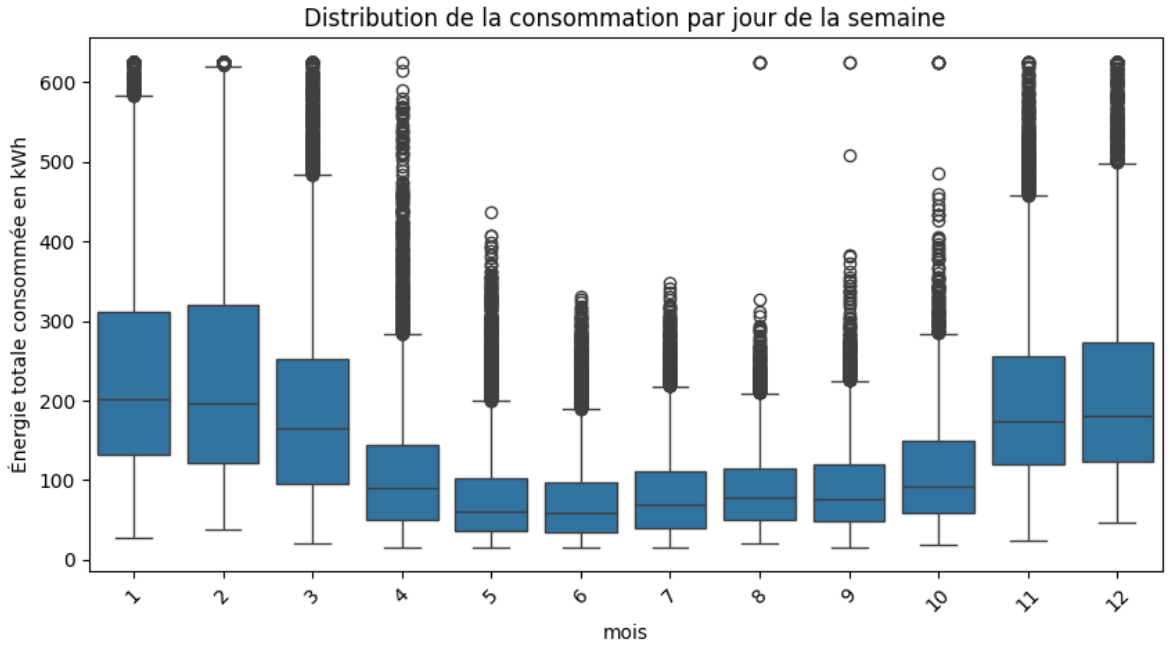

Figure 3.10 : Variation mensuelle de la consommation énergétique totale.

La Figure 3.10 présente des boîtes à boxplots mensuelles de la consommation électrique. Chaque boîte illustre les valeurs minimales, maximales, les quartiles et la médiane de la consommation mensuelle. Elle montre que la consommation électrique connaît une augmentation très remarquable durant les mois d'hiver, en particulier autour de décembre, de janvier, de février et de mars, où le chauffage est constamment utilisé. En revanche, on observe une baisse pendant l'été, notamment de mai à septembre, avec des niveaux parfois très bas. Cette observation démontre une dépendance à l'électricité pour le confort thermique. Ce comportement suit un cycle annuel régulier, affichant clairement une saisonnalité. Il faut aussi noter qu'il y a des pics imprévus à certaines périodes spécifiques. Ces fluctuations sont probablement influencées par le comportement des utilisateurs ou l'utilisation d'équipements particulièrement énergivores. Deux facteurs sont impliqués dans la consommation excessive, les saisons et le comportement des occupants. Pour voir cela de plus près, il est nécessaire de

visualiser la consommation horaire et celle suivant les températures, qui sont des variables très corrélées.

Consommation par heure de la journée

Ce graphique en boxplots illustre la distribution de la consommation énergétique totale en kWh selon l'heure locale, de 0 h à 23 h. Chaque boîte représente la répartition des valeurs de consommation pour une heure donnée, incluant la médiane, les quartiles et les éventuelles valeurs aberrantes. La Figure 3.11 et la Figure 3.12 permettent d'analyser la consommation et les pics électriques moyens journaliers pour voir les heures de pointe et comprendre les comportements des utilisateurs afin de réfléchir à des stratégies d'économie de la consommation. Elles sont aussi utiles pour détecter des tendances spécifiques, identifier des variations entre les différentes heures de la journée et repérer les utilisations abusives ou les anomalies dans la consommation.

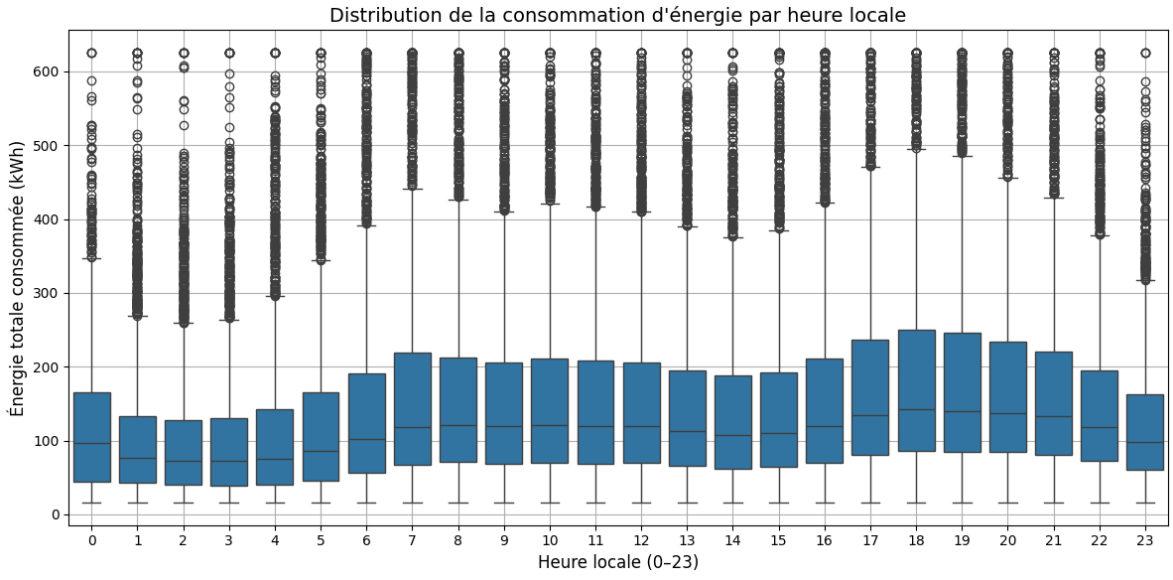

Figure 3.11 : Variation horaire de la consommation énergétique totale.

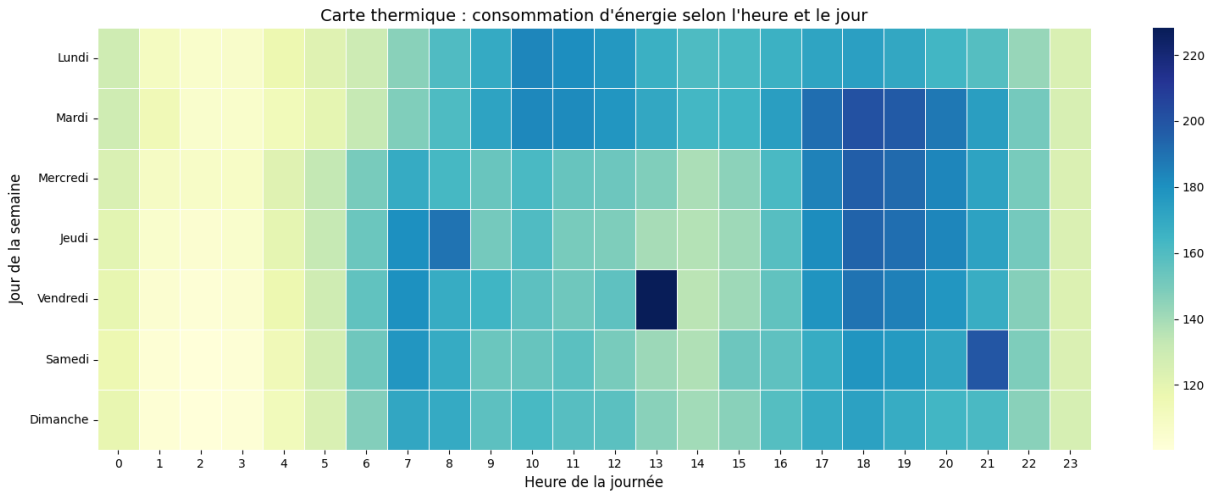

Figure 3.12 : Pics horaire selon les jours de la semaine.

Sur la Figure 3.11, La consommation électrique reste particulièrement faible entre 23 h et 5 h du matin, ce qui indique naturellement les heures de sommeil et une utilisation faible des appareils. À partir de 6 h, on observe une montée progressive de la consommation, en grande partie due aux activités matinales. En soirée, entre 17 h et 20 h, on remarque un pic de consommation très haut, correspondant au moment où la plupart des ménages connectés sont chez eux. La préparation des repas, l'éclairage et l'utilisation d'appareils électroniques comme les téléviseurs participent à cette hausse. Puis, après 22 h, la consommation commence à diminuer à nouveau, ce qui signale la période de sommeil.

Sur la Figure 3.12, la carte thermique signale que les pics précédemment remarqués, entre 17 h et 20 h, sont influencés par deux jours de la semaine, notamment le mardi, le mercredi et le jeudi. Les samedis également à 21 h, un pic est signalé. Ces heures de pointe nécessitent une analyse dans le but de repérer les activités énergivores et de proposer une solution plus économe. Soit déplacer les activités effectuées à des heures libres de consommation ou encore l'utilisation de l'énergie renouvelable pendant ces heures de pointe.

Une autre remarque qui attire l'attention est la journée du vendredi à 13 h, qui présente un pic presque inhabituel. Cette anomalie peut être due à un dysfonctionnement ou à une mauvaise habitude d'utilisation. Une analyse plus approfondie auprès des utilisateurs peut aider à mieux comprendre afin de la régulariser.

Ces analyses révèlent que le comportement des utilisateurs est bel et bien un facteur à améliorer pour optimiser la consommation ainsi qu'une maintenance curative des systèmes de consommation et de collecte.

Consommation en fonction des températures

La Figure 3.13 utilise trois axes pour illustrer la consommation énergétique en fonction des températures intérieures et extérieures. L'axe X, situé horizontalement à gauche, représente la température intérieure moyenne en degrés Celsius, qui varie entre 16 °C et 26 °C, des niveaux pour le confort thermique. L'axe Y, horizontal à droite, montre la température extérieure moyenne, allant de -30 °C à +35 °C, ce qui couvre les variations saisonnières de l'année, du froid intense en hiver à la chaleur. L'axe Z, vertical, mesure la consommation énergétique totale en kilowattheures. La couleur des points sur le graphique signifie que les zones jaunes indiquent une consommation élevée, tandis que les zones violettes représentent une consommation faible. Cette visualisation montre des variations de consommation en fonction des écarts thermiques. Cette combinaison d'axes et de couleurs aide à repérer les tendances ou anomalies dans la consommation thermique.

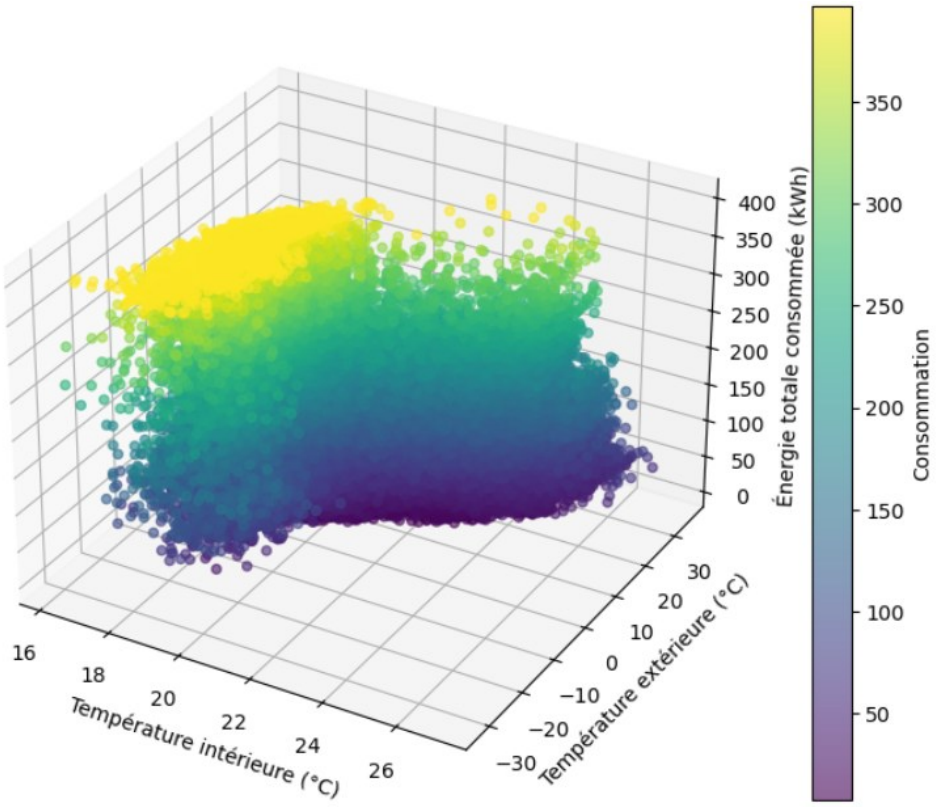

Figure 3.13 : Visualisation 3D de la consommation en fonction des températures.

On observe sur la Figure 3.13, que lorsque la température extérieure est très basse, notamment en dessous de 0 °C, la consommation d'énergie augmente considérablement. Cela s'explique facilement par une utilisation excessive du chauffage pour maintenir l'intérieur des ménages au chaud. Concernant la température intérieure, dès qu'elle est basse dans la plage des 16–18 °C, la consommation est un peu plus élevée, surtout si les conditions extérieures sont également froides. Cela est lié au fait que le système de chauffage consomme davantage pour maintenir une température minimale constante face à des écarts thermiques importants.

En outre, lorsque la température extérieure est modérée, autour de 10–20 °C, la consommation énergétique reste nettement plus faible, et ce, indépendamment de la température intérieure. Cela reflète une période où les besoins en chauffage ou en climatisation sont peu importants.

En gros, on retient que la consommation énergétique domestique est principalement influencé par la température extérieure, avec des pics marqués durant l'hiver, période où le chauffage est fortement sollicité. La température intérieure, bien que secondaire, agit en complément, et c'est avant tout l'écart thermique entre l'intérieur et l'extérieur qui joue un rôle déterminant dans les fluctuations de consommation. Par ailleurs, le comportement des occupants apparaît comme un facteur clé, notamment à travers la fréquence et les horaires d'utilisation des équipements électriques, qui influencent directement la consommation globale. Les pics observés en soirée illustrent cette corrélation entre présence humaine et intensification de l'usage des appareils électriques. À l'inverse, la période estivale ne montre pas de variations significatives, ce qui laisse penser que l'usage de la climatisation reste relativement modéré dans les ménages étudiés. Ces résultats offrent des perspectives concrètes pour la suite de l'étude, notamment dans l'élaboration de stratégies d'optimisation énergétique. Ils ouvrent la voie à des solutions durables visant à réduire la consommation, en mettant l'accent sur une meilleure régulation thermique et sur la sensibilisation aux usages quotidiens, dans une approche globale de gestion intelligente et responsable de l'énergie domestique.

3.3 PRÉDICTIONS

La prédiction étant une étape très importante de cette recherche, elle est expérimentée d'une manière méthodique pour avoir de meilleurs résultats. Ces résultats sont indispensables pour atteindre l'objectif de cette recherche et permettre également l'utilisation de cette méthodologie dans d'autres études qui visent à optimiser la consommation d'électricité grâce à la prédiction. Pour cela, dans un premier temps, plusieurs modèles sont entraînés pour faire une prédiction sur l'ensemble de données traité et une évaluation de leur performance a permis de présenter les résultats primaires. Ensuite, une attention est portée à l'ingénierie des caractéristiques. Enfin, une optimisation des modèles avec l'ajustement des hyperparamètres est faite grâce à diverses techniques, puis les nouvelles performances ont permis de comparer l'ensemble des résultats et de tirer une conclusion sur les modèles de prédiction les mieux adaptés à ce type d'ensemble de données de consommation domestique.

3.3.1 MODELISATION ET ÉVALUATION DES ALGORITHMES

Pour entraîner et évaluer les modèles de prédiction, les données prétraitées ont été divisées selon deux méthodes, choisies en fonction du type de modèle et de leur complexité de calcul. D'une part, la méthode de séparation classique a été utilisée pour les modèles d'apprentissage profond. Elle a permis de répartir l'ensemble des données en deux sous-ensembles de 80 % pour l'entraînement et 20 % pour le test. Cette approche simple et rapide est adaptée aux modèles nécessitant une grande puissance de calcul. D'autre part, pour les modèles d'apprentissage automatique, la validation croisée par k-fold a été privilégiée. Dans cette méthode, les données sont divisées aléatoirement en k sous-ensembles.

Modélisations

Considérant l'étude de la littérature, plusieurs algorithmes, allant des plus simples aux plus avancés, ont été sélectionnés pour cette tâche en raison de leurs performances distinctes. Notamment, la régression linéaire, elle établit et ajuste une équation linéaire aux données (Fumo & Rafe Biswas, 2015). C'est un modèle basique qui est très souvent exploré dans les tâches de prédiction, puisqu'il offre la simplicité d'utilisation et d'interprétabilité. Elle repose sur l'hypothèse qu'il existe une relation linéaire entre la variable cible et les variables explicatives (Lin et al., 2022). Le modèle ajuste l'équation (3.1) :

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta pxp + \epsilon \qquad \textbf{(3.1)}$$

où y est la variable cible donc l'énergie totale consommée, xi représentent les variables explicatives du modèle, βi les coefficients estimés associés à chaque variable explicative, et ϵ l'erreur résiduelle.

L'équation (3.1) représente le modèle de régression linéaire multiple, c'est-à-dire utilisant plusieurs variables explicatives, dans lequel la variable cible qui est l'énergie totale consommée est exprimée comme une combinaison linéaire des variables explicatives xi, pondérées par leurs coefficients respectifs βi. Le terme $\beta 0$ correspond à l'ordonnée à l'origine, avec ε désigne l'erreur résiduelle, c'est-à-dire la part de variation de y non expliquée par le modèle. L'objectif de la régression est d'estimer les coefficients βi de manière à minimiser l'écart global entre les valeurs observées et les valeurs prédites.

Le code du modèle a été implémenté à l'aide de la classe LinearRegression de la bibliothèque scikit-learn. Les paramètres par défaut ont été conservés afin de garantir une

configuration standard du modèle. Plus précisément, les paramètres comme fit_intercept=True, copy_X=True, n_jobs=None et positive=False sont utilisés.

Ensuite, la forêt aléatoire constitue une méthode d'apprentissage par ensemble reposant sur la construction de multiples arbres de décision à partir de sous-échantillons aléatoires des données (Pham et al., 2020). Cette approche vise à réduire la variance du modèle sans accroître le biais, en agrégeant les prédictions issues de chaque arbre. Ce mécanisme d'agrégation permet de limiter le surapprentissage et d'améliorer la robustesse globale des prédictions.

Pour une tâche de régression, comme celle-ci, le résultat final est la moyenne des prédictions de tous les arbres. La structure mathématique de ce modèle est représentée par l'équation (3.2).

$$\hat{y} = \frac{1}{N_{arbres}} \sum_{i=1}^{N_{arbres}} Ti(X) \qquad \textbf{(3.2)}$$

Où $\hat{y}$ est la valeur prédite de la consommation d'énergie totale, $Ti\ (X)$ représente la prédiction de l'arbre i pour les variables explicatives $X = (x1, x2, \dots, xp)$, et Narbres le nombre total d'arbres dans la forêt. L'objectif de la forêt aléatoire est de combiner les prédictions de tous les arbres pour obtenir une estimation plus précise de la variable cible qui est l'énergie totale consommée. Elle permet également de gérer des variables de types différents et de capturer des relations complexes entre elles.

Le code du modèle a été implémenté à l'aide de la classe RandomForestRegressor de la bibliothèque scikit-learn. Les paramètres retenus pour l'implémentation sont, le nombre d'arbres dans la forêt qui est n_estimators est affecté à 100 et pour garantir la reproductibilité

des résultats, le paramètre random_state est affecté à 42. Les autres paramètres par défaut ont été conservés afin de garantir une configuration standard du modèle. La standardisation des variables a été réalisée à l'aide de StandardScaler avant l'apprentissage, pour optimiser la convergence et la stabilité du modèle.

Puis le XGBoost, il s'agit d'un algorithme qui construit un ensemble d'arbres de décision de manière séquentielle en corrigeant à chaque itération les erreurs commises par les arbres précédents (El Houda et al., 2022). C'est un modèle performant et rapide, grâce à l'optimisation de la fonction de perte et sa régularisation intégrée qui limitent le surapprentissage. L'équation (3.3) généralise bien le modèle.

$$\hat{y}i = \sum_{k=1}^{k} fk\,(xi)\,, fk \in F \qquad \textbf{(3.3)}$$

Où $\hat{y}i$ est la valeur prédite pour l'observation i, et chaque fk correspond à un arbre de décision appartenant à l'ensemble F des fonctions possibles. Le modèle optimise une fonction objective composée d'une erreur de prédiction et d'un terme de régularisation pour contrôler la complexité.

Le code du modèle a été implémenté à l'aide de la classe XGBRegressor de la bibliothèque XGBoost. Les principaux hyperparamètres spécifiés incluent n_estimators fixé à 100, learning_rate égale à 0.1, et random_state défini à 42 afin d'assurer la reproductibilité des résultats. Les autres paramètres, tels que max_depth, subsample et ceux non explicitement mentionnés, ont été conservés à leurs valeurs par défaut. Une standardisation préalable des données a été réalisée pour homogénéiser les échelles des variables explicatives.

Quant au CatBoost, il utilise des approches du gradient boosting, qui, selon les études précédemment vues, pourraient se montrer plus décisives pour la prédiction. La méthode de renforcement de gradient utilise un ensemble de modèles faibles qui, collectivement, forment un modèle plus fort. Comme XGBoost, il repose sur l'agrégation d'arbres de décision construits séquentiellement, mais il se distingue par son schéma d'ordonnancement aléatoire et ses techniques de régularisation innovantes qui améliorent la généralisation (Olu-Ajayi et al., 2022). L'équation (3.4) généralise bien ce modèle **(3.4)**

$$\hat{y} = \sum_{t=1}^{T} \eta \cdot h_t(x)$$

où $h_t(x)$ représente l'arbre ajouté à l'itération t, et η le taux d'apprentissage. CatBoost optimise directement une fonction de perte également différentiable en ajustant progressivement les prédictions.

Le modèle a été implémenté à l'aide de la classe CatBoostRegressor. Les principaux paramètres spécifiés sont, verbose égale à 0, afin de masquer les sorties durant l'entraînement, et random_state définit à 42, utilisé pour assurer la reproductibilité des résultats. Les autres paramètres, tels que iterations, learning_rate et depth, ont été conservés à leurs valeurs par défaut, garantissant ainsi une configuration standard du modèle. Contrairement à XGBoost, CatBoost prend en charge nativement les variables catégorielles ; toutefois, dans le cadre de cette étude, les données étaient déjà numériques et préalablement standardisées.

Enfin, les modèles RNN et LSTM, appartenant à la famille de l'apprentissage profond, ont été testés pour leur capacité à prédire les séries temporelles. Leur architecture leur permet

de capter des relations complexes dans les données temporelles. Pour cette étude, les séquences d'entrée correspondent à des fenêtres glissantes de 24 observations horaires. Dans le cas du RNN, un modèle SimpleRNN de 64 neurones avec la fonction d'activation tanh a été implémenté, suivi d'une couche Dense produisant la prédiction finale. Le modèle a été entraîné avec l'optimiseur Adam, une fonction de perte MSE, un nombre d'époques fixé à 10 et une taille de lot de 64. Concernant le LSTM, la même structure d'entrée a été utilisée, mais la couche récurrente est remplacée par une couche LSTM de 64 neurones, permettant de mieux modéliser les dépendances de long terme grâce à ses mécanismes de mémoire interne.

Évaluation de performance

L'évaluation est le processus qui consiste à mesurer la performance d'un modèle. Pour son application, les métriques telle que MSE, RMSE, MAE, MAPE, CV-RMSE, et R^2 servent de mesure des performances réalisées par les algorithmes.

Dans cette recherche, la validation croisée k-fold a été appliquée avec k=10, une valeur qui a démontré une meilleure fiabilité lors des essais. L'ensemble des données a été divisé en dix groupes de taille égale. À chaque itération, neuf groupes ont été utilisés pour l'entraînement du modèle, tandis que le groupe restant a servi à son évaluation. Ce processus a été répété dix fois, en alternant le groupe de validation, afin que chaque groupe soit utilisé une fois comme jeu de test. La moyenne des performances obtenues a ensuite été calculée pour obtenir une estimation plus fiable de la capacité de généralisation des modèles.

La moyenne des carrés des erreurs (MSE) est une métrique couramment utilisée pour mesurer la différence quadratique moyenne entre les valeurs prédites et les valeurs réelles

(Mathumitha et al., 2024). Dans une tâche de prédiction où l'on veut minimiser les grandes erreurs, le MSE est un meilleur choix. Car l'erreur est élevée au carré, ce qui fait que les grandes erreurs sont amplifiées et les écarts entre la prédiction et la valeur réelle sont beaucoup plus remarquables sur la valeur finale du MSE. Elle est calculée à l'aide de l'équation (3.5).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2 \quad \text{où,} \qquad \textbf{(3.5)}$$

n est le nombre d'observations,

yi est la valeur réelle pour l'observation i,

$\hat{y}i$ est la valeur prédite par le modèle pour l'observation i.

La grande limite du MSE est sa sensibilité aux valeurs extrêmes des données, ce qui peut fausser l'évaluation ou l'optimisation des modèles.

C'est en ce moment qu'intervient l'erreur absolue moyenne (MAE) qui donne directement une idée de la qualité des prédictions de façon simple à comprendre. Par exemple, pour la prédiction de la consommation électrique en kWh, si le MAE dans est égal à 10, alors le modèle fait une erreur de 10 kWh sur chaque prédiction. Cela donne une indication claire de la performance du modèle qui prédit 10 kWh près de la valeur réelle. Selon (Mathumitha et al., 2024b), l'analyse du MAE calcule l'erreur absolue moyenne, c'est-à-dire la différence entre les valeurs réelles et prédites, sans tenir compte du signe, ce qui réduit l'impact des grandes erreurs. Il est la métrique qui traite toutes les erreurs de la même manière, d'où il est plus performant lorsque les données contiennent beaucoup de valeurs aberrantes.

Plus la valeur du MAE est faible, plus le modèle se rapproche de la perfection. Cette caractéristique en fait un indicateur essentiel dans les processus de comparaison et d'optimisation des modèles prédictifs. Sa forme mathématique est donnée par la formule (3.6).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|yi - \hat{y}i| \text{ où,} \qquad \textbf{(3.6)}$$

n est le nombre d'observations,

$\boldsymbol{yi}$ est la valeur réelle pour l'observation i,

$\boldsymbol{\hat{y}i}$ est la valeur prédite par le modèle pour l'observation i.

Après l'analyse de l'erreur absolue moyenne (MAE), il est aussi important d'évaluer dans quelle mesure le modèle explique la variabilité des données réelles.

Il s'agit à ce niveau du coefficient de détermination R^2 qui donne la proportion des variations des données que le modèle a pu capturer. Par exemple, un modèle évalué avec un R^2 de 0,85 montre que 85 % des variations des données étaient capturées par le modèle, et donc 15 % des variations le restent à l'œuvre d'autres facteurs. Dans une étude de (Mathumitha et al., 2024), R^2 est une mesure de la qualité de l'ajustement d'un modèle qui compare l'erreur du modèle à la variance totale des données réelles. Plus la valeur de R^2 est proche de 1, plus le modèle est performant, car il capture bien les variations des données. Il se traduit par la formule (3.7).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi-\hat{y}i)^2}{\sum_{i=1}^{n}(yi-\bar{y}i)^2} \qquad \text{où :} \qquad \textbf{(3.7)}$$

$\boldsymbol{n}$: nombre d'observations,

$\boldsymbol{yi}$: valeur réelle de l'observation *i*,

$\boldsymbol{\bar{y}i}$: moyenne des valeurs réelles,

$\boldsymbol{\hat{y}i}$: valeur prédite par le modèle.

Ces métriques donnent assez d'information sur la performance d'un modèle, ce qui permet, dans le cadre de ce mémoire, d'évaluer et de présenter les performances des modèles expérimentés.

Résultat des prédictions

Tableau 3.5 : Performances des modèles entraînés non optimisés

MODÈLES	MSE	MAE	RMSE	R2
Régression linéaire	5068.80	51.65	71.19	0.67
Forêt aléatoire	501.81	12.83	22.40	0.97
XGBoost	680.98	16.71	26.09	0.96
CatBoost	504.27	14.25	22.45	0.97
RNN	3199.61	40.40	56.57	0.79
LSTM	2422.43	32.43	49.21	0.84

Les résultats expérimentaux du Tableau 3.3 montre que, les modèles d'apprentissage automatique basés sur l'approche ensembliste (forêt aléatoire, CatBoost, XGBoost) dépassent

de loin le modèle de régression linéaire et les modèles d'apprentissage profond (RNN, LSTM). Dans le cas présent, le meilleur modèle est la forêt aléatoire, avec les valeurs les plus faibles de MSE, MAE et RMSE, et un R2 élevé (0,97). Ce qui signifie qu'il a une très bonne capacité d'apprendre de nos données sans être optimisé. Toutefois, lorsqu'on prête attention aux modèles d'apprentissage profond, le LSTM donne une bonne performance et dépasse le RNN. Il serait donc intéressant da faire une optimisation des modèles en appliquant l'ajustement d'hyperparamètres et l'ingénierie des caractéristiques, afin d'explorer le plein potentiel des modèles.

3.3.2 INGÉNIERIE DES CARACTERISTIQUES

La première phase d'optimisation des modèles a consisté à appliquer différentes techniques d'ingénierie des caractéristiques afin d'améliorer leur performance prédictive. Deux approches ont été testées : l'Analyse en Composantes Principales (ACP) et la création manuelle de nouvelles variables.

Application de l'ACP

L'ACP a été appliquée en retenant les trois premières composantes principales, qui expliquent 92 % de la variance totale des données. Après intégration dans les algorithmes de prédiction, les résultats indiquent que l'impact de l'ACP est limité. Par exemple, le coefficient de détermination R^2 de la régression linéaire est passé de 0,67 à 0,71. Pour les autres modèles, les valeurs de R^2 sont restées stables, avec uniquement une légère variation du MAE (kWh). Cette étape, bien que rapide en temps d'exécution, n'a pas permis d'obtenir des gains de performance significatifs.

Création manuelle de nouvelles variables

Une deuxième approche a permis d'enrichir l'ensemble de données en créant des variables supplémentaires de type temporel, physique et statistique. Les variables temporelles incluent la distinction entre jours ouvrables et fins de semaine, le numéro de semaine pour capter d'éventuels effets saisonniers, ainsi que l'encodage des saisons sous forme de variables catégorielles. Les variables physiques comprennent la température ressentie et l'écart à la température de consigne. Enfin, les variables statistiques intègrent la moyenne mobile et l'écart-type mobile de la consommation énergétique en kWh, calculés sur une fenêtre glissante de 7 jours afin de capturer les tendances et la variabilité à court terme.

Parmi celles-ci figurent la nouvelle variable, interaction_temp, définie comme le produit entre la température intérieure et la consigne ; la diff_temp, représentant l'écart entre la consigne et la température extérieure en degré ; l'humidity_irradiance_ratio, calculée comme le rapport entre l'humidité et l'irradiance augmentée d'une unité ; et la temperature_variation, correspondant à la différence entre la température intérieure et extérieure. Le prétraitement des données a inclus le remplissage des valeurs manquantes ainsi que la standardisation via StandardScaler. Les paramètres du modèle sont restés identiques à ceux de la version simple, avec les valeurs par défaut et random_state égale à 42. Les résultats ont montré une amélioration notable des métriques, en particulier du MAE et du R^2, ce qui indique que l'ajout de ces variables a permis au modèle de mieux capturer les comportements complexes et les interactions entre variables.

En résumé, après avoir testé séparément les deux approches, il apparaît que, pour ce type de jeu de données, l'ACP est très rapide mais n'apporte presque aucun bénéfice à la

performance des modèles. En revanche, la création manuelle de nouvelles variables, intégrée dans les algorithmes, a montré une nette amélioration, bien que son exécution soit significativement plus lente. Étant donné que l'objectif de cette étape est d'optimiser la précision des modèles, l'approche offrant les meilleurs résultats en termes de performance, comme l'indique le Tableau 3.6, est retenue.

Tableau 3.6 : Performances des modèles entraîné avec l'ingénierie des caractéristiques

MODÈLES	MSE	MAE	RMSE	R2
Régression linéaire	4445.75	49.06	66.67	0.71
Forêt aléatoire	489.78	12.70	22.13	0.97
XGBoost	652.76	16.39	25.54	0.96
CatBoost	495.81	14.13	22.26	0.97
RNN	1994.80	30.90	44.66	0.87
LSTM	1610.58	26.62	40.13	0.89

L'ingénierie des caractéristiques a conduit à une amélioration des performances des modèles. La régression linéaire montre une légère progression, mais demeure limitée face à la complexité des données. Les modèles d'arbres de décision, notamment la forêt aléatoire et CatBoost, conservent leur supériorité avec une précision intéressante. XGBoost reste toujours rapide et aussi performant. Les réseaux neuronaux montrent un léger changement intéressant, mais ils affichent toujours des résultats inférieurs aux modèles d'arbres de

décision. Ces observations renforcent la pertinence des modèles d'arbres. Dans le but d'améliorer les modèles DL et de rendre l'entraînement plus rapide que celui avec l'ingénierie des caractéristiques, l'optimisation des hyperparamètres pourrait aider à affiner davantage les résultats.

3.3.3 AJUSTEMENT DES HYPERPARAMETRES

Dans le contexte de cette étude, l'optimisation des hyperparamètres a été mise en œuvre afin d'optimiser l'efficacité des modèles prédictifs. Trois méthodes distinctes ont été explorées en fonction des spécificités de chaque modèle.

Recherche en grille

La recherche en grille a été appliquée aux modèles classiques tels que CatBoost, XGBoost, la forêt aléatoire et la régression linéaire, dans le but de minimiser l'erreur quadratique moyenne (RMSE). Toutefois, elle s'est révélée particulièrement coûteuse en temps de calcul, en raison du nombre élevé d'itérations nécessaires pour couvrir l'espace des paramètres.

Optimisation bayésienne

Dans le cadre de cette recherche, l'optimisation bayésienne a été appliquée principalement aux modèles classiques. Son objectif était de trouver automatiquement les combinaisons d'hyperparamètres qui maximisent la performance du modèle mesurée par le coefficient de détermination R^2.

Le principe pratique appliqué repose sur la définition d'un espace de recherche pour les hyperparamètres critiques des modèles. Par exemple pour CatBoost, les paramètres

explorés incluaient iterations, learning_rate, depth, l2_leaf_reg et border_count, tandis que pour XGBoost, il s'agissait de n_estimators, learning_rate, max_depth, reg_alpha et reg_lambda. La fonction objective utilisée était la moyenne du coefficient de détermination R^2 obtenue par validation croisée 3-fold sur le jeu d'entraînement. Afin de permettre à gp_minimize de maximiser cette métrique, la valeur négative du R^2 était retournée. La recherche s'effectuait de manière itérative, à chaque étape, le modèle était entraîné avec une combinaison d'hyperparamètres proposée par le processus d'optimisation bayésienne, puis évalué. Après un nombre défini d'itérations par exemple n_calls = 30 pour CatBoost, le modèle retenu correspondait à celui ayant obtenu le meilleur R^2 moyen. Sur le plan pratique, cette approche a permis de réduire significativement le nombre d'évaluations nécessaires par rapport à une recherche en grille exhaustive. Les performances du modèle se sont améliorées, avec une augmentation du R^2 et une réduction du MAE et du RMSE. Bien que cette technique n'ait pas permis d'obtenir des résultats convaincants avec les modèles d'apprentissage profond, une autre approche a donc été expérimentée spécifiquement pour leur cas.

Hyperbande

Pour les modèles RNN et LSTM, la technique Hyperband a été utilisée pour optimiser les hyperparamètres rapidement et efficacement. Son application commence par l'évaluation d'un grand nombre de combinaisons d'hyperparamètres générées aléatoirement, avec un nombre restreint d'arbres ; dans notre cas, on a commencé avec *n_estimators* égale à 50. À chaque itération, seules les combinaisons les plus prometteuses sont conservées, tandis que les autres sont progressivement écartées. Simultanément, le nombre d'arbres alloué est doublé à chaque étape. Finalement, la meilleure configuration retenue est utilisée pour

entraîner un modèle final avec un nombre d'arbres plus élevé, *n_estimators* égale à 200, pour s'assurer de sa performance et sa rapidité.

L'expérimentation de cette technique d'ajustement des hyperparamètres a permis d'obtenir ces nouvelles performances optimisées comme le montre le Tableau 3.7.

Tableau 3.7 : Performances des modèles optimisés

MODÈLES	MSE	MAE	RMSE	R2
Régression linéaire	4453.16	48.82	66.73	0.71
Forêt aléatoire	554.37	13.64	23.55	0.96
XGBoost	386.85	11.47	19.66	0.97
CatBoost	373.78	11.18	19.33	0.98
RNN	2611.71	35.67	51.10	0.83
LSTM	1864.64	28.84	43.18	0.88

Dans cette dernière phase, où les modèles ont été optimisés automatiquement, les résultats montrent que certains modèles répondent très bien à cette approche. C'est le cas de XGBoost et surtout de CatBoost, qui atteint les meilleures performances globales (R2 = 0,98 et MAE faible). L'optimisation leur permet de mieux exploiter la structure des données. À l'inverse, des modèles RNN et LSTM qui réagissent moins bien à cette phase. Leurs performances diminuent légèrement par rapport à celles obtenues par l'ingénierie des caractéristiques, probablement à cause de la sensibilité au sur-ajustement lié aux réglages

automatiques. Pour les modèles de régression linéaire et la forêt aléatoire, les performances sont limitées. On retient que l'efficacité, l'optimisation dépendent beaucoup de la nature du modèle et de sa sensibilité aux paramètres internes.

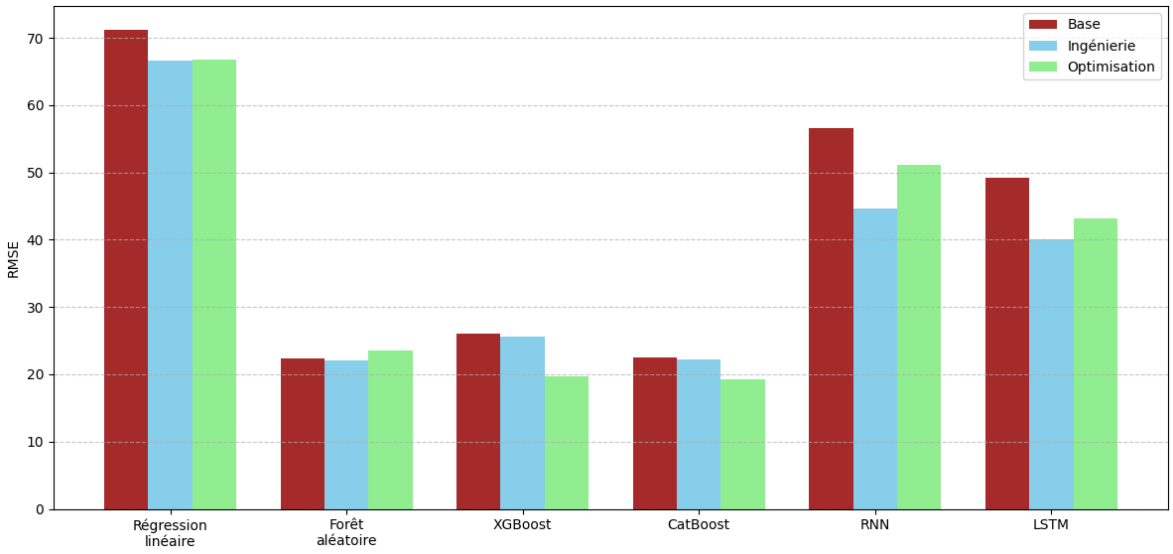

Figure 3.14 : Comparaison des performances des modèles à chaque étape.

En général, les résultats obtenus à travers les trois phases de prédiction, comme le montre la Figure 3.14, mettent en évidence plusieurs des informations clés, en cohérence avec les littératures récentes sur l'optimisation des modèles d'apprentissage. Premièrement, la phase de prédiction de base montre que les modèles d'ensemble comme la forêt aléatoire, XGBoost et CatBoost offrent d'excellents résultats dès leurs configurations par défaut. contrairement aux modèles simples (régression linéaire) et aux modèles d'apprentissage profond non optimisés. Deuxièmement, l'ajout de variables manuellement créées permet d'améliorer les performances de tous les modèles, en particulier des réseaux neuronaux RNN et LSTM. Cette étape montre que le RNN et LSTM captent mieux les effets temporels et saisonniers dans les données. Cela confirme l'information selon laquelle l'importance d'une

bonne représentation des entrées pour la qualité des prédictions est pertinente. Enfin, l'optimisation des hyperparamètres a permis d'exploiter toute la performance des modèles, en particulier le XGBoost et le CatBoost, qui ont enregistré les meilleures performances finales avec un RMSE et un MAE très bas. Cela confirme la sensibilité de ces modèles à la puissance des techniques d'optimisation, comme l'optimisation bayésienne. En revanche, les modèles d'apprentissage profond tel que le RNN et le LSTM, bien qu'améliorés par l'ingénierie, ont eu leurs performances légèrement diminuer après optimisation, ce qui suggère que des ajustements plus spécifiques ou des ressources plus importantes seraient nécessaires pour stabiliser leur entraînement.

3.4 EXPLICABILITÉ DU MODÈLE CHOISI

Le modèle CatBoost a démontré tout au long de la prédiction une belle performance, ce qui amène à comprendre comment il a appris avec les données. L'idée de l'explicabilité est d'interpréter l'apprentissage du modèle choisi, d'identifier les relations entre les entrées et la sortie. Cette technique, déjà décrite dans le chapitre 2, à la base des études antérieures, a montré ses forces et ses faiblesses tant avec les modèles d'apprentissage automatique qu'avec les apprentissages profonds. L'algorithme du SHAP est appliqué pour interpréter les résultats de ce modèle considéré comme plus performant pour la prédiction. À chaque caractéristique de l'échantillon prédit, une valeur SHAP est attribuée, reflétant à la fois les influences positives et négatives. Pour évaluer leur importance, on calcule la moyenne des valeurs SHAP absolues de chaque caractéristique, puis on les classe en ordre décroissant afin de produire un graphique statistique illustrant une hiérarchie, comme le montre la figure 3.17. L'analyse SHAP montre que la caractéristique (*tstats_intelligents_connectés*) est le facteur le

plus important contribuant à la consommation d'énergie, suivie par la température consignée, le nombre de clients connectés et la température intérieure.

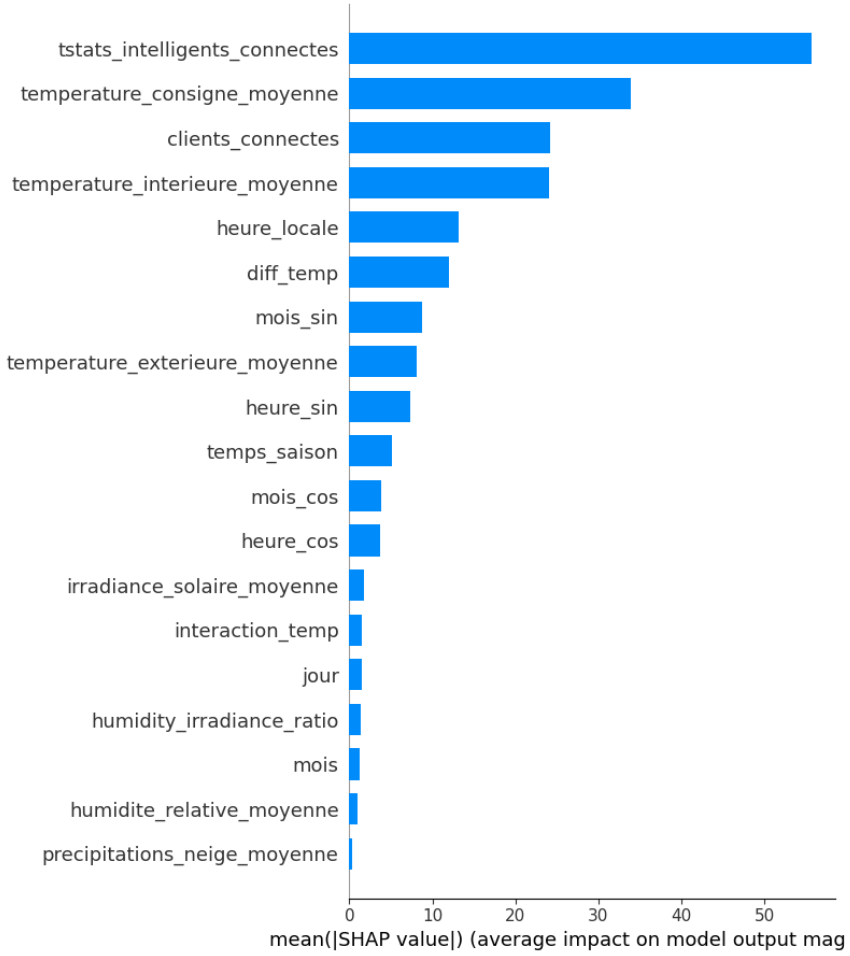

Figure 3.15 : Résultats de l'analyse SHAP du CatBoost.

L'analyse explicative par SHAP de la figure 3.15, explique de manière claire comment le modèle a pu à une consommation de 162,08 kWh pour une situation précise. D'abord, il y a des facteurs qui ont augmenté significativement la prédiction, notamment "la feature 2", qui est le nombre de thermostats intelligents connectés. Il ajoute 27,74 kWh à la prédiction. Cette régulation thermique implique une utilisation importante des systèmes de chauffage. De même, "la feature 0" qui est le nombre de clients connectés, ajoute également +11,95 kWh,

ce qui est cohérent avec un contexte de forte occupation, donc plus d'utilisation des appareils électriques. Ensuite, "la feature 14", qui représente la température de consigne, diminue de –17,7 kWh la prédiction et "la feature 1", la température intérieure moyenne, diminue également de –6,13 kWh. Cela implique une faible utilisation du chauffage parce que la température intérieure est déjà assez chaude réduisant les besoins en énergie. Cette analyse montre que le modèle réagit bien aux différentes conditions, en reproduisant un comportement énergétique réel. SHAP prouve donc que la prédiction finale n'est pas arbitraire, mais résulte d'un raisonnement logique basé sur des contributions et l'aspect des variables.

Tableau 3.8 : Variables explicatives SHAP

Feature	Valeur approx.	Variable réelle
Feature 2	+27.74	tstats_intelligents_connectes
Feature 14	-17.7	temperature_consigne_moyenne
Feature 0	+11.95	clients_connectes
Feature 11	-7.58	mois_cos
Feature 7	+6.98	heure_locale
Feature 1	-6.13	temperature_interieure_moyenne
Feature 13	+5.61	heure_cos
Feature 9	+4.18	jour
Feature 18	-3.16	temperature_variation

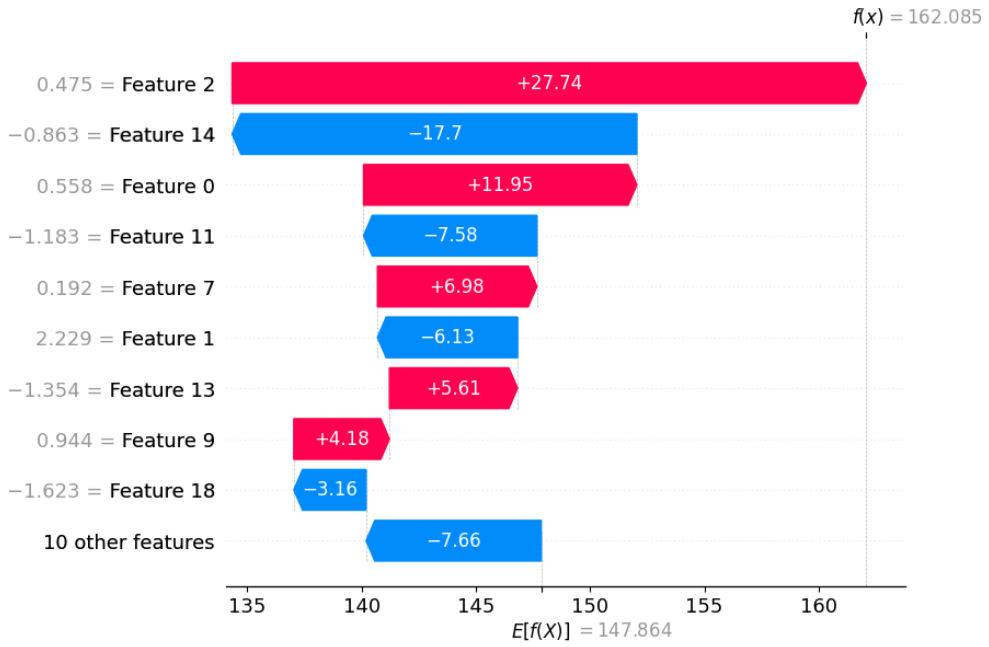

Figure 3.16 : Résumé de SHAP en cascade du CatBoost.

La figure 3.16 représente un graphique récapitulatif SHAP appliqué au modèle CatBoost. Cette visualisation montre l'impact général de chaque variable sur la prédiction globale. Les variables sont classées du haut vers le bas selon leur impact, mesuré par SHAP. Sur chaque prédiction, le rouge montre la valeur élevée de la variable et le bleu la valeur de la variable. Un déplacement vers la droite traduit une contribution positive à la prédiction, tandis qu'un déplacement vers la gauche correspond à une influence négative. Ainsi, on observe que la variable *tstats_intelligents_connectés*, le nombre de thermostats intelligents connectés, quand elle est à droite, elle est en rouge et confirme qu'elle impacte fortement la prédiction. De la même manière, quand la variable *temperature_interieure_moyenne* est à droite, elle est en bleu et confirme également que, lorsqu'elle est élevée, la prédiction baisse, donc moins de besoins en chauffage, ce qui est logique. Généralement, les couleurs, les directions et les effets visibles dans cette figure 3.17, confirment que le modèle réagit de

façon cohérente aux conditions thermiques et temporelles, ce qui renforce la confiance dans ses prédictions.

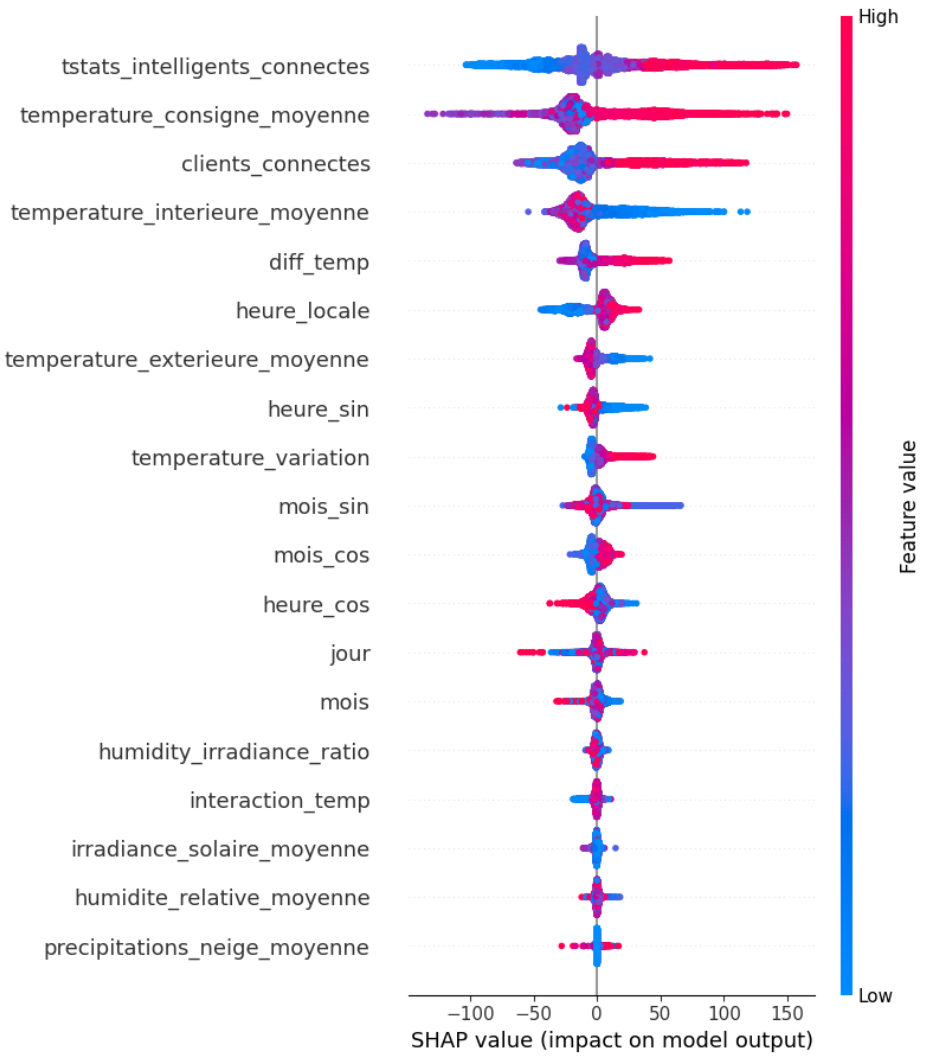

Figure 3.17 : Diagramme récapitulatif SHAP du CatBoost.

En somme, l'analyse explicative réalisée avec SHAP a permis de mieux comprendre le fonctionnement du modèle CatBoost et de justifier ses prédictions. Grâce aux visualisations du diagramme récapitulatif et de résumé de SHAP, il a été démontré que les prédictions du modèle reposent sur une combinaison logique de variables, en lien direct avec les

comportements attendus de consommation électrique résidentielle. Les variables importantes identifiées, telles que le nombre de thermostats intelligents connectés, la température intérieure, le nombre de clients connectés ou la température de consigne, agissent de manière réelle dans la prédiction. Cette transparence dans les décisions du modèle renforce non seulement la confiance dans les résultats, mais également sa pertinence pour une utilisation pratique dans des systèmes de gestion énergétique résidentielle.

3.5 MODÈLE DE FONDATION

L'apprentissage des modèles a nécessité beaucoup de temps et de ressources. Bien vrai que le modèle Catboost fonctionne bien, son optimisation a pris un temps d'exécution de 1784,801 secondes avec un CPU (unité centrale de traitement) pour trouver les meilleurs hyperparamètres, à savoir : iterations = 3288, learning_rate = 0,063, depth = 10, l2_leaf_reg = 6,37 et border_count = 116. Ce délai peut être amplifié par certains facteurs externes, tels que la qualité de la connexion réseau et la puissance du CPU. Sur un ensemble de données plus grand, le modèle peut être limité par le temps ou les ressources disponibles. Pour cela, on explore également des modèles pré-entraînés appelés modèles de fondation.

Les modèles fondamentaux sont des IA entraînées sur d'immenses quantités de données, ce qui les rend capable de faire des prédictions de façon rapide sans entraînement sur les données qui leur sont fournies (Zhou et al., 2024). Dans ce cas, l'objectif n'est pas de créer de nouveaux, mais utiliser ceux qui sont déjà disponibles pour faire la prédiction. Pour cela, deux modèles de fondation spécialisée aux séries temporelles sont choisis pour faire la prédiction plus rapidement, c'est-à-dire sans entraînement et nécessitant moins de ressources.

Le choix de ces modèles est fait en fonction de la performance de prédiction des données de consommation dans le secteur de l'énergie et de leur fonction principale qui est la prédiction des séries temporelles. Ce sont les modèles TimeGPT et TimeFM dont les fonctions et leur application seront détaillées dans la suite.

TimeGPT

TimeGPT est un modèle de fondation dédié spécialement aux séries temporelles. Il fait la prédiction avec précision sur des séries jamais vues pendant l'entraînement de base, comme le modèle GPT avec le langage (Garza et al., 2024). C'est un modèle qui a une architecture de transformer et qui a été entrainé sur des centaines de milliards de données dans des contextes variés. Cette architecture inclut l'encodage positionnel, l'attention masquée pour la prédiction, la convolution pour enrichir les représentations et une sortie probabiliste pour gérer l'incertitude. Pour son utilisation pour notre prédiction, l'ensemble de données a été nettoyé à nouveau afin de le transformer en deux colonnes (variable temporelle et variable cible) pour une prédiction univariée, parce que TimeGPT est à la base un modèle univarié. En plus, la variable temporelle a été formatée sur une fréquence journalière d'une part et mensuelle d'autre part, au lieu de la fréquence horaire de base, parce que cette dernière était irrégulière, or ce modèle fonctionne sur une série régulière. Ensuite, on utilise une clé API pour initialiser Nixtla afin d'effectuer la prédiction sur l'ensemble de la série. Puis, une prédiction rapide sans ajustement est faite en premier lieu sur les trente prochains jours avec la fréquence journalière et en second lieu sur les trois prochains mois avec la fréquence mensuelle. Les fenêtres de prédiction de trente jours et de trois mois sont choisies, car elles offrent une meilleure performance aux modèles. Enfin, les métriques des prédictions sont calculées et affichées dans le Tableau 3.9. Ces métriques sont obtenues en

comparant les prédictions de TimeGPT pour les trente jours à venir avec les trente vraies valeurs correspondantes situées à la fin de la série réelle. Les résultats de base obtenus sans fine-tuning ou l'utilisation de variables exogènes comme la météo sont importants. Cela montre déjà combien ce modèle est puissant et peut être utilisé dans la prédiction énergétique en temps réel, de façon rapide et à long et court terme.

Tableau 3.9 : Performance du modèles TimeGPT (Modèles Pré-entrainés)

Horizon de prédiction	Fréquence	MAE	RMSE	MAPE (%)
30 jours	Journalière	6.41	9.03	8.95
3 mois	Mensuelle	9.00	9.55	12.99

Cette comparaison montre que la fréquence journalière offre les meilleures performances. Le MAPE journalier (8,95 %) est inférieur au MAPE mensuel observé (12,99 %). Cela s'explique par une meilleure compréhension des variations à court terme et des tendances locales des données journalières, ce qui permet à TimeGPT de produire des prédictions à court terme plus précises. La fréquence mensuelle offre également l'avantage d'une planification à long terme plus lisible. Par conséquent, le choix de la granularité dépendra donc des objectifs finaux du système de gestion énergétique.

TimeFM

Afin de comparer le potentiel des modèles de fondation pour la prédiction de séries temporelles énergétiques, on a également évalué les performances du modèle TimesFM, récemment créé par Google DeepMind. Comme TimeGPT, TimesFM est aussi un modèle

préentraîné pour la prédiction automatisée des séries temporelles. Selon (Das et al., 2024), TimesFM adopte une architecture basée sur une segmentation en blocs temporels et une attention masquée. Contrairement aux Transformers classiques, il n'utilise pas d'encodage positionnel. Cela lui permet de capturer efficacement des motifs temporels tout en conservant la scalabilité. En sortie, ce modèle génère à la fois des prédictions ponctuelles et des quantiles, intégrant ainsi la gestion de l'incertitude dans la prédiction (Goel et al., 2025). Il démontre une performance impressionnante en zéro-shot sur divers benchmarks publics couvrant plusieurs domaines et granularités (Das et al., 2024). C'est donc un modèle de fondation performant, ce qui fait aussi l'objet de son choix. Dans le cadre de cette étude, le modèle a été exploité en mode zero-shot, sans réentraînement ni fine-tuning pour équilibrer la comparaison avec TimeGPT. L'ensemble de données a été restructuré également selon le format attendu par le modèle suivant les colonnes de variable temporelle et cible et agrégé selon deux fréquences distinctes, journalière et mensuelle. Ensuite, la prédiction est faite sur les deux différentes séries afin de comparer sa performance à long et à court terme, puis, de la même manière que l'approche du TimeGPT, les métriques sont calculées. Les résultats, présentés dans le tableau 3.8, révèlent que TimesFM fournit aussi une prédiction précise et fiable, confirmant son potentiel dans les systèmes de prédiction énergétique automatisée.

Tableau 3.10 : Performance du modèles TimesFM

Horizon de prédiction	Fréquence	MAE	RMSE	MAPE (%)
30 jours	Journalière	7.72	9.18	11.55
3 mois	Mensuelle	11.90	12.40	17.12

Le Tableau 3.10 montre également que la fréquence journalière offre les meilleures performances. Le MAPE journalier (11,55 %) est inférieur au MAPE mensuel (17,12 %). Exactement comme le TimeGPT, il génère des prédictions plus précises à court terme. Mais sa performance à long terme n'est pas négligeable, elle peut être aussi utile pour des données plus denses tout en maintenant cette performance. Avec ces résultats, on remarque aussi que le choix de la fenêtre de prédiction est important pour assurer une belle performance.

Étude comparative du TimeGPT et TimesFM

D'abord, il faut noter que l'implémentation des deux modèles ainsi que leur architecture sont différentes. TimeGPT est simple avec la clé API tandis que TimesFM demande une mise en œuvre plus avancée mais offre une personnalisation locale plus poussée. TimeGPT, développé par Nixtla, est accessible via une API simple avec clé, ne nécessite aucun entraînement local et prédit à partir d'une série univariée formatée avec les colonnes timestamp et value. Il est idéal pour une utilisation rapide, sans code complexe. En revanche, TimesFM, développé par Google DeepMind, s'utilise localement avec PyTorch, nécessite une configuration plus technique. Il travaille aussi sur des séries univariées structurées et offre plus de contrôle sur les paramètres du modèle.

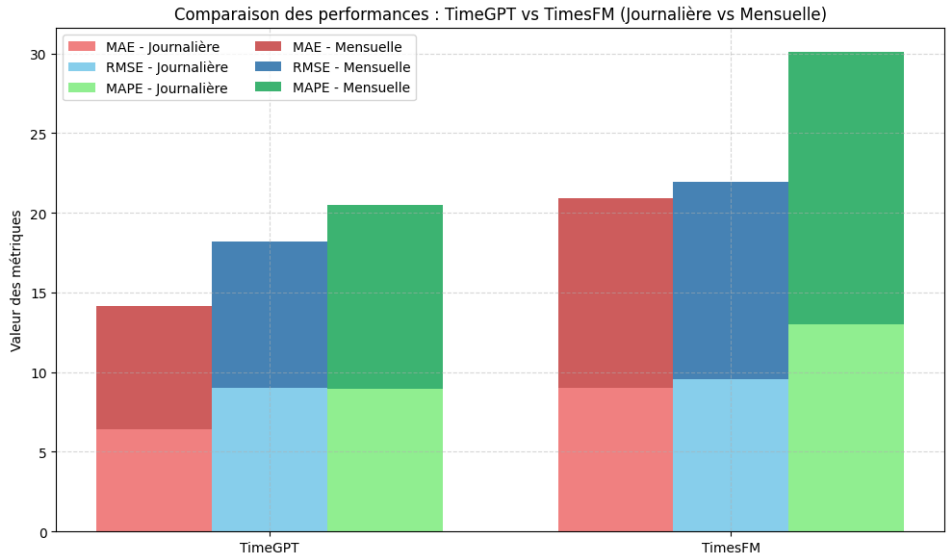

Figure 3.18 : Comparaison des performances TimeGPT et TimesFM

Ensuite, comme le montre la Figure 3.18, les deux modèles montrent une meilleure performance en fréquence journalière. Le MAPE passe de 8,95 % à 12,99 % pour TimeGPT, et de 11,55 % à 17,12 % pour TimesFM, quand on passe de la granularité journalière à mensuelle. Cette observation indique une perte de précision liée à la moyenne mensuelle, qui prouve l'efficacité des modèles pour la prédiction à court terme sur des données de consommation énergétique. La différence entre la performance journalière et mensuelle est plus significative pour TimesFM que pour TimeGPT. Cela pourrait suggérer que TimesFM est plus sensible à la réduction de la granularité temporelle, ou que son architecture nécessite davantage de points de données pour bien modéliser les tendances. Pour les deux horizons (30 jours et 3 mois), TimeGPT affiche clairement des valeurs inférieures de MAE, RMSE et MAPE. Cela suggère sa capacité plus efficace que celle su TimesFM pour prédire sans entraînement.

Enfin, on conclut que, TimeGPT surpasse TimesFM dans les conditions expérimentales actuelles, avec une meilleure précision aussi bien à court terme qu'à long

terme. Le choix des fréquences de pas régulier journalière ou mensuelle reste optimal pour les deux modèles en termes de précision. Cependant, TimesFM pourrait offrir des performances comparables ou supérieures s'il était ajusté spécifiquement aux données d'entrée. Cette étude comparative permet de mieux orienter le choix du modèle en fonction des contraintes d'implémentation, des objectifs temporels et des exigences de précision.

3.6 ÉTUDE COMPARATIVE GLOBALE

La comparaison globale des performances des modèles a été réalisée avec la métrique RMSE. Les résultats sur la Figure 3.19 montrent que le modèle de fondation TimeGPT en fréquence journalière offre la meilleure précision avec un RMSE de 9, ce qui fait de lui le modèle recommandé. Du côté des modèles classiques d'apprentissage automatique, CatBoost et XGBoost, après optimisation, ont montré une forte capacité prédictive avec des RMSE respectifs de 19 et 20 ; ils seront importants dans certains cas d'études. Cependant, la régression linéaire s'est montrée nettement moins performante avec un RMSE de 67, démontrant ses limites dans le traitement des séries temporelles non linéaires. Enfin, parmi les modèles d'apprentissage profond, LSTM avec un RMSE de 40 surpasse RNN, mais demeure globalement moins performant que les modèles d'ensemble ou de fondations. Ces résultats mettent en évidence la pertinence des modèles préentraînés ainsi que des méthodes d'ensemble améliorées pour obtenir des prédictions fiables avec ce type de données.

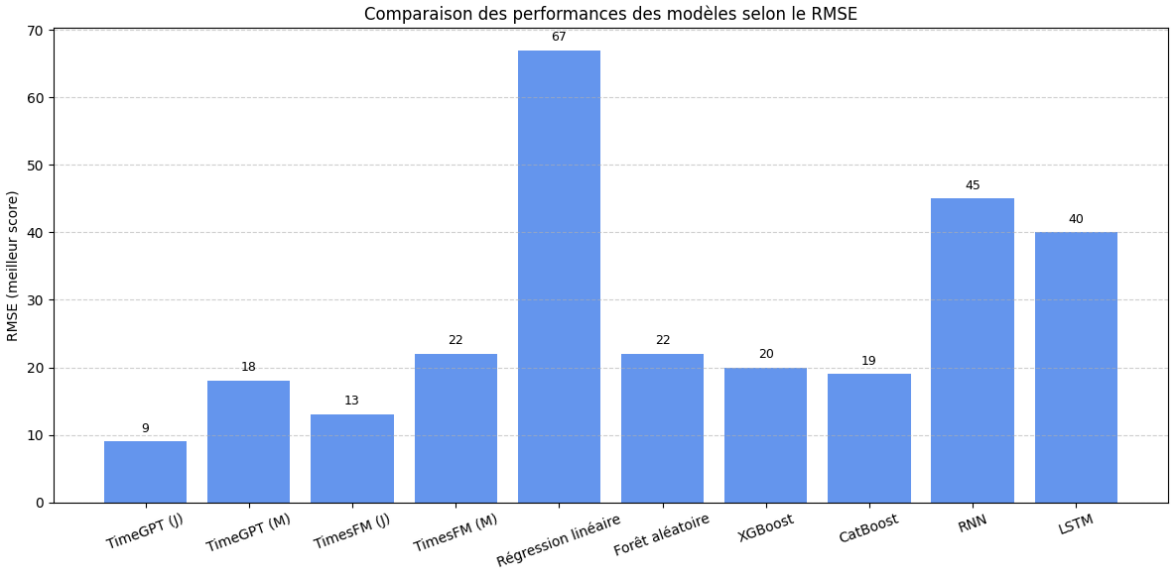

Figure 3.19: Comparaison des performances globales

En se basant sur la comparaison des performances des modèles selon le RMSE, il apparaît clairement que les modèles TimeGPT et TimeFM journalier offrent les meilleurs scores, avec un RMSE respectivement de 9 et 13 pour TimeGPT et TimeFM, ce qui est nettement inférieur à celui des autres modèles. Bien que le modèle CatBoost présente également une performance compétitive avec un RMSE de 19, son optimisation a nécessité un temps et des ressources considérables, ce qui pourrait limiter son application sur des ensembles de données plus volumineux. Pour cette raison, le choix des modèles de fondation spécialisés sur les séries temporelles, TimeGPT et TimeFM, est privilégié : ils permettent des prédictions rapides sans nécessiter d'entraînement intensif, tout en conservant une excellente précision, ce qui les rend particulièrement adaptés à la problématique de prédiction de la consommation énergétique étudiée dans ce mémoire.

CHAPITRE 4

CONCLUSION

Cette section présente les conclusions de ce mémoire. Elle commence par une synthèse du travail réalisé, suivie d'un aperçu des principales contributions. Ensuite, les perspectives de recherches futures sont abordées. Enfin, la section se clôture par l'apport spécifique de ce mémoire.

Le présent mémoire étudie la prédiction de la consommation électrique résidentielle à partir de données électriques et météorologiques, en mettant l'accent sur l'utilisation de modèles d'apprentissage automatique. Le chapitre 1 présente le contexte du sujet et pose les bases conceptuelles, les objectifs et les aspects techniques nécessaires à la compréhension de l'étude. Le chapitre 2, consacré à la revue de littérature, explore les travaux existants dans le domaine de la prédiction énergétique, les méthodes classiques et modernes utilisées, ainsi que les défis liés à la performance des modèles pour une prédiction suffisamment précise afin de contribuer à l'optimisation de la consommation. Le chapitre 3 présente les données utilisées ainsi que les étapes de préparation et de transformation nécessaires pour les rendre exploitables. Il décrit également le processus de développement des modèles de prédiction, incluant des modèles classiques (régression linéaire, forêt aléatoire, XGBoost, CatBoost) et des modèles d'apprentissage profond (RNN, LSTM), ainsi que les techniques d'optimisation mises en œuvre. Ensuite, il compare les résultats obtenus à différentes étapes de prédiction de base, avec ingénierie des caractéristiques et avec optimisation, afin d'analyser la réaction des modèles. L'explicabilité du meilleur modèle, qui a affiché les meilleures performances, est explorée à l'aide de la technique SHAP. De plus, il explore les modèles de fondation

TimeGPT et TimesFM pour évaluer leur efficacité sur les données énergétiques, offrant ainsi une alternative aux modèles classiques face aux contraintes de ressources et de temps. Cette étude s'est révélée intéressante en raison des nombreuses étapes à maîtriser, de la préparation des données passant par l'explicabilité des modèles jusqu'à l'exploration des modèles préentrainés. Cependant, il a démontré qu'une prédiction fiable de la consommation énergétique peut être réalisée grâce à des modèles ensemblistes comme le CatBoost.

4.1 REVUE DES CONTRIBUTIONS

Les travaux présentés dans ce mémoire apportent plusieurs contributions significatives au domaine de la prédiction énergétique résidentielle. Tout d'abord, cette étude fait partie des rares ayant combiné à la fois des modèles fondamentaux ainsi que des approches d'apprentissage automatique et profond, dans un contexte réel, afin de développer, évaluer et comparer la prédiction de la consommation énergétique.

La première contribution importante est la constitution d'un jeu de données prétraité et enrichi par des techniques d'ingénierie de caractéristiques, pouvant aider pour la poursuite d'études similaires. De plus, l'application de méthodes telles que l'optimisation bayésienne et Hyperband, ainsi que l'utilisation d'approches d'explicabilité comme SHAP, a permis d'affiner la compréhension de l'impact des variables sur les prédictions, offrant ainsi une transparence sur le fonctionnement du modèle CatBoost sur ce type de données.

La deuxième contribution est les résultats de SHAP qui montrent les utilisations à ajuster pour favoriser l'optimisation de la consommation électrique domestique.

La troisième contribution de cette étude met en évidence les enjeux liés à l'intégration de modèles de fondation dans les systèmes de gestion énergétique domestique, ouvrant ainsi la voie à de futurs développements dans les environnements intelligents. Elle propose également un cadre méthodologique reproductible, facilitant d'éventuels déploiements concrets dans le domaine de l'optimisation énergétique résidentielle.

De manière générale, les contributions de ce travail sont multiples et touchent plusieurs domaines. Sur le plan scientifique, il s'agit d'une avancée pour les recherches autour des modèles génératifs appliqués à la prédiction énergétique. Du côté des concepteurs de maisons intelligentes, cette étude constitue une ressource pour rendre leurs systèmes plus efficaces, en vue de réduire la consommation d'énergie. Enfin, les résultats sont aussi pertinents pour les distributeurs d'électricité, qui peuvent s'en servir pour mieux ajuster leur production.

4.2 IMPACTS ATTENDUS

Les résultats de cette recherche apporteront des améliorations significatives sur les plans environnemental, économique et technologique.

Sur le plan technologique, des systèmes de contrôles intelligents du chauffage et de la climatisation peuvent être conçus à partir des prédictions pour anticiper et limiter la consommation d'énergie afin d'augmenter l'efficacité opérationnelle dans les zones inoccupées du bâtiment pendant les périodes de forte demande (Shah et al., 2022). En plus, les systèmes de gestion intelligente de réponse à la demande connaitront une grande innovation. C'est-à-dire que ces systèmes peuvent ajuster automatiquement les paramètres des appareils en fonction des besoins anticipés, améliorer l'efficacité énergétique globale et mieux répondre aux variations de la demande en temps réel.

Sur le plan environnemental, L'optimisation de la consommation énergétique, notamment le contrôle de la température, des lumières et des dispositifs de stockage d'énergie, va réduire les émissions de gaz à effet de serre générées par les bâtiments résidentiels (Giannelos et al., 2024). Selon une étude internationale très récente, l'application d'un contrôle prédictif simple, basé sur les conditions météorologiques et les émissions, peut engendrer des économies d'énergie significatives et une réduction des émissions allant jusqu'à 25 %, tout en préservant le confort thermique. Étant donné que les opérations énergétiques des bâtiments représentent 28 % des émissions mondiales de carbone, la mise en œuvre de pratiques de gestion durable des bâtiments offre un potentiel considérable pour réaliser des économies et répondre aux préoccupations croissantes liées au changement climatique (Hepf et al., 2024).

Sur le plan économique, l'efficacité énergétique permet d'utiliser moins d'énergie pour les mêmes services tels que l'éclairage, le chauffage et le refroidissement, ce qui diminue les factures d'électricité pour les consommateurs (W. Chen et al., 2023). L'efficacité énergétique réduit la dépendance aux infrastructures coûteuses et en optimisant l'utilisation des ressources, ainsi les pays évitent de nouvelles constructions énergétiques. De plus, une meilleure efficacité énergétique permet de diminuer les coûts, en consommant moins pour le même niveau de production. Ainsi la gestion des ressources et les dépenses énergétiques nationales sont renforcées (F. Liu et al., 2023).

4.3 TRAVAUX FUTURS

Cette étude présente certaines limites, notamment en ce qui concerne l'utilisation directe de modèles de fondation comme TimeGPT dans un contexte de prédiction multivariée. Bien que TimeGPT démontre de solides performances en prédiction univariée, son application à des contextes plus complexes, comme celui de la consommation énergétique résidentielle influencée par des facteurs météorologiques et comportementaux, reste limitée. Une perspective prometteuse est celle proposée par (Garza et al., 2024) à travers le modèle TiMF, qui combine TimeGPT avec un perceptron multicouche (MLP). Cette architecture hybride permet l'intégration de variables exogènes sans réentraîner les poids du modèle fondation. Les résultats rapportés dans des contextes industriels montrent une nette amélioration de la précision prédictive, ce qui confirme l'intérêt d'un tel cadre pour les prévisions énergétiques contextuelles.

D'un autre côté, l'étude a également expérimenté TimesFM, un modèle exécutable localement. Contrairement à TimeGPT, TimesFM pourrait être soumis à un fine-tuning, permettant de l'adapter plus finement aux spécificités des données locales, ce qui n'a pas encore été réalisé ici, mais constitue un autre axe prometteur.

En parallèle, le modèle classique CatBoost a montré une excellente performance dans ce travail, avec un R^2 de 0,98. Toutefois, il pourrait réagir autrement sur d'autres types d'ensemble de données. Pour cela, plusieurs pistes peuvent être envisagées pour pousser encore plus loin les capacités de CatBoost, en testant l'effet du fine-tuning sur des sous-groupes de données comme les clusters saisonniers ou les profils de consommation pour créer des modèles personnalisés. Intégrer d'autres méthodes d'explicabilité plus affinée,

notamment LIME, afin d'analyser les interactions entre variables. Combiner CatBoost à des approches hybrides, par exemple en exploitant les sorties de TimeGPT ou de TimesFM comme nouvelles variables pour enrichir ses prédictions.

Enfin, des développements futurs pourraient viser la création d'une plateforme intégrée, facilitant la transformation des données et l'obtention de prédictions exploitables.

LISTE DE RÉFÉRENCES

Abd El-Aziz, R. M. (2022). Renewable power source energy consumption by hybrid machine learning model. *Alexandria Engineering Journal*, *61*(12), 9447-9455. https://doi.org/10.1016/j.aej.2022.03.019

AbuBaker, M. (2021). HOUSEHOLD ELECTRICITY LOAD FORECASTING TOWARD DEMAND RESPONSE PROGRAM USING DATA MINING TECHNIQUES IN A TRADITIONAL POWER GRID. *International Journal of Energy Economics and Policy*, *11*(4), 132-148. https://doi.org/10.32479/ijeep.11192

Aguirre-Fraire, B., Beltrán, J., & Soto-Mendoza, V. (2024). A comprehensive dataset integrating household energy consumption and weather conditions in a north-eastern Mexican urban city. *Data in Brief*, *54*, 110452. https://doi.org/10.1016/j.dib.2024.110452

Akgündoğdu, A., Öz, I., & Uzunoğlu, C. P. (2019). Signal quality based power output prediction of a real distribution transformer station using M5P model tree. *Electric Power Systems Research*, *177*, 106003. https://doi.org/10.1016/j.epsr.2019.106003

Al Kez, D., Foley, A., Lowans, C., & Del Rio, D. F. (2024). Energy poverty assessment : Indicators and implications for developing and developed countries. *Energy Conversion and Management*, *307*, 118324. https://doi.org/10.1016/j.enconman.2024.118324

Albahli, S., Shiraz, M., & Ayub, N. (2020). Electricity Price Forecasting for Cloud Computing Using an Enhanced Machine Learning Model. *IEEE Access*, *8*, 200971-200981. https://doi.org/10.1109/ACCESS.2020.3035328

Ali, U., Bano, S., Shamsi, M. H., Sood, D., Hoare, C., Zuo, W., Hewitt, N., & O'Donnell, J. (2024). Urban building energy performance prediction and retrofit analysis using data-driven machine learning approach. *Energy and Buildings*, *303*, 113768. https://doi.org/10.1016/j.enbuild.2023.113768

Amasyali, K., & El-Gohary, N. M. (2018). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, *81*, 1192-1205. https://doi.org/10.1016/j.rser.2017.04.095

Amin, A., & Mourshed, M. (2024). Weather and climate data for energy applications. *Renewable and Sustainable Energy Reviews*, *192*, 114247. https://doi.org/10.1016/j.rser.2023.114247

Bai, Z. (2024). Residential electricity prediction based on GA-LSTM modeling. *Energy Reports*, *11*, 6223-6232. https://doi.org/10.1016/j.egyr.2024.06.010

Banik, R., Das, P., Ray, S., & Biswas, A. (2021). Prediction of electrical energy consumption based on machine learning technique. *Electrical Engineering*, *103*(2), 909-920. https://doi.org/10.1007/s00202-020-01126-z

Barth, D., Mautor, T., De Moissac, A., Watel, D., & Weisser, M.-A. (2021). Optimisation of electrical network configuration : Complexity and algorithms for ring topologies. *Theoretical Computer Science*, *859*, 162-173. https://doi.org/10.1016/j.tcs.2021.01.023

Berardi, U., & Jafarpur, P. (2020). Assessing the impact of climate change on building heating and cooling energy demand in Canada. *Renewable and Sustainable Energy Reviews*, *121*, 109681. https://doi.org/10.1016/j.rser.2019.109681

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, *13*(null), 281-305.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197-227. https://doi.org/10.1007/s11749-016-0481-7

Bibri, S. E., & Krogstie, J. (2020). Environmentally data-driven smart sustainable cities : Applied innovative solutions for energy efficiency, pollution reduction, and urban metabolism. *Energy Informatics*, *3*(1), 29. https://doi.org/10.1186/s42162-020-00130-8

Bourhnane, S., Abid, M. R., Lghoul, R., Zine-Dine, K., Elkamoun, N., & Benhaddou, D. (2020). Machine learning for energy consumption prediction and scheduling in smart buildings. *SN Applied Sciences*, *2*(2), 297. https://doi.org/10.1007/s42452-020-2024-9

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis : Forecasting and Control*. John Wiley & Sons.

Chassagnon, G., Vakalopolou, M., Paragios, N., & Revel, M.-P. (2020). Deep learning : Definition and perspectives for thoracic imaging. *European Radiology*, *30*(4), 2021-2030. https://doi.org/10.1007/s00330-019-06564-3

Chen, T., & Guestrin, C. (2016). XGBoost : A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Chen, W., Alharthi, M., Zhang, J., & Khan, I. (2023). The need for energy efficiency and economic prosperity in a sustainable environment. *Gondwana Research*, S1342937X23001077. https://doi.org/10.1016/j.gr.2023.03.025

Chen, Y., Xu, P., Chu, Y., Li, W., Wu, Y., Ni, L., Bao, Y., & Wang, K. (2017). Short-term electrical load forecasting using the Support Vector Regression (SVR) model to

calculate the demand response baseline for office buildings. *Applied Energy*, *195*, 659-670. https://doi.org/10.1016/j.apenergy.2017.03.034

Chévez, P., Barbero, D., Martini, I., & Discoli, C. (2017). Application of the k-means clustering method for the detection and analysis of areas of homogeneous residential electricity consumption at the Great La Plata region, Buenos Aires, Argentina. *Sustainable Cities and Society*, *32*, 115-129. https://doi.org/10.1016/j.scs.2017.03.019

Chujai, P., Kerdprasop, N., & Kerdprasop, K. (2013). Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models. *Hong Kong*.

Čistý, M., Danko, M., Kohnová, S., Považanová, B., & Trizna, A. (2024). Machine Learning Enhanced by Feature Engineering for Estimating Snow Water Equivalent. *Water*, *16*(16), 2285. https://doi.org/10.3390/w16162285

Côté, P.-O., Nikanjam, A., Ahmed, N., Humeniuk, D., & Khomh, F. (2024). Data cleaning and machine learning : A systematic literature review. *Automated Software Engineering*, *31*(2), 54. https://doi.org/10.1007/s10515-024-00453-w

Darne, O., & Diebolt, C. (2007). La Reichsbank, 1876-1920. Une analyse institutionnelle et cliométrique. *Revue européenne des sciences sociales*, *XLV-137*, 203-212. https://doi.org/10.4000/ress.233

Das, A., Kong, W., Sen, R., & Zhou, Y. (2024). *A decoder-only foundation model for time-series forecasting* (arXiv:2310.10688). arXiv. https://doi.org/10.48550/arXiv.2310.10688

Dil, Aakash Ramchand. (2025). *Testing for Time Series Stationarity : A Practical Guide to the Dickey-Fuller and Augmented Dickey-Fuller Tests*. https://doi.org/10.2139/ssrn.5287311

Dinmohammadi, F., Han, Y., & Shafiee, M. (2023). Predicting Energy Consumption in Residential Buildings Using Advanced Machine Learning Algorithms. *Energies*, *16*(9), 3748. https://doi.org/10.3390/en16093748

Dostmohammadi, M., Pedram, M. Z., Hoseinzadeh, S., & Garcia, D. A. (2024). A GA-stacking ensemble approach for forecasting energy consumption in a smart household : A comparative study of ensemble methods. *Journal of Environmental Management*, *364*, 121264. https://doi.org/10.1016/j.jenvman.2024.121264

El Houda, B. N., Lakhdar, L., & Abdallah, M. (2022). Time Series Analysis of Household Electric Consumption with XGBoost Model. *2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, 1-6. https://doi.org/10.1109/PAIS56586.2022.9946913

ElShawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2021). Interpretability in healthcare : A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, *37*(4), 1633-1650. https://doi.org/10.1111/coin.12410

Emissions Gap Report 2021. (2021). https://www.unep.org/resources/emissions-gap-report-2021

Falkner, S., Klein, A., & Hutter, F. (2018). BOHB: Robust and efficient hyperparameter optimization at scale. *International conference on machine learning*, 1437-1446.

Fathi, S., Srinivasan, R., Fenner, A., & Fathi, S. (2020). Machine learning applications in urban building energy performance forecasting : A systematic review. *Renewable and Sustainable Energy Reviews*, *133*, 110287. https://doi.org/10.1016/j.rser.2020.110287

Fumo, N., & Rafe Biswas, M. A. (2015). Regression analysis for prediction of residential energy consumption. *Renewable and Sustainable Energy Reviews*, *47*, 332-343. https://doi.org/10.1016/j.rser.2015.03.035

Garza, A., Challu, C., & Mergenthaler-Canseco, M. (2024). *TimeGPT-1* (arXiv:2310.03589). arXiv. https://doi.org/10.48550/arXiv.2310.03589

Gellert, A., Fiore, U., Florea, A., Chis, R., & Palmieri, F. (2022). Forecasting Electricity Consumption and Production in Smart Homes through Statistical Methods. *Sustainable Cities and Society*, *76*, 103426. https://doi.org/10.1016/j.scs.2021.103426

Giannelos, S., Bellizio, F., Strbac, G., & Zhang, T. (2024). Machine learning approaches for predictions of CO2 emissions in the building sector. *Electric Power Systems Research*, *235*, 110735. https://doi.org/10.1016/j.epsr.2024.110735

Goel, A., Pasricha, P., & Kanniainen, J. (2025). *Time-Series Foundation AI Model for Value-at-Risk Forecasting* (arXiv:2410.11773). arXiv. https://doi.org/10.48550/arXiv.2410.11773

Güneralp, B., Zhou, Y., Ürge-Vorsatz, D., Gupta, M., Yu, S., Patel, P. L., Fragkias, M., Li, X., & Seto, K. C. (2017). Global scenarios of urban density and its impacts on building energy use through 2050. *Proceedings of the National Academy of Sciences*, *114*(34), 8945-8950. https://doi.org/10.1073/pnas.1606035114

Han, B., Zhang, S., Qin, L., Wang, X., Liu, Y., & Li, Z. (2022). Comparison of Support Vector Machine, Gaussian Process Regression and Decision Tree Models for Energy Consumption Prediction of Campus Buildings. *2022 8th International Conference on Hydraulic and Civil Engineering: Deep Space Intelligent*

Development and Utilization Forum (ICHCE), 689-693. https://doi.org/10.1109/ICHCE57331.2022.10042664

Han, Z., Zhao, J., Leung, H., Ma, K. F., & Wang, W. (2021). A Review of Deep Learning Models for Time Series Prediction. *IEEE Sensors Journal*, *21*(6), 7833-7848. https://doi.org/10.1109/JSEN.2019.2923982

Hepf, C., Gottkehaskamp, B., Miller, C., & Auer, T. (2024). International Comparison of Weather and Emission Predictive Building Control. *Buildings*, *14*(1), 288. https://doi.org/10.3390/buildings14010288

Hertel, L., Collado, J., Sadowski, P., Ott, J., & Baldi, P. (2020). Sherpa : Robust hyperparameter optimization for machine learning. *SoftwareX*, *12*, 100591. https://doi.org/10.1016/j.softx.2020.100591

Hsu, P.-C., Gao, L., & Hwang, Y. (2025). Comparative study of LSTM and ANN models for power consumption prediction of variable refrigerant flow (VRF) systems in buildings. *International Journal of Refrigeration*, *169*, 55-68. https://doi.org/10.1016/j.ijrefrig.2024.10.020

Huuki, H., Ruokamo, E., Kopsakangas-Savolainen, M., Belonogova, N., Sridhar, A., & Honkapuro, S. (2024). House and socio-demographic features vs. Electricity consumption time series in main heating mode classification. *The Electricity Journal*, *37*(2), 107373. https://doi.org/10.1016/j.tej.2024.107373

Hydro-Québec. (2024, novembre). *Consommation électrique de la clientèle participant à un programme de gestion locale de la demande de puissance*. https://donnees.hydroquebec.com/explore/dataset/consommation-clients-evenements-pointe/analyze/?flg=fr-fr

Injadat, M., Salo, F., Nassif, A. B., Essex, A., & Shami, A. (2018). Bayesian Optimization with Machine Learning Algorithms Towards Anomaly Detection. *2018 IEEE Global Communications Conference (GLOBECOM)*, 1-6. https://doi.org/10.1109/GLOCOM.2018.8647714

Kelany, O., Aly, S., & Ismail, M. A. (2020). Deep Learning Model for Financial Time Series Prediction. *2020 14th International Conference on Innovations in Information Technology (IIT)*, 120-125. https://doi.org/10.1109/IIT50501.2020.9299063

Kemal, M. S., & Olsen, R. L. (2016). *Adaptive Data Collection Mechanisms for Smart Monitoring of Distribution Grids* (arXiv:1608.06510). arXiv. https://doi.org/10.48550/arXiv.1608.06510

Khalil, M., McGough, A. S., Pourmirza, Z., Pazhoohesh, M., & Walker, S. (2022). Machine Learning, Deep Learning and Statistical Analysis for forecasting building energy consumption—A systematic review. *Engineering Applications of Artificial Intelligence*, *115*, 105287. https://doi.org/10.1016/j.engappai.2022.105287

Kim, H., Park, S., & Kim, S. (2023). Time-series clustering and forecasting household electricity demand using smart meter data. *Energy Reports*, *9*, 4111-4121. https://doi.org/10.1016/j.egyr.2023.03.042

Klyuev, R. V., Morgoev, I. D., Morgoeva, A. D., Gavrina, O. A., Martyushev, N. V., Efremenkov, E. A., & Mengxu, Q. (2022). Methods of Forecasting Electric Energy Consumption : A Literature Review. *Energies*, *15*(23), 8919. https://doi.org/10.3390/en15238919

Kolluru, V. K., Challagundla, Y., Chintakunta, A. N., Roy, B., Bermak, A., & M, R. D. S. (2024). AI-Driven Energy Optimization : Household Power Consumption

Prediction With LSTM Networks and PyTorch-Ray Tune in Smart IoT Systems. *2024 International Conference on Microelectronics (ICM)*, 1-6. https://doi.org/10.1109/ICM63406.2024.10815802

Kumar, A. (2020, juin 30). Introduction to the Gradient Boosting Algorithm. *Analytics Vidhya*. https://medium.com/analytics-vidhya/introduction-to-the-gradient-boosting-algorithm-c25c653f826b

Lauzon, D., & Gloaguen, E. (2024). Quantifying uncertainty and improving prospectivity mapping in mineral belts using transfer learning and Random Forest : A case study of copper mineralization in the Superior Craton Province, Quebec, Canada. *Ore Geology Reviews*, *166*, 105918. https://doi.org/10.1016/j.oregeorev.2024.105918

Lei, L., Chen, W., Wu, B., Chen, C., & Liu, W. (2021). A building energy consumption prediction model based on rough set theory and deep learning algorithms. *Energy and Buildings*, *240*, 110886. https://doi.org/10.1016/j.enbuild.2021.110886

Li, X., Wang, Z., Yang, C., & Bozkurt, A. (2024). An advanced framework for net electricity consumption prediction : Incorporating novel machine learning models and optimization algorithms. *Energy*, *296*, 131259. https://doi.org/10.1016/j.energy.2024.131259

Lien, S. K., & Rajasekharan, J. (2024). Automatic standard building category classification from smart meter data – A supervised learning approach. *Energy and Buildings*, *325*, 114954. https://doi.org/10.1016/j.enbuild.2024.114954

Lin, Y., Liu, J., Gabriel, K., Yang, W., & Li, C.-Q. (2022). Data-Driven Based Prediction of the Energy Consumption of Residential Buildings in Oshawa. *Buildings*, *12*(11), 2039. https://doi.org/10.3390/buildings12112039

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI : A Review of Machine Learning Interpretability Methods. *Entropy*, *23*(1), 18. https://doi.org/10.3390/e23010018

Liu, F., Sim, J., Sun, H., Edziah, B. K., Adom, P. K., & Song, S. (2023). Assessing the role of economic globalization on energy efficiency : Evidence from a global perspective. *China Economic Review*, *77*, 101897. https://doi.org/10.1016/j.chieco.2022.101897

Liu, J.-L., Wang, K., Xiahou, Q.-R., Liu, F.-M., Zou, J., & Kong, Y. (2019). China's long-term low carbon transition pathway under the urbanization process. *Advances in Climate Change Research*, *10*(4), 240-249. https://doi.org/10.1016/j.accre.2019.12.001

Lovatti, B. P. O., Nascimento, M. H. C., Neto, Á. C., Castro, E. V. R., & Filgueiras, P. R. (2019). Use of Random forest in the identification of important variables. *Microchemical Journal*, *145*, 1129-1134. https://doi.org/10.1016/j.microc.2018.12.028

Lu, Y., Vijayananth, V., & Perumal, T. (2025). Smart home energy prediction framework using temporal Kolmogorov-Arnold transformer. *Energy and Buildings*, *335*, 115529. https://doi.org/10.1016/j.enbuild.2025.115529

Mariano-Hernández, D., Hernández-Callejo, L., García, F. S., Duque-Perez, O., & Zorita-Lamadrid, A. L. (2020). A Review of Energy Consumption Forecasting in Smart Buildings : Methods, Input Variables, Forecasting Horizon and Metrics. *Applied Sciences*, *10*(23), 8323. https://doi.org/10.3390/app10238323

Mathumitha, R., Rathika, P., & Manimala, K. (2024). Intelligent deep learning techniques for energy consumption forecasting in smart buildings : A review. *Artificial Intelligence Review*, *57*(2), 35. https://doi.org/10.1007/s10462-023-10660-8

Matos, M., Almeida, J., Gonçalves, P., Baldo, F., Braz, F. J., & Bartolomeu, P. C. (2024). A Machine Learning-Based Electricity Consumption Forecast and Management System for Renewable Energy Communities. *Energies*, *17*(3), 630. https://doi.org/10.3390/en17030630

Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent Neural Networks : A Comprehensive Review of Architectures, Variants, and Applications. *Information*, *15*(9), 517. https://doi.org/10.3390/info15090517

Minoli, D., Sohraby, K., & Occhiogrosso, B. (2017). IoT Considerations, Requirements, and Architectures for Smart Buildings—Energy Optimization and Next-Generation Building Management Systems. *IEEE Internet of Things Journal*, *4*(1), 269-283. https://doi.org/10.1109/JIOT.2017.2647881

Mohammed, A. S., Asteris, P. G., Koopialipoor, M., Alexakis, D. E., Lemonis, M. E., & Armaghani, D. J. (2021). Stacking Ensemble Tree Models to Predict Energy Performance in Residential Buildings. *Sustainability*, *13*(15), 8298. https://doi.org/10.3390/su13158298

Nagauri, M. R. (2020, décembre 1). *Guide To Ensemble Methods : Bagging vs Boosting*. Analytics India Magazine. https://analyticsindiamag.com/deep-tech/guide-to-ensemble-methods-bagging-vs-boosting/

Nie, P., Roccotelli, M., Fanti, M. P., Ming, Z., & Li, Z. (2021). Prediction of home energy consumption based on gradient boosting regression tree. *Energy Reports*, *7*, 1246-1255. https://doi.org/10.1016/j.egyr.2021.02.006

Nyitrai, T., & Virág, M. (2019). The effects of handling outliers on the performance of bankruptcy prediction models. *Socio-Economic Planning Sciences*, *67*, 34-42. https://doi.org/10.1016/j.seps.2018.08.004

Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F., & Ajayi, S. (2022). Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques. *Journal of Building Engineering*, *45*, 103406. https://doi.org/10.1016/j.jobe.2021.103406

Panorama mondial des émissions de GES 2024. (2024). https://www.statistiques.developpement-durable.gouv.fr/edition-numerique/chiffres-cles-du-climat/fr/5-panorama-mondial-des-emissions-de

Pham, A.-D., Ngo, N.-T., Ha Truong, T. T., Huynh, N.-T., & Truong, N.-S. (2020). Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. *Journal of Cleaner Production*, *260*, 121082. https://doi.org/10.1016/j.jclepro.2020.121082

Premkumar, M., Sowmya, R., Ahmad, O. H., Chandran, R., Tan, C. S., Tengku Juhana, T. H., & Pradeep, J. (2025). Optimizing household energy management with distributed energy resources : A multi-learning-guided stochastic optimization approach. *Energy and Buildings*, *331*, 115323. https://doi.org/10.1016/j.enbuild.2025.115323

Radhoush, S., Whitaker, B. M., & Nehrir, H. (2023). An Overview of Supervised Machine Learning Approaches for Applications in Active Distribution Networks. *Energies*, *16*(16), 5972. https://doi.org/10.3390/en16165972

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Model-Agnostic Interpretability of Machine Learning* (arXiv:1606.05386). arXiv. https://doi.org/10.48550/arXiv.1606.05386

Sayed, H. A., William, A., & Said, A. M. (2023). Smart Electricity Meter Load Prediction in Dubai Using MLR, ANN, RF, and ARIMA. *Electronics*, *12*(2), 389. https://doi.org/10.3390/electronics12020389

Shachee, S. B., Latha, H. N., & Hegde Veena, N. (2022). Electrical Energy Consumption Prediction Using LSTM-RNN. Dans V. Suma, X. Fernando, K.-L. Du, & H. Wang (Éds.), *Evolutionary Computing and Mobile Sustainable Networks* (p. 365-384). Springer. https://doi.org/10.1007/978-981-16-9605-3_25

Shah, S., Iqbal, M., Aziz, Z., Rana, T., Khalid, A., Cheah, Y.-N., & Arif, M. (2022). The Role of Machine Learning and the Internet of Things in Smart Buildings for Energy Efficiency. *Applied Sciences*, *12*(15), 7882. https://doi.org/10.3390/app12157882

Sim, S. E., Tay, K. G., Huong, A., & Tiong, W. K. (2019). Forecasting Electricity Consumption Using SARIMA Method in IBM SPSS Software. *Universal Journal of Electrical and Electronic Engineering*, *6*(5B), 103-114. https://doi.org/10.13189/ujeee.2019.061614

Sivakumar, V. G., Arunfred, N., Anusha, N., Balakrishnan, C., Meenakshi, B., & Sujatha, S. (2024). A Gradient Boosting Algorithm to Predict Energy Consumption for Home Applications. *2024 2nd International Conference on Computer, Communication and Control (IC4)*, 1-5. https://doi.org/10.1109/IC457434.2024.10486226

Stacking in Machine Learning. (2021, avril 17). OpenGenus IQ: Learn Algorithms, DL, System Design. https://iq.opengenus.org/stacking-in-machine-learning/

Sun, S., Li, G., Chen, H., Guo, Y., Wang, J., Huang, Q., & Hu, W. (2017). Optimization of support vector regression model based on outlier detection methods for predicting electricity consumption of a public building WSHP system. *Energy and Buildings*, *151*, 35-44. https://doi.org/10.1016/j.enbuild.2017.06.056

Sun, Y., Haghighat, F., & Fung, B. C. M. (2020). A review of the-state-of-the-art in data-driven approaches for building energy prediction. *Energy and Buildings*, *221*, 110022. https://doi.org/10.1016/j.enbuild.2020.110022

Transition énergétique—Le gouvernement du Québec actualise son Plan directeur en transition, innovation et efficacité énergétiques. (2022). Gouvernement du Québec. https://www.quebec.ca/nouvelles/actualites/details/transition-energetique-le-gouvernement-du-quebec-actualise-son-plan-directeur-en-transition-innovation-et-efficacite-energetiques-41239

Uddin, I., Awan, H. H., Khalid, M., Khan, S., Akbar, S., Sarker, M. R., Abdolrasol, M. G. M., & Alghamdi, T. A. H. (2024). A hybrid residue based sequential encoding mechanism with XGBoost improved ensemble model for identifying 5-hydroxymethylcytosine modifications. *Scientific Reports*, *14*(1), 20819. https://doi.org/10.1038/s41598-024-71568-z

Ullah, F. U. M., Ullah, A., Haq, I. U., Rho, S., & Baik, S. W. (2020). Short-Term Prediction of Residential Power Energy Consumption via CNN and Multi-Layer Bi-Directional LSTM Networks. *IEEE Access*, *8*, 123369-123380. https://doi.org/10.1109/ACCESS.2019.2963045

Verdonck, T., Baesens, B., Óskarsdóttir, M., & Vanden Broucke, S. (2024). Special issue on feature engineering editorial. *Machine Learning*, *113*(7), 3917-3928. https://doi.org/10.1007/s10994-021-06042-2

Vu, T., Thirunavukkarasu, G. S., Seyedmahmoudian, M., Mekhilef, S., & Stojcevski, A. (2023). Comparative Analysis of Regression Models for Household Appliance Energy Consumption Prediction using Extreme Gradient Boosting. *2023 33rd Australasian Universities Power Engineering Conference (AUPEC)*, 1-6. https://doi.org/10.1109/AUPEC59354.2023.10503204

Wang, J., Xu, J., & Wang, X. (2018). *Combination of Hyperband and Bayesian Optimization for Hyperparameter Optimization in Deep Learning* (arXiv:1801.01596). arXiv. https://doi.org/10.48550/arXiv.1801.01596

Wang, R., Lu, S., & Feng, W. (2020). A novel improved model for building energy consumption prediction based on model integration. *Applied Energy*, *262*, 114561. https://doi.org/10.1016/j.apenergy.2020.114561

Wang, Z., & Srinivasan, R. S. (2017). A review of artificial intelligence based building energy use prediction : Contrasting the capabilities of single and ensemble prediction models. *Renewable and Sustainable Energy Reviews*, *75*, 796-808. https://doi.org/10.1016/j.rser.2016.10.079

Wei, Y., Zhang, X., Shi, Y., Xia, L., Pan, S., Wu, J., Han, M., & Zhao, X. (2018). A review of data-driven approaches for prediction and classification of building energy consumption. *Renewable and Sustainable Energy Reviews*, *82*, 1027-1047. https://doi.org/10.1016/j.rser.2017.09.108

World Energy Outlook 2023 – Analysis—IEA. (2023). https://www.iea.org/reports/world-energy-outlook-2023

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms : Theory and practice. *Neurocomputing*, *415*, 295-316. https://doi.org/10.1016/j.neucom.2020.07.061

Yazdan, M. M. S., Khosravia, M., Saki, S., & Mehedi, M. A. A. (2022). *Forecasting Energy Consumption Time Series Using Recurrent Neural Network in Tensorflow*. https://doi.org/10.20944/preprints202209.0404.v1

Yuan, T., Zhu, N., Shi, Y., Chang, C., Yang, K., & Ding, Y. (2018). Sample data selection method for improving the prediction accuracy of the heating energy consumption. *Energy and Buildings*, *158*, 234-243. https://doi.org/10.1016/j.enbuild.2017.10.006

Zhang, F., Fleyeh, H., & Bales, C. (2022). A hybrid model based on bidirectional long short-term memory neural network and Catboost for short-term electricity spot price forecasting. *Journal of the Operational Research Society*, *73*(2), 301-325. https://doi.org/10.1080/01605682.2020.1843976

Zhang, L., Chen, Y., & Yan, Z. (2023). Predicting the short-term electricity demand based on the weather variables using a hybrid CatBoost-PPSO model. *Journal of Building Engineering*, *71*, 106432. https://doi.org/10.1016/j.jobe.2023.106432

Zhang, X. M., Grolinger, K., Capretz, M. A. M., & Seewald, L. (2018). Forecasting Residential Energy Consumption : Single Household Perspective. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 110-117. https://doi.org/10.1109/ICMLA.2018.00024

Zhou, T., Xia, W., Zhang, F., Chang, B., Wang, W., Yuan, Y., Konukoglu, E., & Cremers, D. (2024). *Image Segmentation in Foundation Model Era : A Survey* (arXiv:2408.12957). arXiv. https://doi.org/10.48550/arXiv.2408.12957