Prédiction de la consommation électrique journalière domestique à l'aide de la météo et du calendrier : approche par régression

1. Mohammed Adil MANI
Master 1 ISI
SUPMTI
Maroc

2e Imane AZZA
Master 1 ISI
SUPMTI
Maroc

3e Yousser LAMKHANTAR
Master 1 ISI
SUPMTI
Maroc

***Résumé*—Nous traitons la prédiction de la consommation électrique journalière (en kWh) d'un logement à Sceaux, France, à partir du jeu de données « Individual Household Electric Power Consumption » et des données météo historiques de l'API Open-Meteo. Nous testons l'hypothèse selon laquelle la température n'est pas le seul critère explicatif : les jours fériés et le calendrier (week-end) influencent également la consommation. Après collecte et fusion des données (dataset, API météo, jours fériés), nous analysons les données puis entraînons trois modèles (Random Forest, Gradient Boosting, régression linéaire) sur le même tableau fusionné. Le Random Forest, avec les variables sélectionnées automatiquement (DJU, lag-1, etc.), offre la meilleure performance (R^2 = 0,66, RMSE = 4,1 kWh, MAE = 3,18 kWh). Nous présentons une étude comparative des trois modèles, puis la prédiction sur une semaine et une discussion de la marge d'erreur journalière (RMSE), réalisées à l'aide du meilleur algorithme retenu (Random Forest).**

***Index Terms*—Consommation électrique, régression, Random Forest, Gradient Boosting, Week-end, Jours Feriés, série temporelle, météo.**

I. Introduction

La prédiction de la consommation électrique domestique aide à la gestion de la demande, aux prévisions de facturation et à l'efficacité énergétique. Nous nous concentrons sur la consommation journalière agrégée (kWh) d'un seul logement à Sceaux, France, à partir de mesures à la minute (jeu UCI « Individual Household Electric Power Consumption ») et de la météo journalière de l'API Open-Meteo. L'objectif est de construire un modèle de régression qui explique et prédit la consommation journalière à partir de la météo et du calendrier, et de documenter l'impact du nettoyage, de l'ingénierie des variables et du filtrage sur les performances (R^2, RMSE, MAE).

II. Hypothèse

Nous posons l'hypothèse que **la température n'est pas le seul critère** qui détermine la consommation d'énergie : **les jours fériés** et le type de jour (week-end ou non) influencent également les usages et donc la consommation électrique. La section IV (Analyse des données) permettra de vérifier cette hypothèse en montrant que les jours de plus forte consommation sont souvent des week-ends en hiver, comme le montrent les figures 2 et 3.

III. Collecte des données

A. Dataset de consommation

Le jeu principal est un fichier texte contenant des mesures à la minute : puissance active et réactive globale (kW), tension (V), intensité globale (A), et trois canaux de sous-comptage (Wh). Le fichier utilise des valeurs séparées par des points-virgules et la virgule comme séparateur décimal ; les valeurs manquantes sont codées « ? ». Les données sont chargées avec `pandas.read_csv` ; les lignes avec valeurs manquantes dans les colonnes numériques clés sont supprimées. Les colonnes `Date` et `Time` sont fusionnées en un index `datetime` pour permettre l'agrégation journalière et la fusion avec l'API. L'énergie journalière (kWh) est obtenue en sommant `Global_active_power/60` sur chaque jour.

B. API météo (Open-Meteo)

La météo journalière historique pour Sceaux (latitude 48,78, longitude 2,29) est demandée à https://archive-api.open-meteo.com/v1/archive sur la même plage de dates que les données du dataset. Les variables récupérées incluent : `temperature_2m_mean/min/max`, `precipitation_sum`, `daylight_duration`, `sunshine_duration` et la durée d'ensoleillement qui sera convertie en heures (`daylight_hours`).

C. Jours fériés (Holidays)

La bibliothèque Python `holidays` permet d'obtenir les jours fériés français pour la période 2006–2010 (`holidays.France(years=[2006,...,2010])`). Une variable binaire `is_holiday` est ajoutée, de même que `is_weekend` (samedi ou dimanche).

D. Fusion des données

Le DataFrame de consommation est fusionné avec le DataFrame météo sur la colonne `date` (`how='left'`) pour constituer le **tableau fusionné** (données + météo) utilisé dans toute la suite.

E. Contenu des variables du jeu de données

Le tableau I décrit le contenu de chaque variable du tableau fusionné (consommation + météo + calendrier + variables dérivées) utilisé pour l'analyse et la modélisation.

TABLE I – Description du contenu de chaque variable du jeu de données fusionné.

Variable	Contenu
`date`	Date du jour.
`energy_kwh`	Consommation électrique journalière (kWh), variable cible.
`mois`	Numéro du mois (1 à 12).
`jour_semaine`	Jour de la semaine.
`is_weekend`	Indicateur week-end (1 si samedi ou dimanche, 0 sinon).
`is_holiday`	Indicateur jour férié français (1 si férié, 0 sinon).
`temperature_2m_mean`	Température moyenne à 2 m (°C), API Open-Meteo.
`temperature_2m_min`	Température minimale à 2 m (°C), API Open-Meteo.
`temperature_2m_max`	Température maximale à 2 m (°C), API Open-Meteo.
`precipitation_sum`	Cumul des précipitations (mm), API Open-Meteo.
`daylight_duration`	Durée du jour entre lever et coucher du soleil (s), API.
`sunshine_duration`	Durée d'ensoleillement (s), API Open-Meteo.
`daylight_hours`	Durée du jour en heures (dérivée : `daylight_duration`/3600).
`DJU`	Degrés-jours unifiés, $\max(0, 18 - T_{\text{mean}})$ (°C·jour).
`sin_mois`, `cos_mois`	Encodage cyclique du mois (saisonnalité).
`energy_kwh_lag1`	Consommation de la veille (kWh), variable décalée.

IV. Analyse des données

La consommation journalière moyenne par mois (toutes années confondues) est représentée en barres : la consommation est nettement plus élevée en hiver (décembre, janvier, février, novembre) et plus faible en été (juillet, août). Comme le montre la figure 1, ce résultat conforte le lien entre la consommation et le froid (chauffage, éclairage, présence).

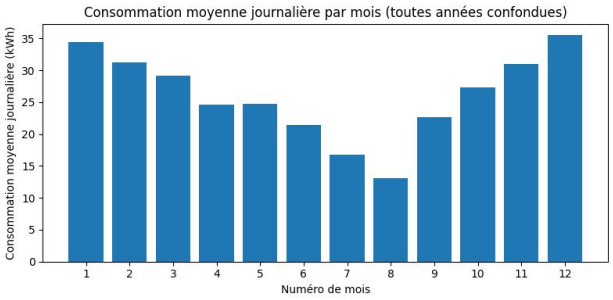

FIGURE 1 – Consommation moyenne journalière par mois (toutes années confondues).

Un graphique en barres par tranche de température, avec le nombre de jours affiché au-dessus de chaque barre, montre que la consommation diminue lorsque la température augmente (froid → forte consommation). Comme l'indique la figure 2, cette visualisation renforce la relation entre le « fait froid » et la « consommation ».

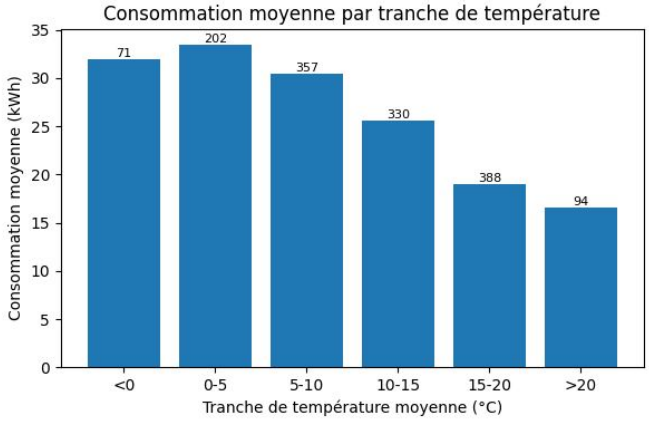

FIGURE 2 – Consommation moyenne par tranche de température. Les nombres au-dessus des barres indiquent le nombre de jours par tranche.

Les dix jours présentant la plus forte consommation sont examinés (`daily.nlargest(10, ''energy_kwh'')`). Comme le montre la figure 3, ces jours sont majoritairement des *week-ends en hiver*. Cette observation confirme l'hypothèse : le calendrier (week-end) et la saison (froid) expliquent les pics de consommation, en plus de la température.

	date	jour_nom	is_weekend	is_holiday	energy_kwh
7	2006-12-23	Saturday	True	False	79.556433
49	2007-02-03	Saturday	True	False	67.162033
10	2006-12-26	Tuesday	False	False	65.568500
64	2007-02-18	Sunday	True	False	63.829367
50	2007-02-04	Sunday	True	False	59.932333
57	2007-02-11	Sunday	True	False	59.520467
105	2007-03-31	Saturday	True	False	58.491833
15	2006-12-31	Sunday	True	False	58.236600
85	2007-03-11	Sunday	True	False	58.010600
36	2007-01-21	Sunday	True	False	56.787700

FIGURE 3 – Top 10 des jours en termes de consommation (week-ends en hiver dominent).

V. Entraînement des modèles

A. Random Forest

1) Premier modèle: Un premier Random Forest est entraîné avec les variables brutes : `temperature_2m_mean/min/max`, `precipitation_sum`, `daylight_hours`, `sunshine_duration`, `jour_semaine`, `is_weekend`, `is_holiday`, `mois`. Le découpage est de 75 %/25 % dans l'ordre chronologique (`shuffle=False`). Les résultats

sont : R^2 = 0,23, RMSE = 7,31 kWh. La prédiction n'est pas satisfaisante.

2) Température non linéaire et calcul du DJU (seuil 18 °C): La **température ambiante** n'est pas linéaire avec la consommation : en froid extrême (par ex. -2 °C), le chauffage augmente fortement la consommation ; en chaleur extrême (par ex. 30 °C), la climatisation peut également l'augmenter ; entre les deux (zone de confort), la consommation est plus faible. La relation température–consommation présente une forme en U. Nous introduisons les **degrés-jours unifiés (DJU)** avec un seuil de confort de 18 °C : $\mathrm{DJU} = \max(0, 18 - T_{\mathrm{mean}})$. Le DJU offre une relation linéaire avec la consommation liée au chauffage (côté froid). Un Random Forest entraîné avec le DJU en remplacement (ou en complément) des seules températures, sur le même découpage 75 %/25 % et l'ensemble non filtré, donne des résultats améliorés : R^2 = 0,30, RMSE = 6,2 kWh.

3) Décalage lag-1: Nous ajoutons la consommation de la veille (`energy_kwh_lag1`) comme variable décalée. Les lignes pour lesquelles le lag est manquant sont supprimées. Nous ajoutons également l'encodage cyclique du mois (`sin_mois`, `cos_mois`) afin que décembre et janvier soient proches. Sur l'ensemble non filtré, les performances s'améliorent : R^2 = 0,40, RMSE = 5,4 kWh, MAE = 4,2 kWh. Un nuage de points réel vs. prédit montre toutefois que la relation n'est pas purement linéaire et que d'autres facteurs (ou valeurs atypiques) entrent en jeu (voir la figure 4).

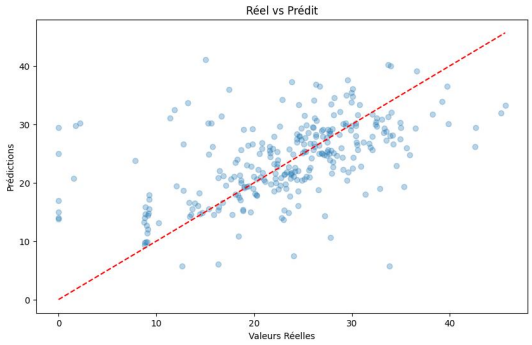

FIGURE 4 – Nuage de points réel vs. prédit (Random Forest, ensemble non filtré). R^2 (plus élevé = mieux).

4) Filtrage et choix automatique des variables: Nous excluons les jours à très faible ou très forte consommation : nous conservons les jours pour lesquels `energy_kwh` $>$ 4 et `energy_kwh` $\leq$ 33 kWh. Nous utilisons l'**importance des variables** du Random Forest pour sélectionner les variables. Le tableau II présente les importances obtenues sur le jeu d'entraînement.

TABLE II – Importance des variables (Random Forest).

Variable	Importance
`energy_kwh_lag1`	0,5096
`daylight_hours`	0,0934
`temperature_2m_min`	0,0644
`sunshine_duration`	0,0584
`temperature_2m_max`	0,0534
`DJU`	0,0459
`temperature_2m_mean`	0,0437
`jour_semaine`	0,0343
`precipitation_sum`	0,0340
`mois`	0,0203
`sin_mois`	0,0149
`is_holiday`	0,0117
`cos_mois`	0,0093
`is_weekend`	0,0067

Pour compléter la discussion du choix des variables, nous calculons la **matrice de corrélation** des variables retenues sur le jeu filtré. Elle donne les coefficients de corrélation de Pearson entre chaque paire de variables : une valeur proche de 1 (resp. -1) indique une corrélation linéaire positive (resp. négative) forte ; une valeur proche de 0 indique une liaison linéaire faible. Comme le montre la figure 7, la diagonale vaut 1 (chaque variable avec elle-même) et la matrice est symétrique. Les variables de température (`temperature_2m_min`, `temperature_2m_max`, `temperature_2m_mean`) sont très fortement corrélées entre elles (proches de 0,98), ce qui signale une possible redondance ; le DJU est très fortement négativement corrélé avec les températures (environ $-0{,}98$ avec `temperature_2m_mean`), ce qui est cohérent avec sa définition (degrés-jours liés au froid). La consommation de la veille (`energy_kwh_lag1`) présente une corrélation positive avec le DJU (environ 0,53) et négative avec les températures et `daylight_hours`, reflétant le lien entre consommation passée, froid et saison. Les variables `sin_mois` et `cos_mois` capturent la saisonnalité (forte corrélation négative entre `daylight_hours` et `cos_mois`). Les variables `jour_semaine` et `precipitation_sum` ont des corrélations faibles avec la plupart des autres, ce qui indique qu'elles apportent une information complémentaire. Cette matrice aide à interpréter le choix des variables et à repérer d'éventuelles colinéarités (voir la figure 7 en annexe).

Le lag-1 domine (environ 51 %). Nous retenons les variables listées ci-dessus. Avec le jeu filtré et un découpage chronologique 75 %/25 %, les **résultats finaux** du Random Forest sont présentés dans le tableau III. En résumé : premier modèle (variables brutes), R^2 = 0,23 et RMSE = 7,31 kWh ; après introduction du DJU (ensemble non filtré), R^2 = 0,30 et RMSE = 6,2 kWh ; après lag-1 et mois cyclique (ensemble non filtré), R^2 = 0,40, RMSE = 5,4 kWh et MAE = 4,2 kWh ; après filtrage et choix des variables, R^2 = 0,66, RMSE = 4,1 kWh et MAE = 3,18 kWh.

TABLE III – Résultats Random Forest (jeu filtré, 75 %/25 %, variables sélectionnées).

Modèle	R^2	RMSE (kWh)	MAE (kWh)
Random Forest (final)	0,66	4,10	3,18

B. Gradient Boosting (GBoost)

Le **Gradient Boosting** est une méthode d'ensemble qui construit séquentiellement des arbres de régression : chaque nouvel arbre corrige les erreurs résiduelles du modèle courant. On minimise une fonction de perte (par ex. l'erreur quadratique) par descente de gradient, ce qui permet d'obtenir un modèle puissant et régularisable (profondeur limitée, taux d'apprentissage, early stopping). Nous avons entraîné un régresseur `HistGradientBoostingRegressor` (scikit-learn) avec les **mêmes variables déjà sélectionnées** que pour le Random Forest (tableau II), sur le même tableau fusionné et avec le même découpage train/test (75 %/25 %, ordre chronologique). Les résultats sont présentés dans le tableau IV.

TABLE IV – Résultats Gradient Boosting (mêmes variables que le Forest, jeu filtré, 75 %/25 %).

Modèle	R^2	RMSE (kWh)	MAE (kWh)
Gradient Boosting	0,661	4,12	3,33

C. Régression linéaire

La **régression linéaire** suppose une relation linéaire entre les variables explicatives et la variable cible : $\hat{y} = \beta_0 + \sum_j \beta_j x_j$. Les coefficients sont estimés par moindres carrés. Il s'agit d'un modèle simple et interprétable, mais il peut sous-performer lorsque la relation est non linéaire (comme pour la relation température–consommation). Nous avons entraîné une régression linéaire sur les **mêmes variables** que pour le Random Forest et le GBoost, avec une mise à l'échelle (`StandardScaler`) et le même découpage 75 %/25 %. Les résultats sont présentés dans le tableau V.

TABLE V – Résultats régression linéaire (mêmes variables, jeu filtré, 75 %/25 %).

Modèle	R^2	RMSE (kWh)	MAE (kWh)
Régression linéaire	0,534	4,83	3,94

D. Comparaison entre les modèles

Une étude comparative des trois modèles (Random Forest, Gradient Boosting, régression linéaire) sur les métriques R^2, RMSE et MAE est présentée dans la figure 5. Comme l'indique cette figure, le Random Forest obtient le R^2 le plus élevé (0,66), suivi du Gradient Boosting (0,661) et de la régression linéaire (0,534) ; il offre également le RMSE et le MAE les plus faibles. Nous retenons donc le **Random Forest** comme meilleur modèle pour la prédiction.

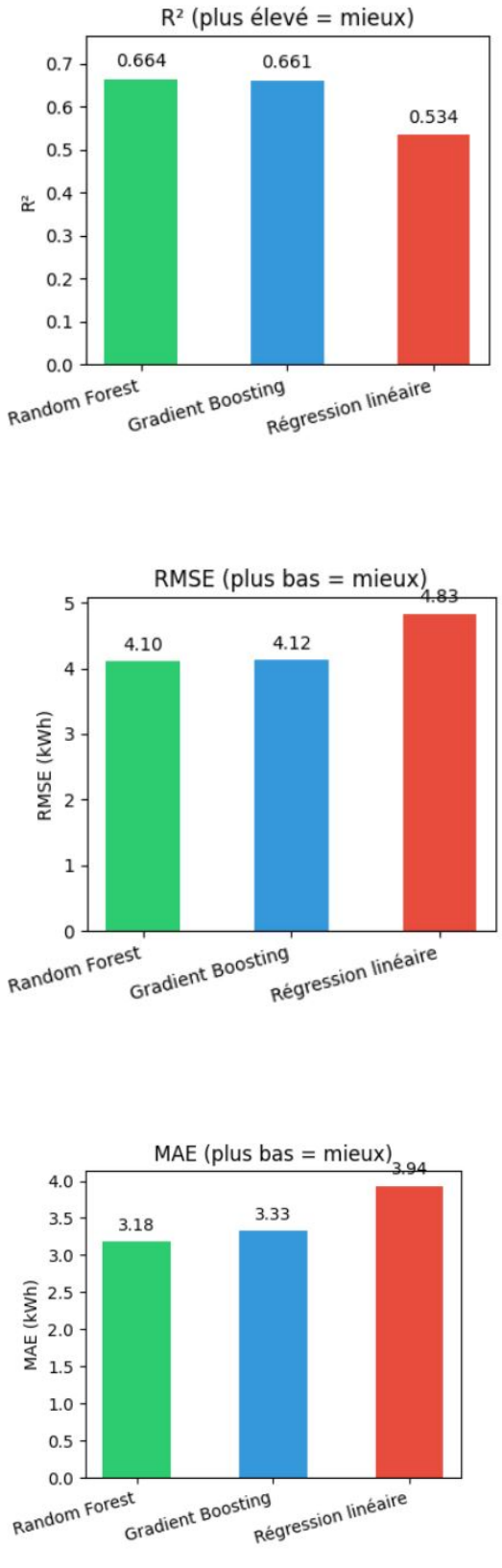

FIGURE 5 – Étude comparative des performances des trois modèles (Random Forest, Gradient Boosting, régression linéaire).

E. Distribution des erreurs

La distribution des erreurs (résidus : réel − prédit) sur le jeu de test est centrée autour de zéro, avec un pic pour les petites erreurs, ce qui indique que la plupart des jours sont prédits avec une erreur limitée (voir la figure 6).

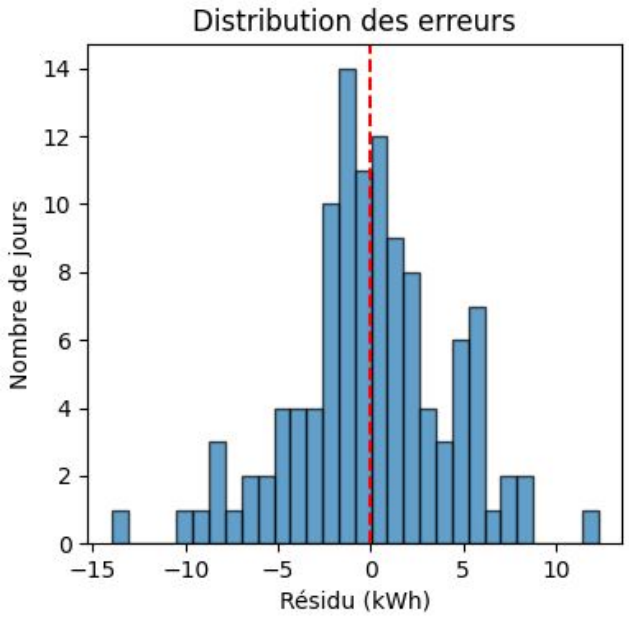

FIGURE 6 – Distribution des erreurs (résidus en kWh). La ligne verticale à zéro indique l'absence d'erreur.

VI. PRÉDICTION

Le modèle Random Forest retenu est utilisé pour prédire la consommation des sept prochains jours. Les dates sont celles renvoyées par l'API de prévision Open-Meteo. Pour une prévision future, la consommation de la veille (lag-1) n'étant pas disponible, nous utilisons une valeur de bootstrap (dernière consommation connue) pour le premier jour, puis nous chaînons les prédictions (lag-1 = prédiction du jour précédent). Le tableau VI présente les prédictions journalières.

TABLE VI – Prédictions de consommation (kWh) sur une semaine (dates API forecast).

Date	Consommation prédite (kWh)
2026-01-29	28,42
2026-01-30	27,87
2026-01-31	27,73
2026-02-01	27,95
2026-02-02	27,31
2026-02-03	27,01
2026-02-04	27,16

D'après le tableau VI, la moyenne sur la semaine est de 27,64 kWh/jour et la consommation totale sur la semaine est de 193,46 kWh.

La **marge d'erreur journalière** du modèle retenu est donnée par le RMSE : 4,1 kWh par jour. Cela signifie qu'en moyenne, l'erreur de prédiction sur la consommation d'un jour est de l'ordre de 4,1 kWh. Pour un tarif électrique usuel de 0,15 €/kWh (tarif bleu France), cette marge d'erreur journalière correspond à environ $4{,}1 \times 0{,}15 \approx 0{,}62$ € par jour.

VII. LIMITES, RISQUES ET BIAIS

- **Un seul foyer** : le modèle est ajusté sur un logement à Sceaux ; il ne garantit pas un transfert à d'autres sites ou ménages.
- **Variable décalée** : l'utilisation de `energy_kwh_lag1` nécessite la connaissance de la consommation de la veille ; une prévision « une semaine à l'avance » sans lag demanderait d'autres variables.
- **API météo** : les données d'archive sont supposées correctes ; des erreurs ou des lacunes pourraient biaiser le modèle.
- **Aucune causalité** : une forte corrélation entre le froid et la consommation ne prouve pas la causalité ; des facteurs non observés (par ex. la présence) peuvent intervenir.

VIII. RECOMMANDATIONS ET PISTES D'AMÉLIORATION

- **Granularité horaire** : refaire l'analyse à l'échelle horaire avec une météo horaire pour capturer les profils intra-journaliers.
- **Hyperparamètres** : régler le Random Forest ou le GBoost par validation croisée temporelle (`TimeSeriesSplit`).
- **Interprétabilité** : utiliser SHAP ou les courbes de dépendance partielle pour clarifier l'effet du DJU, du lag et du week-end ou du jour férié.
- **Usage opérationnel** : pour des prévisions sans lag, entraîner un modèle n'utilisant que la météo et le calendrier et rapporter ses performances séparément.

IX. CONCLUSION

Nous avons collecté et fusionné les données de consommation avec la météo (Open-Meteo) et le calendrier (jours fériés, week-end). L'analyse exploratoire a confirmé l'hypothèse : la température n'est pas le seul critère ; les jours de plus forte consommation sont souvent des week-ends en hiver. Nous avons entraîné trois modèles (Random Forest, Gradient Boosting, régression linéaire) sur le même tableau fusionné, avec les variables dérivées (DJU, lag-1, mois cyclique) et le filtrage des jours atypiques. L'étude comparative (R^2, RMSE, MAE) a conduit à retenir le Random Forest (R^2 = 0,66, RMSE = 4,1 kWh, MAE = 3,18 kWh). La prédiction sur une semaine a été illustrée ; la marge d'erreur journalière (RMSE = 4,1 kWh/jour, soit environ 0,62 €/jour pour un tarif de 0,15 €/kWh) caractérise la précision du modèle au niveau quotidien.

ANNEXE

La figure 7 présente la matrice de corrélation des variables (Pearson) sur le jeu filtré.

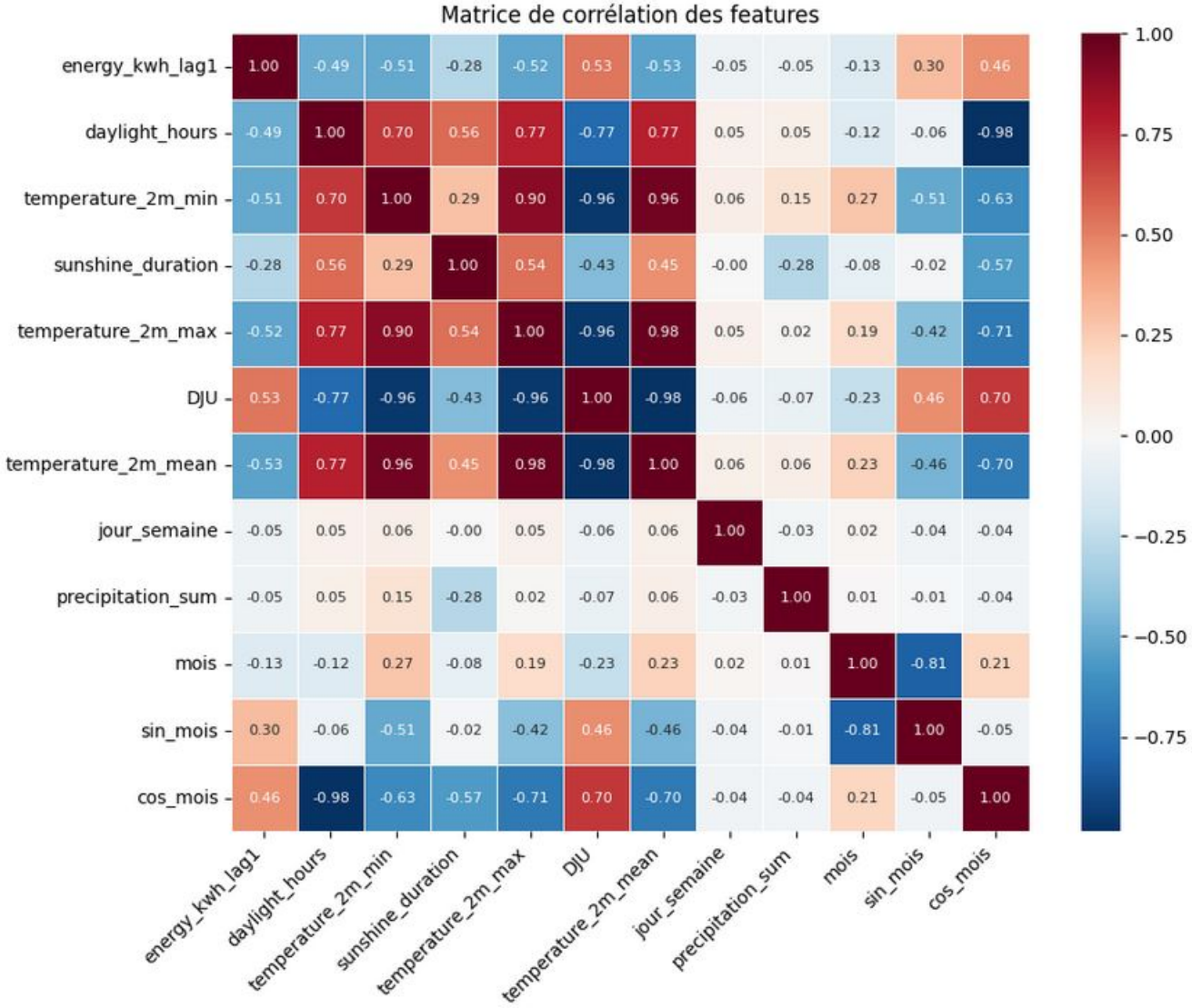

FIGURE 7 – Matrice de corrélation des variables (Pearson) sur le jeu filtré.

RÉFÉRENCES

[1] UCI Machine Learning Repository, ≪ Individual household electric power consumption data set. ≫ [En ligne]. Disponible : https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption

[2] Open-Meteo, ≪ Historical Weather API. ≫ [En ligne]. Disponible : https://open-meteo.com/en/docs/historical-weather-api

[3] Bibliothèque Python `holidays`. [En ligne]. Disponible : https://pypi.org/project/holidays/

[4] scikit-learn : Machine Learning in Python. [En ligne]. Disponible : https://scikit-learn.org/