**Projet Energie**



Projet de Data Analyse

Formation DataScientest

Session septembre 2023



Sommaire

[**Introduction au projet 4**](#_bk5f2v2lqmy7)

[Le contexte 4](#_3drz3jitisj7)

[Du point de vue technique 4](#_98xty8v1i31l)

[Du point de vue économique 4](#_mxt3451cnic2)

[Du point de vue scientifique 5](#_rka14bma0809)

[Les objectifs 5](#_9staddi1idqj)

[Quels sont les principaux objectifs du projet ? 5](#_vzf02vd2wsk3)

[Quel est le niveau de compétence de l’équipe sur le sujet ? 5](#_5zwbnfj7zl9p)

[Enzo 5](#_dp8dbvrtkll9)

[Damien 6](#_c45f8lvxoka5)

[Charlène 6](#_1rd99hiqu19u)

[Pacôme 6](#_hx8noi912xy7)

[**Compréhension et manipulation des données 7**](#_iobofzq4td1u)

[Le cadre 7](#_qz3lamgjy8)

[Dataset éCO2mix 7](#_hmnbw0y891rp)

[Dataset températures régionales 7](#_xswjrvnmywmg)

[Dataset températures nationales 7](#_qj50pq87ubhq)

[La pertinence 8](#_pajxhrko23x)

[Les variables les plus pertinentes 8](#_u2cmseunpeph)

[Quelle est la variable cible ? 9](#_c9dbxovndyus)

[Les particularités du jeu de données 9](#_lzptx1vq4vrg)

[Les limites du jeu de données 9](#_bltl7dd3n2h6)

[Pre-processing et feature engineering 10](#_wfkep12je3cb)

[Visualisations et Statistiques 10](#_hu1oiqdfhzu2)

[Relations entre les différentes variables 10](#_xkcf5dqjlo9e)

[Relations entre les variables explicatives 11](#_6hkebt997q1l)

[Relations entre les variables explicatives et la variable cible 12](#_1tlgd9z7akkw)

[Quel lien entre consommation et production d’énergie ? 12](#_qd9ph6cwoxmd)

[Quelle est la répartition de la production d’électricité des régions ? 13](#_glxt6nlxddwo)

[Quelle est la spécialité de chaque région ? 16](#_wgh1dwkzovg0)

[Comment est répartie la production d’électricité en France ? 17](#_97i65z89q5ws)

[Quelle est la proportion d’énergies renouvelables dans la production française ? 18](#_2hw3m5uc0zxj)

[Les mois ont-ils une influence sur la consommation d’électricité ? 18](#_gb8u2vqnwoqo)

[Les jours et les horaires ont-ils une influence sur la consommation d’électricité ? 19](#_c6gd4loedvpu)

[Distribution des données 20](#_flbswvdvy39k)

[Traitement des outliers 21](#_9p0ntbemyvzj)

[Analyses statistiques 22](#_izwxe74h3fal)

[1er analyse statistique 23](#_8wlegydvow5d)

[2ème analyse statistique 23](#_m4n5t09t6mhz)

[**Introduction à la modélisation 25**](#_pjdz5h9ediyo)

[À quel type de problème de machine learning votre projet s’apparente-t-il ? 25](#_f8yrlmoei87s)

[À quelle tâche de machine learning votre projet s’apparente-t-il ? 25](#_av2v5jq5c2h6)

[Quelle est la métrique de performance principale utilisée pour vos modèles ? 25](#_b2zs7k3jynp2)

[Avez-vous utilisé d’autres métriques de performances ? 25](#_d0mbbx39l2ci)

[**La modélisation au niveau national 26**](#_ee6qn3wehq1n)

[Le modèle de classification 26](#_d9ht4pbaqemj)

[Le choix du modèle de régression 28](#_qct63dxhqhi2)

[Quel algorithme avez-vous retenu et pourquoi ? 31](#_din1qgiqswfz)

[Avez-vous utilisé des techniques d’optimisation de paramètres ? 31](#_p4f4kp4g53zn)

[Avez-vous testé des modèles avancés ? 32](#_lmzwfslljlmu)

[Interprétation des résultats 32](#_57nxkvuirh0p)

[Avez-vous analysé les erreurs de votre modèle, avec quelles améliorations ? 32](#_v4mygj16hodw)

[Avez-vous utilisé des techniques d’interprétabilité ? 33](#_hdcbaark5rkx)

[Qu’est ce qui a (ou non) engendré une amélioration de vos performances ? 33](#_onqrpl7fctx)

[**La modélisation au niveau régional 34**](#_869shp8zy9iw)

[Choix du modèle et optimisation 34](#_iwnmmfeue39i)

[Quels algorithmes avez-vous essayés ? 34](#_l2ve9f5mezfq)

[Quel algorithme avez-vous retenu et pourquoi ? 36](#_vrfcfg5g63n)

[Avez-vous utilisé des techniques d’optimisation de paramètres ? 37](#_ckvndnojp6er)

[Avez-vous testé des modèles avancés ? 37](#_8jgagxhryhxk)

[Interprétation des résultats 38](#_jfl588fhglgn)

[Avez-vous analysé les erreurs de votre modèle, avec quelles améliorations ? 38](#_6x2zhw1hu335)

[Avez-vous utilisé des techniques d’interprétabilité ? 38](#_dct3k5swk8tv)

[Qu’est ce qui a (ou non) engendré une amélioration de vos performances ? 40](#_nqinjmyrwioi)

[Des prédictions pour 2022 et 2023 40](#_myy0i40hlt)

[Des prédictions sans machine learning 41](#_rwp0a9rquh8k)

[Conclusion 43](#_iap7vud2c7g1)

[Bibliographie 46](#_3bvr14gyefa1)

[Table des pièces jointes 47](#_yb4hl26611k4)

# 

# Introduction au projet

L'acquisition des [données éco2mix régionales](#_3bvr14gyefa1), couvrant la période de janvier 2013 à mai 2022, s'avère essentielle dans le contexte de notre projet d’étude, visant à améliorer la gestion de l'énergie. Ces données offrent une fenêtre précieuse pour analyser les implications techniques, économiques et scientifiques de la consommation et de la production d'énergie, tant au niveau national qu'au niveau régional. Notre objectif principal est de mieux comprendre le phasage entre la consommation et la production énergétique, avec un accent particulier sur la prévention des risques de black-out, en se penchant sur les dynamiques régionales et les filières de production, notamment les énergies nucléaires et renouvelables.

## Le contexte

Ce projet sur l'étude de la consommation d’énergie a été mené dans le cadre de la formation data analyst de Datascientest, session de septembre 2023, avec quatre participants : Damien Lioret, Enzo Tournefier, Charlène Vince, Pacôme Bouadou.

### Du point de vue technique

D'un point de vue technique, l'accès aux données éCO2mix régionales constitue un atout majeur pour notre étude. Ces données nous permettent de scruter de près la corrélation entre la consommation et la production d'énergie aux niveaux national et régional. Cela revêt une importance cruciale pour anticiper et prévenir les situations de black-out. Grâce à ces données, nous sommes en mesure d'effectuer une analyse fine au niveau régional, nous permettant de mieux comprendre les variations saisonnières et géographiques de la consommation d'énergie.

### Du point de vue économique

Sur le plan économique, l'accès aux données éco2mix régionales offre des perspectives inestimables. Ces données nous permettent d'évaluer les coûts associés à la production d'électricité dans différentes régions, ce qui est essentiel pour prendre des décisions stratégiques. En identifiant les filières de production les plus efficaces, en tenant compte des fluctuations de la demande, des subventions et des incitations fiscales, l’entreprise pourra optimiser ses investissements dans les infrastructures énergétiques. De plus, cette connaissance approfondie permettrait de mieux anticiper les coûts de l'électricité et de proposer des tarifs compétitifs à leurs clients, tout en contribuant à la transition vers des sources d'énergie plus durables.

### Du point de vue scientifique

Enfin, d'un point de vue scientifique, l'analyse des données éco2mix régionales représente une opportunité pour approfondir notre compréhension des énergies renouvelables. Ces données nous permettent de cartographier précisément l'implantation des sources d'énergie renouvelable, offrant ainsi une vision claire des régions où elles sont les plus présentes. Cela a un impact significatif sur la démarche de développement durable, permettant aux entreprises de favoriser davantage l'utilisation d'énergies renouvelables là où elles sont les plus adaptées. Cette approche contribue à réduire l’empreinte carbone et à promouvoir une gestion plus responsable des ressources énergétiques.

En résumé, l'intégration des données éco2mix régionales dans notre projet d’étude revêt une importance majeure. Ces données techniques, économiques et scientifiques aident les entreprises à mieux gérer l'énergie au niveau national et régional, à prévenir les risques de black-out et à promouvoir la transition vers des sources d'énergie plus durables, tout en optimisant leurs coûts et en renforçant leur compétitivité.

## Les objectifs

### Quels sont les principaux objectifs du projet ?

* Analyser le phasage entre la consommation et la production énergétique au niveau national et régional, avec un accent sur la prévention des risques de black-out.
* Réaliser une analyse au niveau régional pour déduire des prévisions de consommation.
* Analyser les différentes filières de production d'énergie, en mettant particulièrement l'accent sur l'énergie nucléaire et les sources d'énergie renouvelable.

# Compréhension et manipulation des données

## Le cadre

### Dataset éCO2mix

Nous utiliserons les [données éco2mix régionales](#_3bvr14gyefa1) qui comprennent des informations sur la consommation d'électricité ainsi que la production par filière (énergie nucléaire, énergies renouvelables, etc.) sur une base quotidienne, toutes les 1/2 heures, depuis 2013.

Ces données sont mises à disposition librement, ce qui signifie qu'elles sont accessibles au public sans restriction particulière. Elles sont destinées à être utilisées par un large éventail de parties prenantes, y compris les chercheurs, les entreprises du secteur de l'énergie, et les décideurs publics.

L'ORDRE (Open Data Réseaux Energies) est l'entité qui gère et distribue ces données. Il s’agit d’une collaboration entre les principaux producteurs de l’énergie français, ainsi que des entreprises qui gèrent son transport dans tout le pays.

En ce qui concerne la volumétrie de ces données, la fréquence de collecte toutes les 1/2 heures depuis 2013 génère un ensemble de données substantiel. La volumétrie totale représente : 1 980 288 lignes et 32 colonnes.

### Dataset températures régionales

Nous envisageons également d’ajouter des données climatiques à notre jeu de données initial. Nous pensons en effet que pour prédire au mieux la consommation par région, nous aurons besoin de données de températures. En effet, en France, une grande partie du chauffage domestique est réalisée à l'aide de radiateurs électriques. Plus il fait froid, plus les gens ont besoin de chauffer leur domicile, ce qui entraîne une augmentation significative de la consommation d'énergie. Nous avons donc choisi un jeu de données également fourni par l'ORDRE selon les mêmes conditions de diffusion : [les températures quotidiennes régionales](#_3bvr14gyefa1).

Ce fichier qui comprend 36 712 lignes sur 6 colonnes met à disposition des données de température par région entre 2016 et 2023. Il est mis à jour chaque mois et contient donc des informations très récentes.

### Dataset températures nationales

En complément du jeu de données des températures régionales (décrit ci-dessus), nous ajoutons un jeu de données climatiques à notre jeu de données initial, dans le but de faire des prévisions de consommation et de production à l’échelle nationale et sur l’entièreté de la période étudiée sur notre jeu de données initial, soit entre 2013 et 2022. Ce jeu de données est fourni par la même source que les jeux précédents. Ce fichier renseigne des données de température journalière au niveau national. Ce jeu de données comporte 4 colonnes (Date, Pic journalier consommation (MW), Température moyenne (°C), Température référence (°C)) et des données en lignes allant du 1er janvier 2012 au 31 août 2023. [Voir le jeu de données de températures au niveau national](#_3bvr14gyefa1)

## La pertinence

### Les variables les plus pertinentes

Au regard de nos objectifs, certaines variables se révèlent particulièrement pertinentes pour notre projet. Voici les variables que nous considérons comme étant les plus importantes :

* Consommation d'électricité : La consommation d'électricité est une variable cruciale pour comprendre les besoins énergétiques des régions et pour anticiper les variations de demande, notamment pour prévenir les risques de black-out.
* Production d'électricité par filière : Il est essentiel de surveiller la production d'électricité par filière, en mettant l'accent sur les sources d'énergies nucléaire et renouvelables. Cela nous permet de déterminer la contribution de chaque source à la production totale et d'identifier les régions où elles sont implantées.
* Géolocalisation par région : La localisation des infrastructures de production d'électricité est d'une grande importance pour comprendre les particularités régionales et identifier les zones à haut risque de déséquilibre entre la production et la consommation.
* Données temporelles : La fréquence des données toutes les 1/2 heures permet une analyse détaillée des fluctuations de la consommation et de la production tout au long de la journée et de l'année, ce qui est essentiel pour anticiper les besoins.

En plus des variables présentes dans notre jeu de données d’origine, nous avons identifié des données de températures à ajouter. Pour pouvoir prédire la consommation au niveau national, nous avons ajouté un jeu de données de températures nationales, et nous avons procédé de la même manière au niveau régional avec un dataset de températures différent.

### Quelle est la variable cible ?

La variable cible de notre projet est la consommation d'électricité. Si on reprend la liste de nos objectifs, nous devons réaliser une analyse approfondie au niveau régional et au niveau national pour déduire des prévisions de consommation par jour.

### Les particularités du jeu de données

Les particularités de notre jeu de données comprennent la haute résolution temporelle (toutes les 1/2 heures), la granularité régionale (couvrant diverses régions), et la disponibilité sur une longue période (de janvier 2013 à mai 2022).

### Les limites du jeu de données

Cependant nous avons constaté quelques incohérences dans le jeu de données listées ci-dessous

* Les colonnes thermique (MW), Nucléaire (MW), Éolien (MW), Solaire (MW), Hydraulique (MW) rassemblent des données de production d'énergie, pourtant elles contiennent de nombreuses valeurs négatives. Nous avons essayé d’entrer en contact par mail avec l’un des producteurs des données. Seul RTE (Réseau de transport d'électricité) a répondu brièvement à certaines de nos questions et selon eux, il pourrait s’agir du soutirage[[1]](#footnote-1) d’installations de production et donc de valeurs normales.
* Les variables Stockage batterie et Déstockage batterie possèdent de nombreuses données négatives et la deuxième comprenant des valeurs souvent négligeables. Nous nous sommes interrogés sur la qualité des données de ces deux colonnes. Et toujours selon notre contact chez RTE, une erreur d’ajout de colonnes a eu lieu sur la période 2021-2022, les variables Ech. physiques (MW), Stockage batterie, Déstockage batterie, Éolien terrestre et Éolien offshore sont donc fausses sur cette période.
* Les colonnes Éolien terrestre et Éolien offshore devraient détailler la colonne Éolien (MW). Pourtant les chiffres n'ayant souvent aucun rapport entre eux, nous nous sommes interrogés sur la définition de ces deux colonnes. Notre contact chez RTE a confirmé qu’il y avait bien un problème avec ces deux colonnes (voir ci-dessus).
* Les Taux de CHarge (TCH) par filière étant manquants pour une grande partie du fichier, nous envisageons de l’exclure de nos données. De même que les Taux de Couverture (TCO) que l’on peut déduire par calcul à partir de deux autres variables.

## Pre-processing et feature engineering

Le pré-traitement des données est une étape importante dans tout projet d'analyse de données. Voici les premières étapes :

* Suppression de la dernière colonne qui est vide
* Suppression des 12 premières colonnes qui ne possèdent que des NaN
* Suppression des colonnes : TCO, TCH, nature, Ech. physiques, Stockage batterie, Déstockage batterie, Éolien terrestre, Éolien offshore, date - heure
* Suppression de la colonne région, mais conservation de la colonne Code INSEE
* Conversion de la variable Date en datetime
* Création de colonnes jour, mois, année et suppression de la colonne Date
* Conversion des NAN en 0 dans les colonnes : nucléaire, hydraulique, éolien, solaire, bioénergie
* Groupby par jour afin de faire nos prédictions sur une journée
* Création de deux nouveaux jeux de données qui fusionnent des données de températures : au niveau national (période 2013-2022) et régional (période 2016-2022) sur des périodes différentes suivant les données que nous avons pu récupérer.
* Sur nos deux jeux de données nous allons réaliser un Standard Scaler pour uniformiser toutes nos données numériques

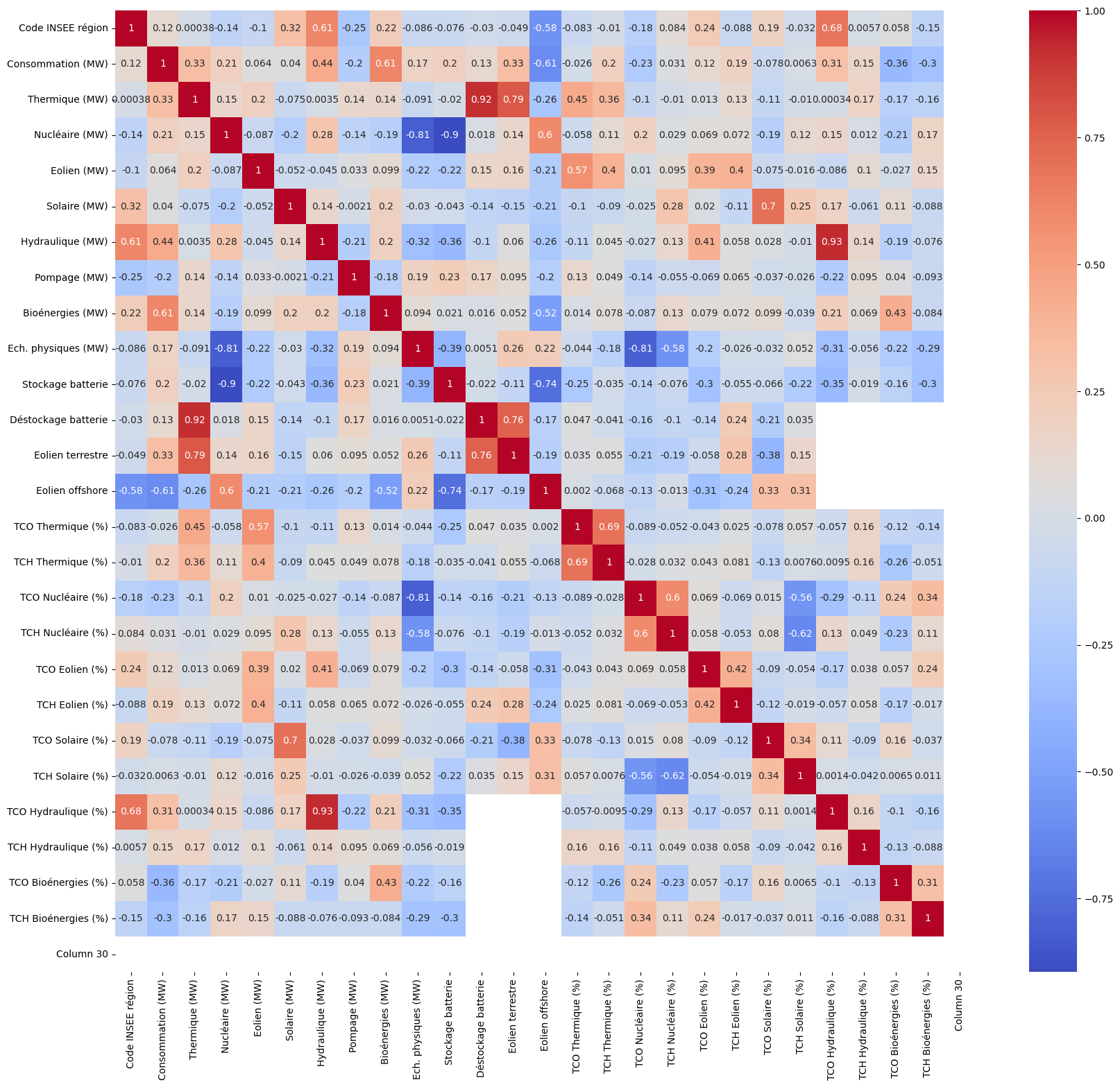
Nous n’envisageons pas de techniques de réduction de dimension dans la partie de modélisation, notre nombre de colonnes étant assez restreint.

## Visualisations et Statistiques

Les analyses visuelles et statistiques sont essentielles pour comprendre les relations entre les variables, identifier des tendances, et préparer les données pour la modélisation. Voici un aperçu des analyses que nous avons menées dans le cadre de notre projet.

### Relations entre les différentes variables

Nous avons effectué des analyses de corrélation pour comprendre les relations entre différentes variables, telles que la consommation d'électricité, la production par filière, la géolocalisation, et d'autres caractéristiques. Si la matrice de corrélation (ci-dessus) nous a permis d’identifier certaines relations, aucune des données de notre premier jeu de données n’était liée à la variable cible.



### Relations entre les variables explicatives

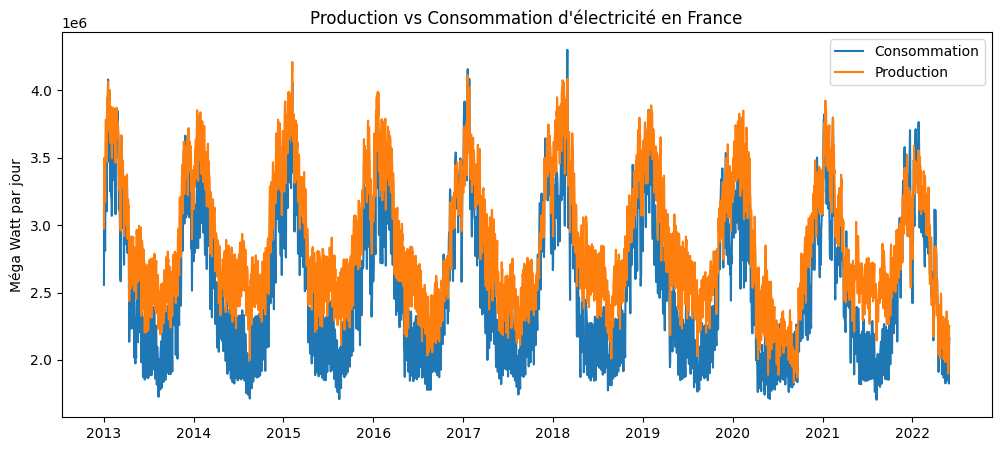
Ayant très peu de variables explicatives, hormis les lieux et dates, nous n’avons pu en déduire des relations.

### Relations entre les variables explicatives et la variable cible

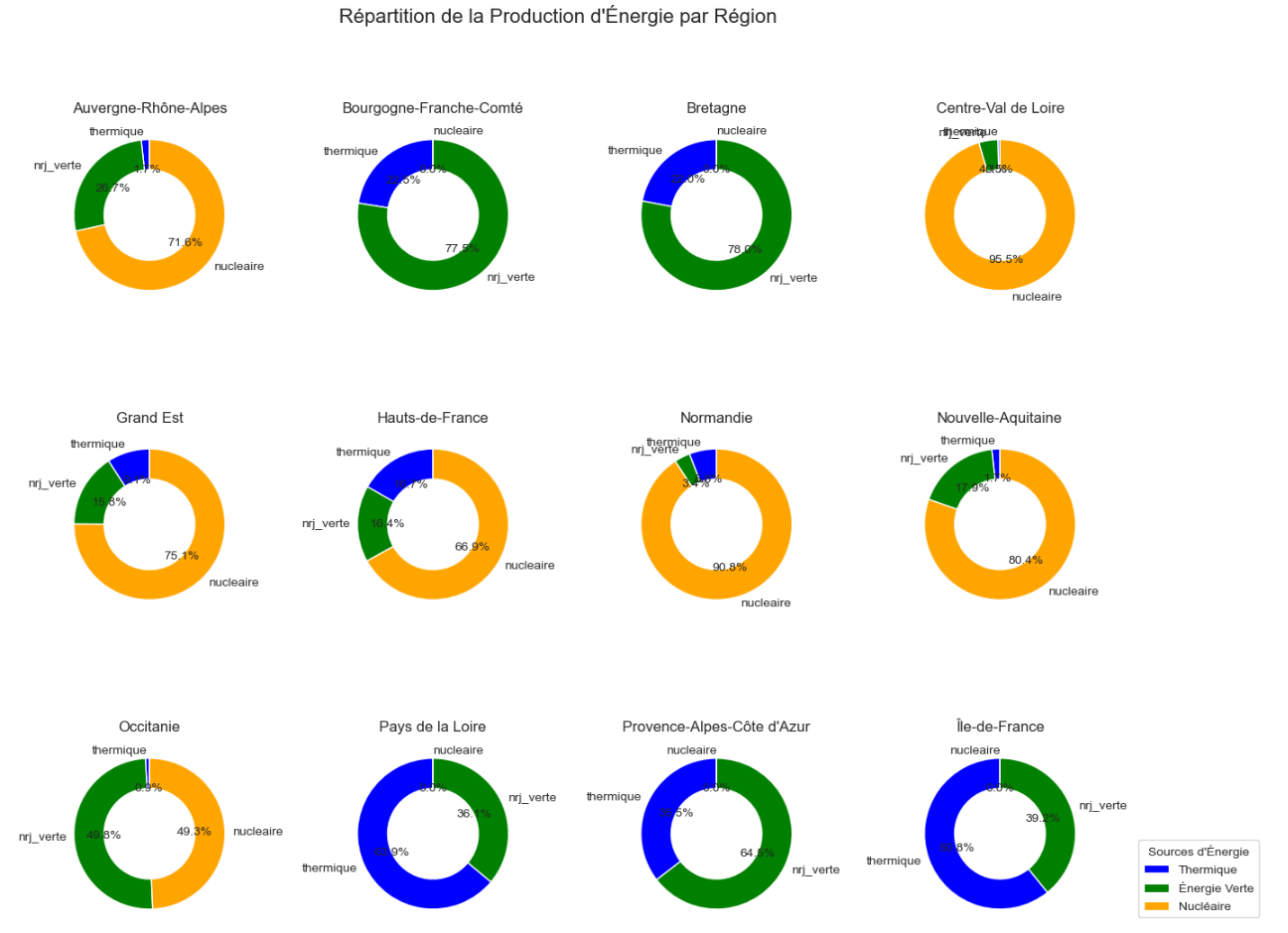
Nous avons étudié comment la consommation d'électricité était influencée par les régions, les heures, les jours, les mois.

#### Quel lien entre consommation et production d’énergie ?

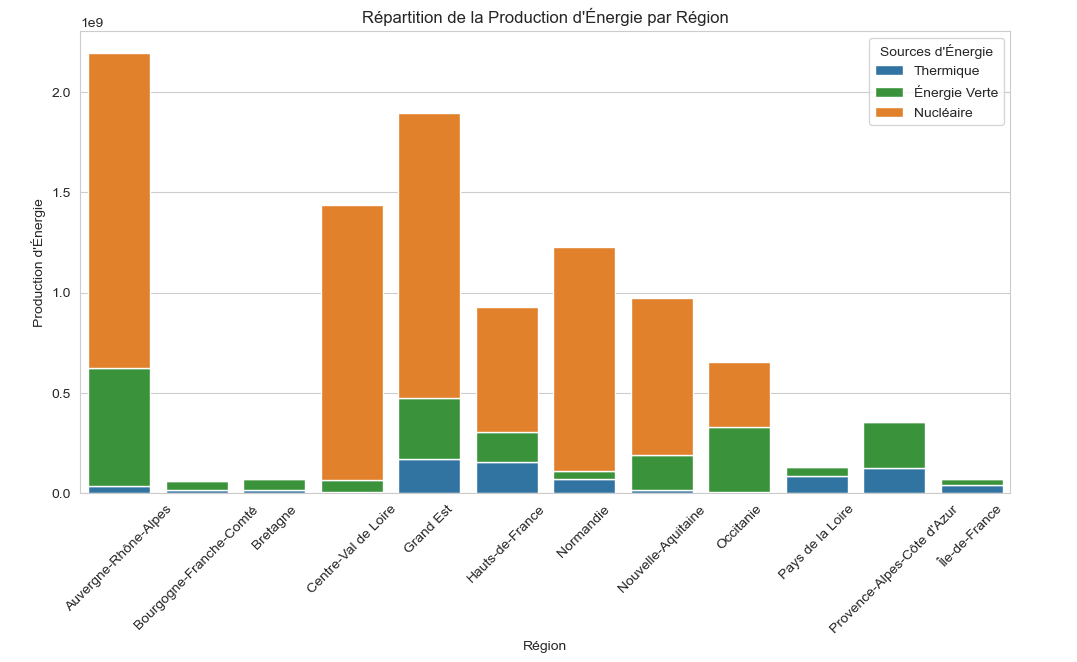
A partir du graphique ci-dessous, on observe en premier lieu que consommation et production d'énergie sont liées. Plus les habitants consomment, plus le pays produit de l'électricité. Ce graphique permet aussi d'appréhender la saisonnalité du couple production-consommation d'énergie. En hiver, les Français consomment plus, notamment à cause du chauffage, en conséquence la production d'électricité est plus importante. L'été, le constat est inversé : la population consomme moins, donc la France produit moins. On remarque enfin que la production d'électricité est en baisse durant l'été 2020 et durant l'hiver 2022, tandis que la consommation se maintient à peu près au même niveau, chaque année.



#### Quelle est la répartition de la production d’électricité des régions ?

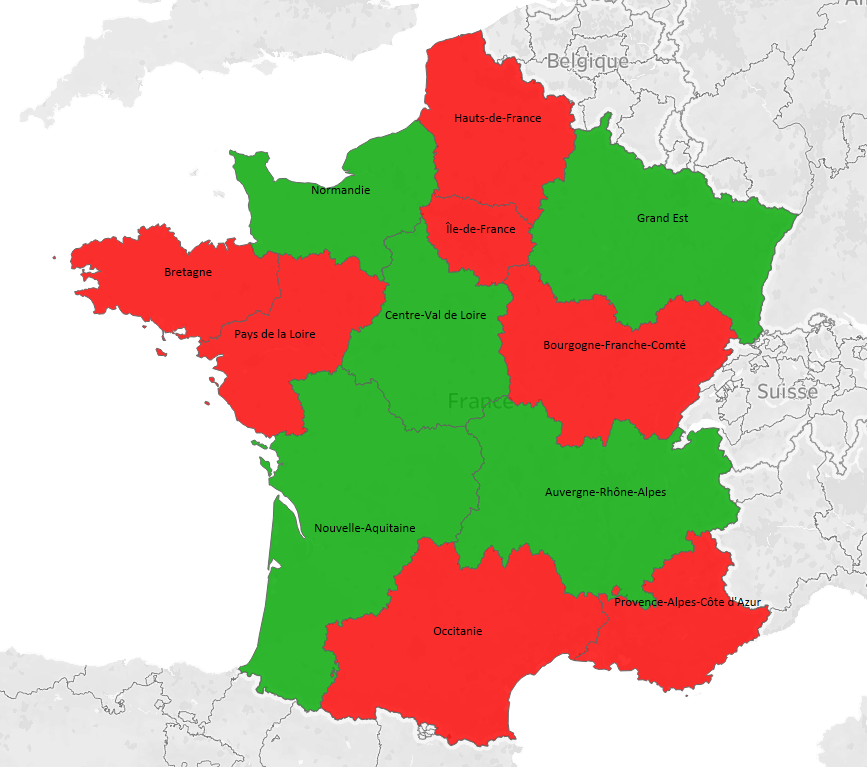


A l’aide du graphique ci-dessus, on observe que la production d’énergie est complètement différente selon les régions, certaines régions produisent de façon quasi exclusive de l'électricité à partir d’énergie renouvelable.



Avec le graphique ci-dessus on remarque que l’électricité produite par les énergies vertes est largement en dessous de la quantité d’électricité produite par le nucléaire. On peut donc se poser la question suivante : est-ce que les régions qui produisent en majorité de l’électricité verte sont autonomes (c’est-à-dire produisent-elles assez d'électricité pour subvenir à leurs besoins de consommation) ? La région PACA par exemple, produit-elle assez d'électricité pour sa population ?

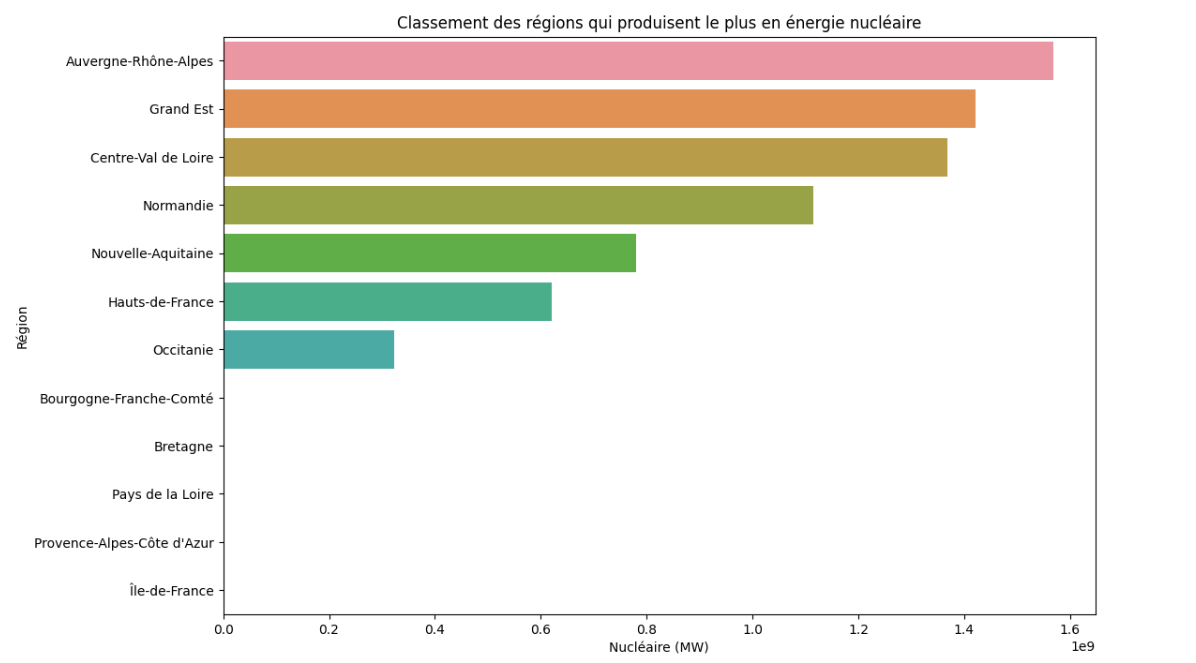
La carte ci-dessous qui vient compléter le graphique précédent, permet de répondre à nos questions. Les régions qui produisent seulement de l'électricité verte ne sont pas autonomes, même en couplant cette production des énergies fossiles. Seule une importante production d’énergie nucléaire permet à des régions comme le Grand-Est ou le Centre-Val-de-Loire d'être indépendantes énergiquement parlant.



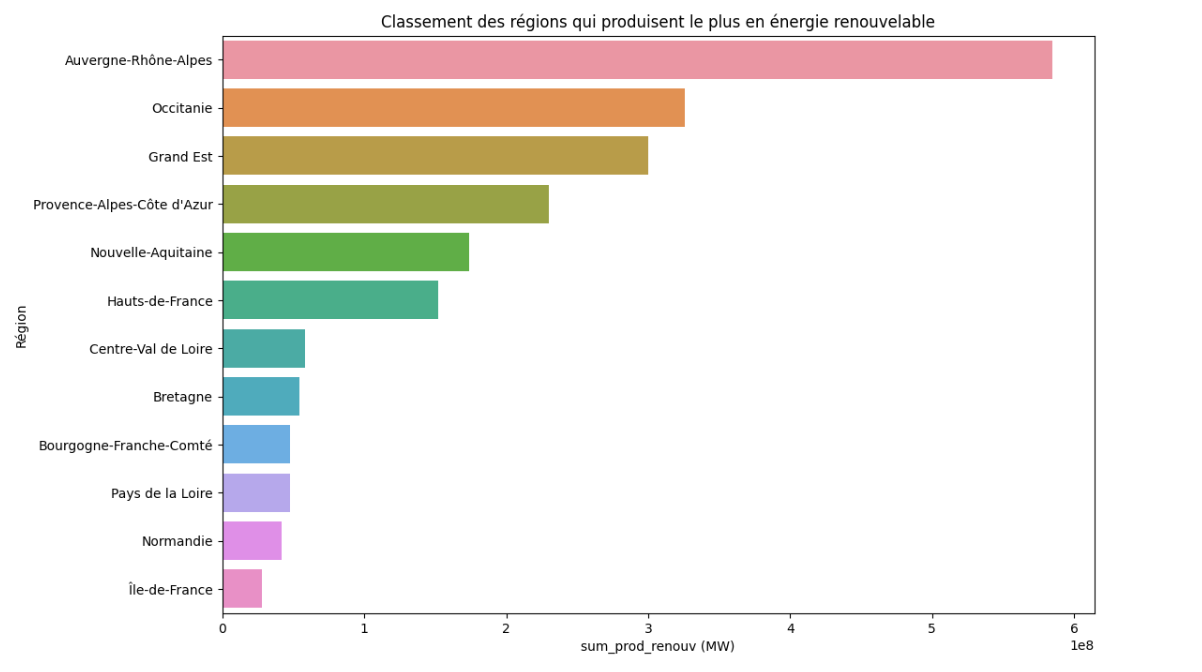
Carte des régions qui produisent plus qu’elles ne consomment

Avec ces séries de graphiques, on peut conclure que pour l’instant la production d'électricité issue d'énergies vertes ne suffit pas à couvrir les besoins de consommation par région. En revanche, les échanges d’énergie entre régions permettent de couvrir les besoins nationaux de la France.

#### Quelle est la spécialité de chaque région ?

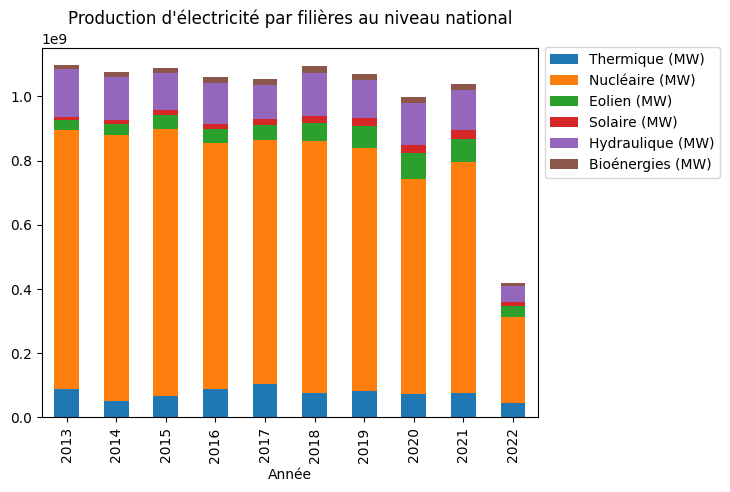


Dans ce graphique qui représente le classement des régions qui produisent le plus d’énergie nucléaire, on remarque que la région Auvergne-Rhône-Alpes est la région la plus productrice d'énergie nucléaire, on recense sur cette région 3 des 8 plus grandes centrales nucléaires de l'hexagone. Nous constatons que 3 autres régions sont très productrices d'énergie nucléaire, le Grand Est, le Centre Val de Loire et la Normandie avec 2 à 3 centrales nucléaires par région évoquée. Enfin, nous observons que 5 régions ne produisent pas d'énergie nucléaire (Bourgogne-Franche-Comté, Bretagne, Pays de la Loire, Ile de France, Provence-Alpes-Côte d'Azur).



Dans ce graphique qui représente le classement des régions qui produisent le plus d’énergie renouvelable, la région la plus productrice d'énergie renouvelable est la région Auvergne-Rhône-Alpes avec plus de 50% de production (en MW) que tout autre région, notamment grâce à l'énergie hydraulique. En suivant, nous observons que 2 régions sont sur des valeurs équivalentes, l'Occitanie et le Grand-Est, pour la première grâce à la production d'énergie hydraulique et solaire, pour le second grâce au développement de plusieurs types d'énergie renouvelable. Enfin, nous retrouvons des régions qui produisent très peu d'énergies renouvelables en comparaison avec celles évoquées, la Bourgogne-Franche-Comté, la Bretagne, le Centre Val de Loire, les Pays de la Loire, l'Ile de France et la Normandie.

#### Comment est répartie la production d’électricité en France ?

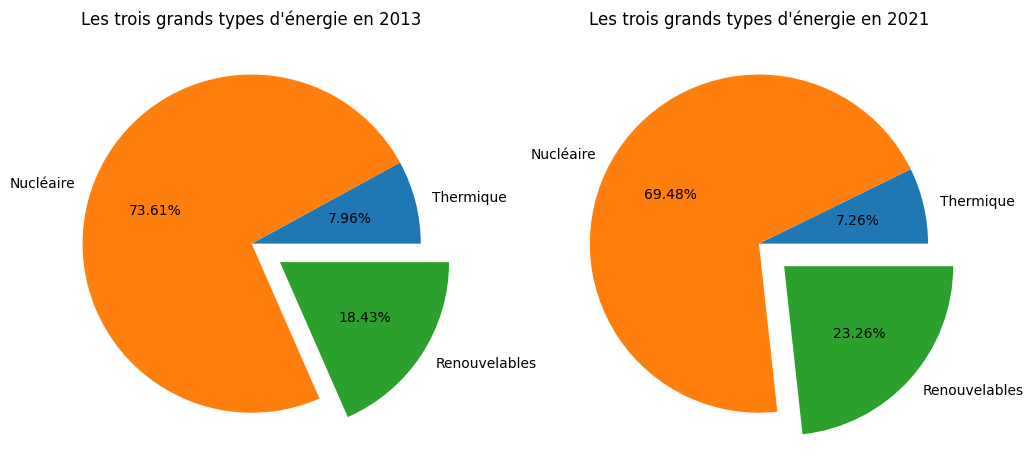


Sur ce graphique, on remarque que la principale source de production d'électricité en France est le nucléaire. Le nucléaire représente 72 % de la production d'énergie de notre pays sur la période étudiée. Viennent ensuite la production hydraulique et la production thermique des énergies fossiles.

On observe également que la part de la production nucléaire baisse, tandis que celles des énergies renouvelables augmentent entre 2013 et 2021. Comme vu sur le graphique précédent, la production a particulièrement baissé durant l'été 2020.

Attention, nous ne pouvons pas nous fier à la colonne 2022, car le dataset s'arrête au mois de mai 2022.

#### Quelle est la proportion d’énergies renouvelables dans la production française ?



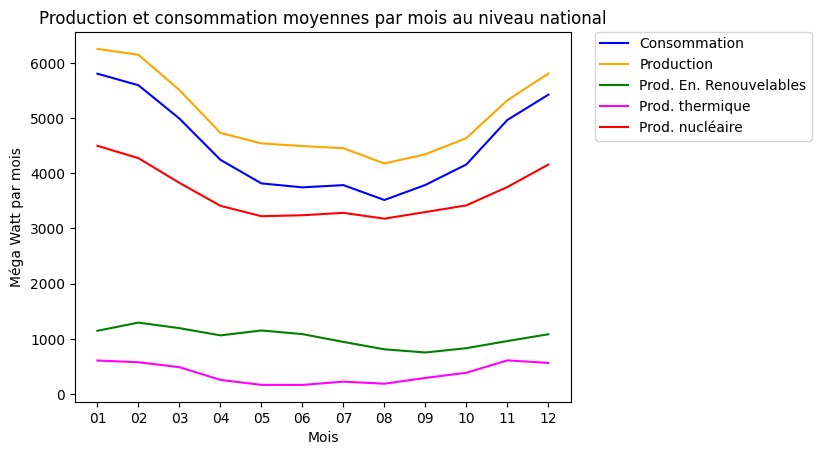
Sur ce focus, qui vient compléter le graphique précédent, on observe que la part des énergies renouvelables dans la production d'électricité en France a augmenté de près de 5% entre 2013 et 2021, et cela au détriment de la production nucléaire.

Dans les énergies renouvelables, nous regroupons : l'éolien, le solaire, l'hydraulique et les bioénergies. Et pour rappel, voici la définition des bioénergies : cette filière comprend les biogaz, les biocombustibles solides (bois énergie et autres biocombustibles solides) et les déchets ménagers et papetiers.

#### Les mois ont-ils une influence sur la consommation d’électricité ?

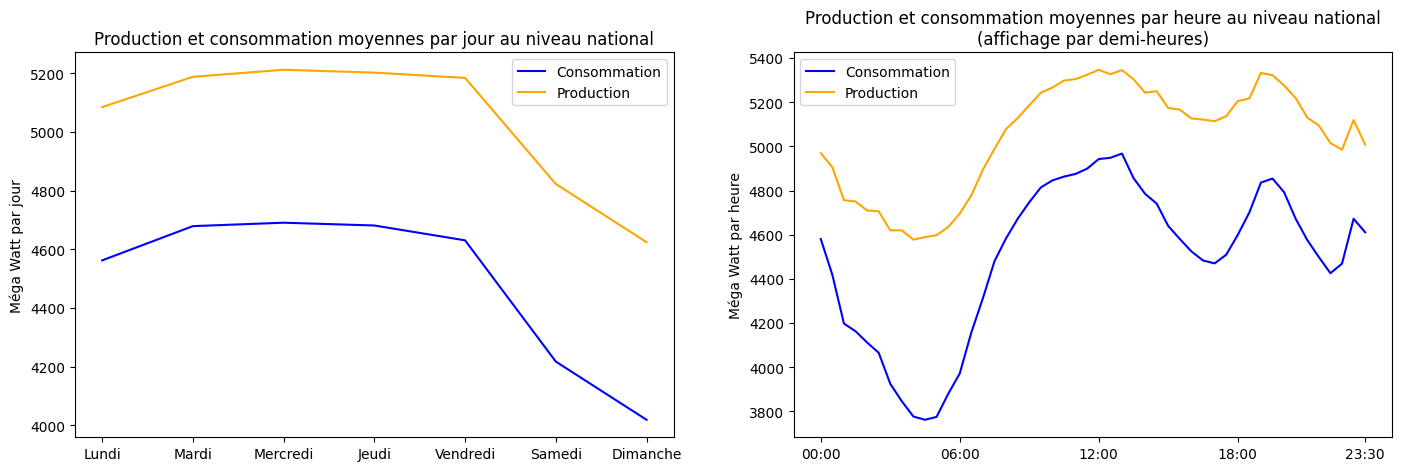
Ce graphique permet de bien observer la saisonnalité de la consommation d'électricité. Sans surprise, on consomme plus en décembre, janvier, février, qui sont les mois d’hiver, en raison de l’utilisation du chauffage. A l’inverse, le mois d’août est le mois où la consommation est la plus basse.

On observe aussi que la production nationale d'électricité semble toujours supérieure à la consommation, donc peu de risque de black-out. Nous savons que la France vend notamment de l'énergie à ses voisins.



On peut également en déduire que le pays adapte sa production d'électricité aux besoins de ses habitants. Tandis que les énergies renouvelables ont une production à peu près stable sur les différents mois de l'année, c'est la production d'énergie nucléaire (ou thermique) qui permet cette adaptabilité.

#### Les jours et les horaires ont-ils une influence sur la consommation d’électricité ?

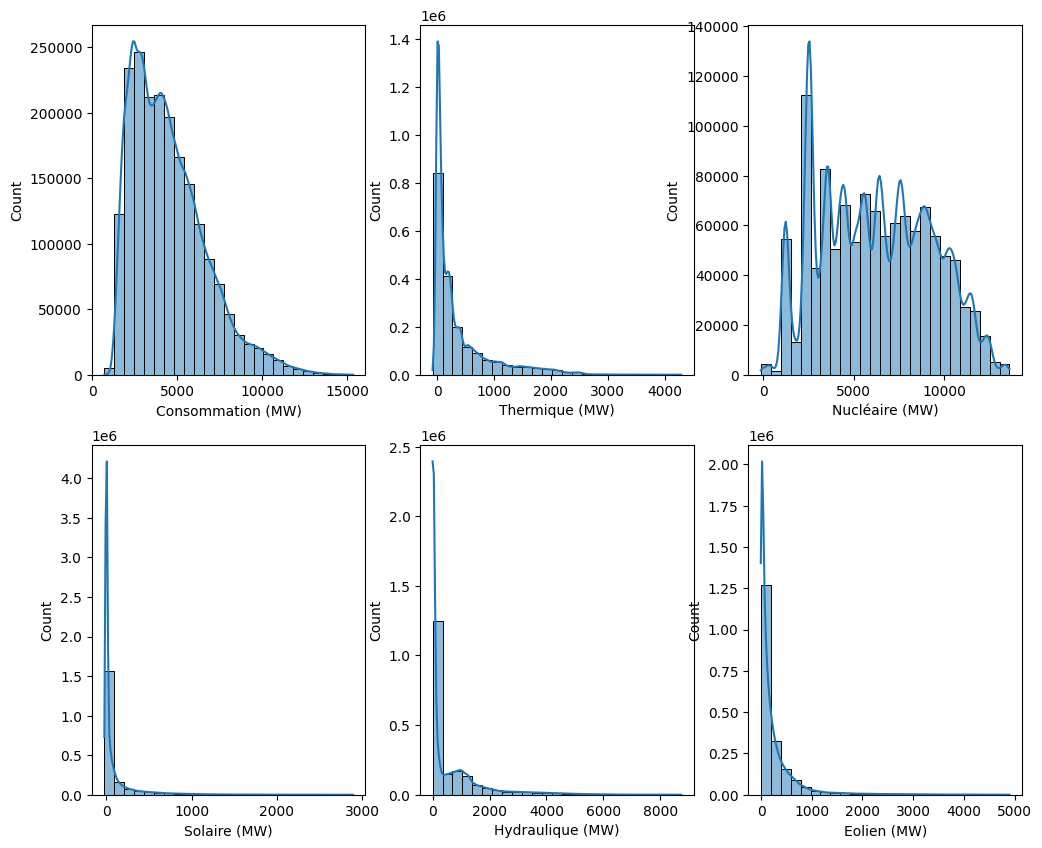


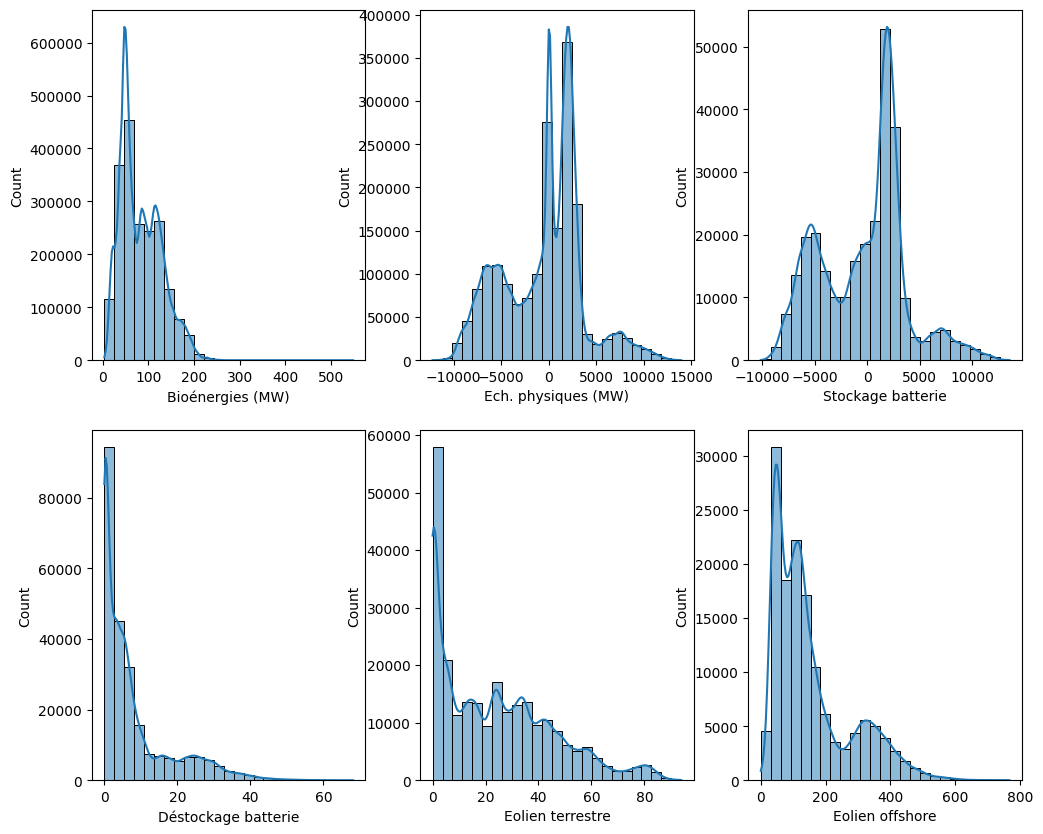
A l’aide de ces deux graphiques, on observe que le couple consommation-production dépend également des jours de la semaine. Ainsi, la France consomme moins d'électricité les week-ends et particulièrement le dimanche. Cela peut s'expliquer par le fait que le dimanche est pour la majorité des Français un jour non travaillé et par conséquent de nombreuses entreprises sont à l'arrêt, consommant moins.

Sans surprise également, les pics de consommation d’électricité se situent entre 11h et 13h, puis entre 19h et 20h. Ce qui correspond aux périodes de préparation des repas. En revanche, les horaires où la France consomme le moins se situent entre 4h et 5h, le matin.

### Distribution des données

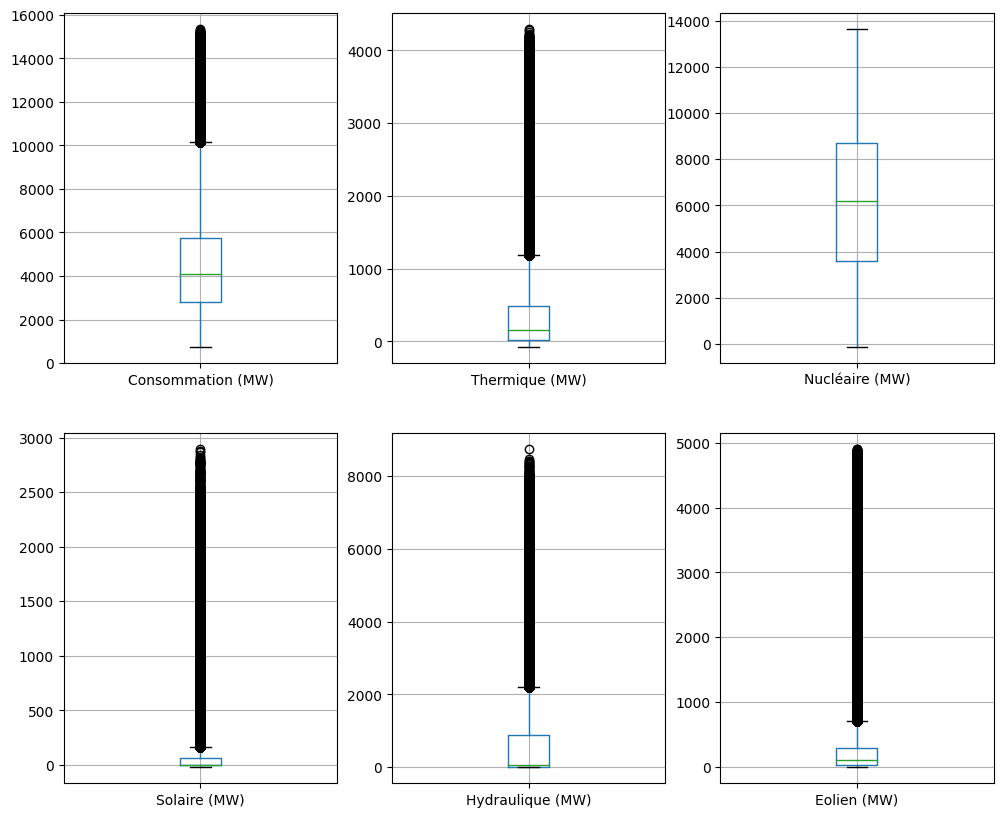
Nous avons examiné la distribution des données, en utilisant des histogrammes et des diagrammes de boîtes pour identifier la répartition des valeurs, la présence d'outliers, et les pics de consommation ou de production.



******

### Traitement des outliers

Nous avons décidé de conserver les outliers qui correspondent souvent à des pics de consommation ou de production d’énergie. Concernant les valeurs aberrantes que nous avons pu relever, c’est-à-dire les valeurs de production négatives d’énergie. Selon notre contact chez RTE (Réseau de transport d'électricité), il pourrait s’agir du soutirage d’installations de production et donc de valeurs normales. Pour ce qui est des erreurs de saisie signalées par ce même contact, nous avons choisi de nous séparer des variables incluant ces erreurs.



### Analyses statistiques

Pour confirmer nos observations visuelles, nous allons utiliser des tests statistiques tels que des tests de signification (t-tests, ANOVA) pour évaluer la relation entre les variables et des modèles de régression pour quantifier les effets des variables explicatives sur la variable cible.

À partir de ces analyses, nous allons tirer plusieurs conclusions importantes qui orientent notre phase de modélisation à savoir si :

Les énergies renouvelables ont une corrélation positive avec la consommation d'électricité dans certaines régions, indiquant une augmentation de la demande liée à ces sources d'énergie.

Les variations saisonnières ont une forte influence sur la consommation d'électricité, ce qui nécessitera une modélisation adaptée.

Pour valider nos visualisations, nous allons procéder à deux tests statistiques. Le premier sur la corrélation entre les variables région et la production d’énergie. Dans le second test, nous allons nous intéresser à l’influence de la temporalité, dans notre étude il s’agit des mois, et la consommation d’énergie.

#### 1er analyse statistique

Choix des hypothèses :

Nous souhaitons savoir s'il existe une corrélation entre la variable région et la variable production, le but étant de savoir si la production est différente selon une localisation en France, la région. Donc 2 hypothèses se dégagent.

Les hypothèses sont :

* H0 : La région n'influence pas la production d'énergie
* H1 : La région influence la production d'énergie

Le choix du test :

Nous allons choisir le test Anova puisque nous devons étudier une interaction entre une variable catégorielle et une variable continue.

Résultat : p value = 0.00085 donc p value < 0,05

L'hypothèse H1 est validée. On peut conclure que la géolocalisation des points de production a un impact significatif sur la production d'électricité. Ces conclusions servent de base pour la phase de modélisation, où nous développerons des modèles prédictifs pour anticiper la consommation d'électricité en fonction des différentes variables explicatives.

#### 2ème analyse statistique

Choix des hypothèses :

Nous souhaitons savoir s'il existe une corrélation entre la variable date (mois) et la variable consommation, le but étant de savoir si la consommation est différente selon le mois de l'année. Donc 2 hypothèses se dégagent.

Les hypothèses sont :

* H0 : Le mois n'influence pas la consommation d'énergie
* H1 : Le mois influence la consommation d'énergie

Le choix du test :

Nous allons choisir le test Anova puisque nous devons étudier une intéraction entre une variable catégorielle et une variable continue.

Résultat : p value = 4.311662907431899e-15 donc p value < 0,05

L'hypothèse H1 est validée. On peut conclure que le mois de l’année a une influence significative sur la consommation d’énergie en France. Selon les différents mois, la consommation est plus ou moins importante. Ce résultat permet d’admettre une corrélation entre ces deux variables. Nous pourrons nous servir de cette corrélation dans la phase de modélisation, où nous développerons des modèles prédictifs pour anticiper la consommation d'électricité en fonction des différentes variables explicatives.

# Introduction à la modélisation

Afin de répondre au mieux à nos objectifs, nous avons choisi d’effectuer des modélisations et de réaliser des prédictions à la fois au niveau national et au niveau régional. A partir de deux jeux de données de températures, l’un national, l’autre régional, nous avons pu préparer deux jeux d’entraînements différents. C’est pourquoi notre rapport de modélisation se divise en deux grandes parties distinctes.

#### À quel type de problème de machine learning votre projet s’apparente-t-il ?

Notre projet est principalement un projet de régression, car notre objectif est de prédire la consommation d'énergie en MW par jour qui est une variable continue. Cependant, nous avons également exploré des modèles de classification pour déterminer l'appartenance ou non des résultats au bon quartile de consommation.

#### À quelle tâche de machine learning votre projet s’apparente-t-il ?

Notre projet ne se limite pas à une seule tâche de machine learning spécifique. Nous abordons la modélisation de données de consommation d'énergie, ce qui s'apparente à une tâche de régression. Cependant, en explorant des modèles de classification, nous nous aventurons dans des domaines tels que la segmentation de la consommation d'énergie, ce qui élargit la portée de notre projet.

#### Quelle est la métrique de performance principale utilisée pour vos modèles ?

Notre principale métrique de performance est le coefficient de détermination (R²). Nous l'avons choisi car il permet d'évaluer la qualité de nos modèles de régression en mesurant leur capacité à expliquer la variance des données observées.

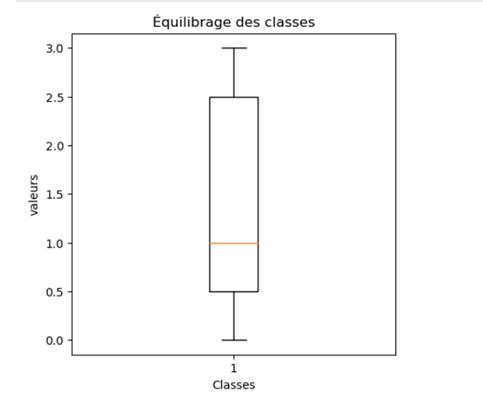
#### Avez-vous utilisé d’autres métriques de performances ?

En plus du coefficient de détermination (R²), nous avons employé d'autres métriques pour évaluer la performance de nos modèles, notamment le Mean Squared Error (MSE), le Mean Absolute Error (MAE), et la Root Mean Squared Error (RMSE). Ces métriques nous ont permis d'obtenir une compréhension plus fine des performances de nos modèles en termes d'erreurs, de précision et de robustesse. Pour mesurer les performances de nos modèles de classification, nous avons également suivi le f1-score, la matrice de confusion et le score de précision.

# La modélisation au niveau national

## Le modèle de classification

Nous avons tenté une approche de classification pour l’étude de la consommation d’énergie sur le plan national. Nous avons effectué un encodage en classe de notre variable cible « Consommation (MW) ». Nous avons découpé la variable en quartile. Nous avons remarqué qu’en augmentant le nombre de classes (nous avons testé 5 et 10 classes), la précision des prédictions devenait plus faible (pas plus de 0.75 sur les coefficients de détermination) et les résultats étaient moins intéressants en diminuant le nombre de classes.



J’ai utilisé la méthode “qcut” pour découper ma classe en 4 classes correspondant aux interquartiles (q1, q2, q3 et q4). Mes classes sont équilibrées avec environ 3200 lignes, chacune de nos classes fait 800 lignes.

À la suite de cela, nous avons testé quatre modèles de classification que sont :

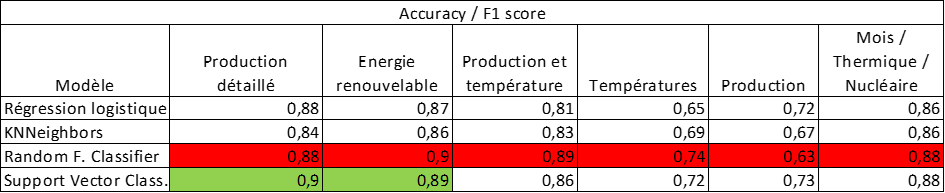
- Le modèle de régression logistique

- Le modèle K Nearest Neighbors (K-NN)

- Le modèle de Random Forest Classifier

- Le modèle de Support Vector Classification

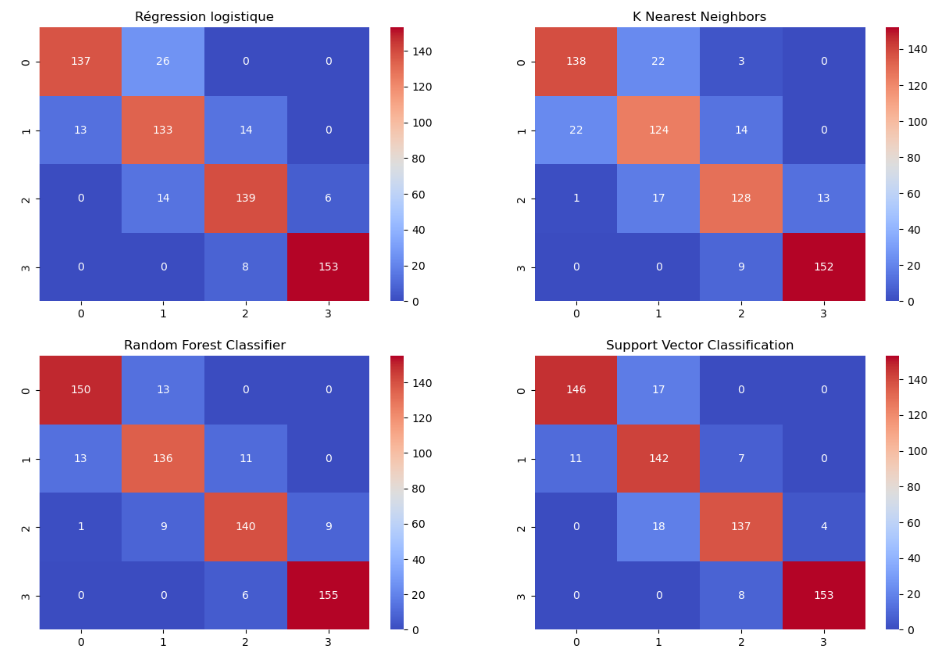
Pour ces quatre modèles, nous avons fait des essais pour gagner en précision en enlevant des variables. Ainsi, vous pouvez retrouver ci-dessous les six essais et les résultats du Accuracy / F1 score.



Dans l’ensemble des tests, nous remarquons des résultats ne dépassant qu’à une seule reprise les 90% et cela malgré plusieurs essais pour gagner en précision. Nous pouvons en déduire qu’il faut garder l’ensemble de nos variables pour tester nos modèles de classification. Le modèle Support Vector Classification est celui qui est le plus intéressant dans les résultats. Tandis que le modèle Random Forest Classifier est en surapprentissage, ce qui est flagrant sur le coefficient de détermination du train puisque tous les coefficients de détermination des modèles sur le train sont égal à 1. Avec une modification de la profondeur maximum du modèle, on peut corriger sur le surapprentissage, sans pour autant obtenir un R² aussi intéressant que pour le SVC ou les scores actuels du Random Forest.

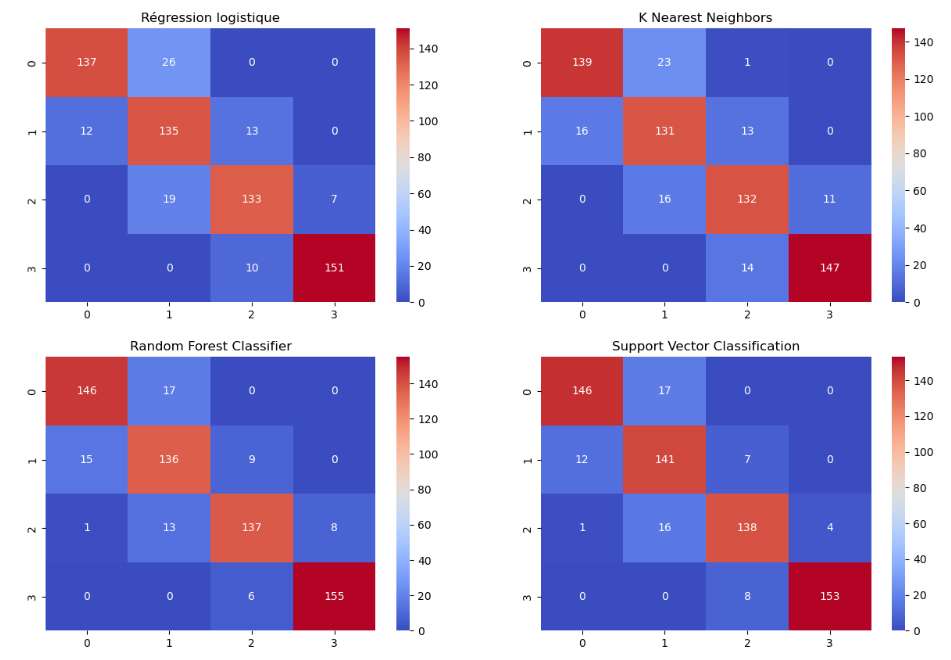
Nous retrouvons ci-dessous un heatmap des matrices de confusion des deux tests les plus performants :

Test 1:



Nous voyons que nos classes sont globalement bien prédites, voire très bien prédites pour la classe 3, les résidus ne sont pas éloignés et ne représentent pas beaucoup de fausses prédictions. On remarque que les valeurs bien prédites sont de 89% pour la classe 0, 84% pour la classe 1, 85% pour la classe 2 et 96% pour la classe 3.

Test 2 :

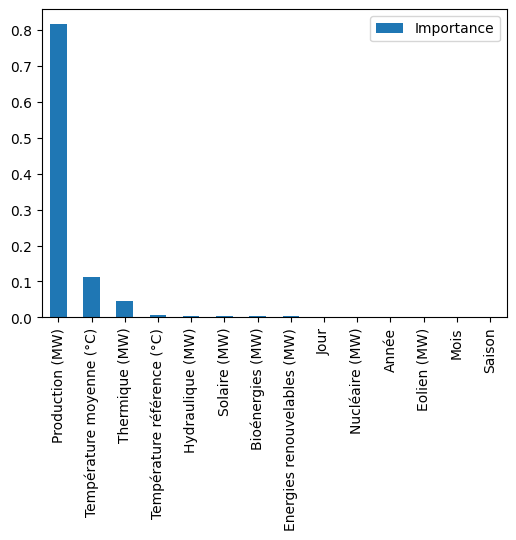


Nous remarquons cette fois-ci que les classes 1 et 2 sont les valeurs qui sont les moins bien prédites en comparaison avec les classes 0 et 3. Pour les classes 1 et 2, les résidus correspondent à 15% et à 16% , pour les classes 0 et 3, les résidus représentent seulement 11% et 5%.Cela permet de voir que nous avons des niveaux de performance corrects, mais au même titre que les coefficients de détermination, les matrices de confusion ne sont pas comparables à celles trouvées dans l’approche de régression.

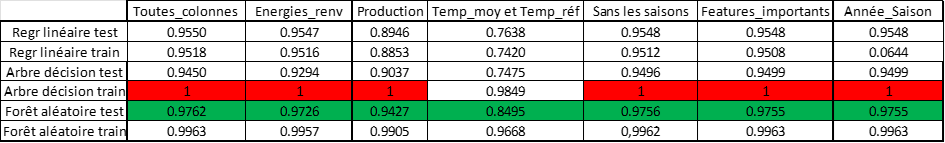
Pour conclure, nous pouvons dire que les modèles de classification sont très corrects, mais ils ne sont pas les modèles plus adaptés à notre problématique. En effet, le fait de classer notre variable cible en 4 classes correspondant aux interquartiles n’est pas adapté aux grands écarts de valeur. C’est pourquoi il est plus judicieux de se concentrer sur l’approche de régression.

## Le choix du modèle de régression

Nous avons examiné divers algorithmes de régression, notamment LinearRegression, DecisionTreeRegressor et RandomForestRegressor. Pour identifier les modèles les plus performants, nous avons évalué ces trois algorithmes sur des ensembles de données comportant différentes configurations de colonnes. Nous avons commencé par inclure toutes les colonnes. Ensuite, nous avons testé les données avec une seule colonne regroupant les énergies renouvelable, une seule colonne pour la production, une sélection ne comprenant que les données de températures moyennes et de références, une sélection sans la colonne saison et enfin, nous avons choisi les cinq premières colonnes en se basant sur les "feature importances" provenant du modèle d'arbre de décision mentionné ci-dessous :



Ci-dessous, vous trouverez les R² obtenus pour les différents modèles de régression cités précédemment :



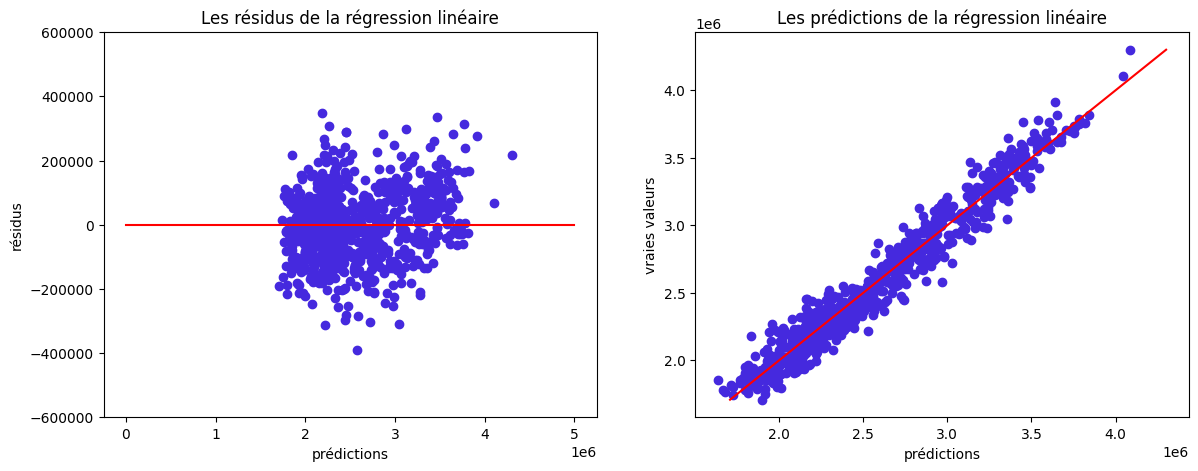
Il est évident que, parmi tous nos essais, le modèle de forêt aléatoire se démarque comme le plus approprié, affichant un R² le plus proche de 1. Parallèlement, nous observons un problème de surapprentissage pour tous les modèles DecisionTreeRegressor lors de nos diverses phases d' entraînement, avec un R² du train égal à 1. Avec une modification de la profondeur maximum du modèle, on peut corriger sur le surapprentissage, sans pour autant obtenir un R² aussi intéressant que pour la forêt aléatoire.

Les graphiques ci-dessous mettent en évidence que le modèle de forêt aléatoire offre la meilleure précision, avec des résidus moins dispersés par rapport aux deux autres modèles. C'est pourquoi nous avons opté pour l'optimisation de ce modèle.

Le modèle de régression linéaire :

*Coefficient de détermination du modèle sur train: 0.9518441844089172*

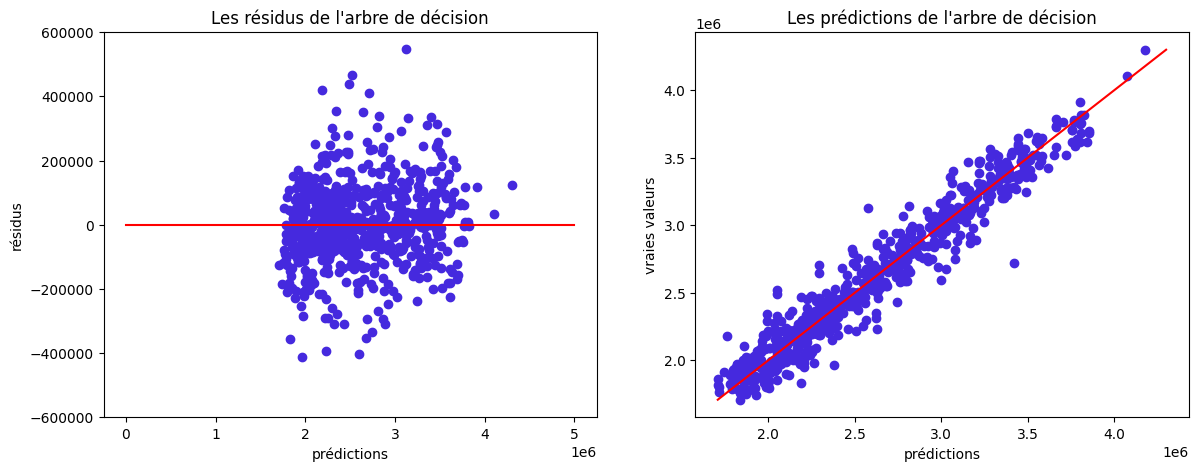
*Coefficient de détermination du modèle sur test: 0.9550433190545952*



Le modèle d'arbre de décision :

*Coefficient de détermination du modèle sur train: 1.0*

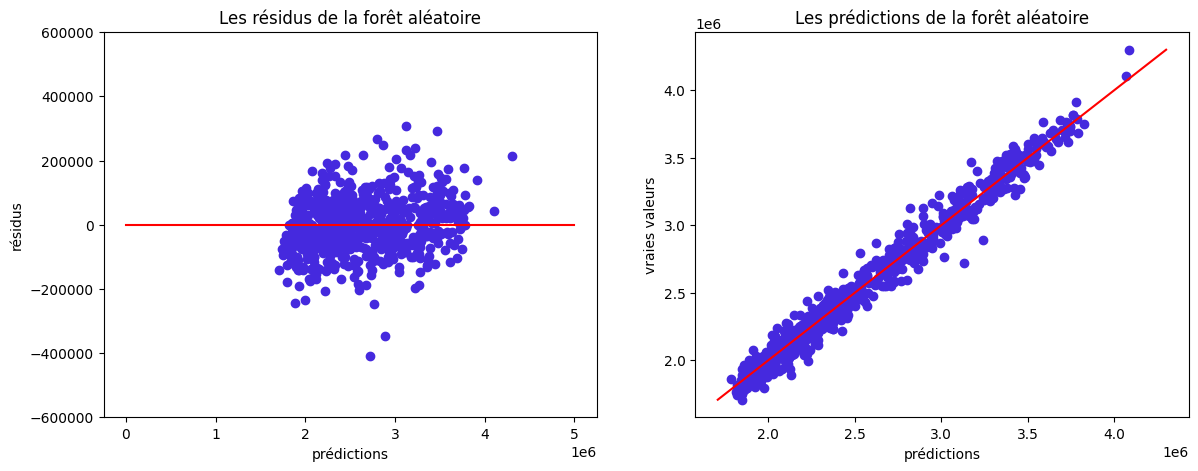
*Coefficient de détermination du modèle sur test: 0.9450855922389325*



Le modèle de forêt aléatoire :

*Coefficient de détermination du modèle sur train: 0.9963077580150688*

*Coefficient de détermination du modèle sur test: 0.97620340345898*



Lorsqu'on examine les métriques de performance globales pour ces trois modèles, il est clair

qu'il existe un problème évident de sur apprentissage pour l'arbre de décision avec un R²=1. En revanche, le modèle de forêt aléatoire semble être la plus prometteuse, ne commettant qu'une erreur de prédiction de 60 794 unités sur les prévisions de consommation d'électricité quotidienne.



#### Quel algorithme avez-vous retenu et pourquoi ?

Le modèle de forêt aléatoire est l’algorithme retenu pour notre étude du fait des résultats obtenus à la suite des différents tests. Elle présente la meilleure précision que ce soit sur les R² et les métriques, ainsi que des résidus moins dispersés. C’est le modèle que nous allons chercher à optimiser par la suite.

#### Avez-vous utilisé des techniques d’optimisation de paramètres ?

Après avoir identifié le modèle de régression de la forêt aléatoire comme le plus pertinent, nous avons entrepris de rechercher les hyper paramètres optimaux de ce modèle en utilisant la méthode de Grid Search. Cependant, il est important de noter que cette méthode prend environ 25 minutes avant de s'exécuter dans notre cas. Grâce à cette approche d'optimisation, nous avons abouti aux hyperparamètres suivants :

*'max\_degree': 30 (profondeur maximal de l'arbre)*

*'max\_features': 'log2' (nombre maximal de caractéristiques)*

*'min\_samples\_leaf' : 1 (nombre minimal d'échantillons pour ancien un nœud terminal)*

*'min\_samples\_split' : 2 (nombre minimal d'échantillons requis pour effectuer une division dans l'arbre)*

*'n\_estimators' : 559 (nombre d'arbres à utiliser pour l'ensemble de la forêt)*

En utilisant ces hyper paramètres optimisés, notre modèle de forêt aléatoire affiche désormais les scores suivants :

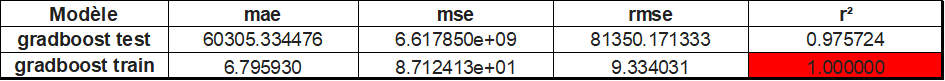
*Mean Squared Error (MSE): 6187581990.801711*

*Mean Absolute Error (MAE): 60684.09990183722*

*R-squared (R2): 0.9776658387254317*

#### Avez-vous testé des modèles avancés ?

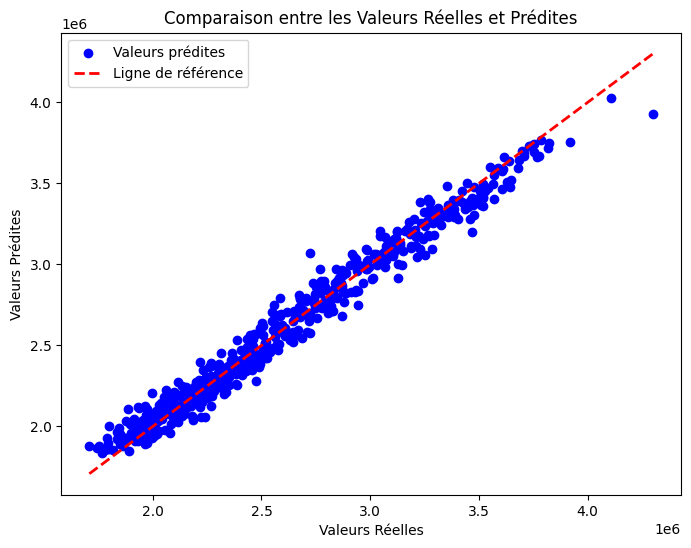
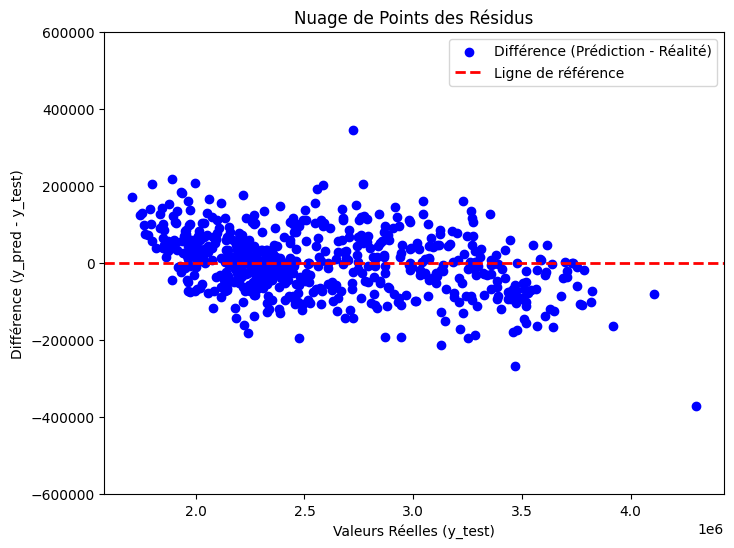
Nous avons opté pour le modèle de Gradient Boosting afin de pousser plus en profondeur nos essais d'apprentissage avancé. Malheureusement, après avoir réalisé plusieurs ajustements de paramètres lors de divers tests d'apprentissage, nous avons constaté qu'il n'a pas mieux performé que le modèle de Forêt aléatoire. Au contraire, il a montré des signes d’overfitting avec un R²=1. Par conséquent, nous maintenons le modèle de Forêt aléatoire pour la suite de notre étude.



## Interprétation des résultats

#### Avez-vous analysé les erreurs de votre modèle, avec quelles améliorations ?

Afin d'analyser les résultats de notre forêt aléatoire optimisée, nous avons procédé à une visualisation des résidus et à une comparaison entre les valeurs prédites et les valeurs réelles. En examinant les résidus, nous pouvons dire que le modèle capture correctement la structure des données avec une légère amélioration du coefficient de détermination R-squared (R²): 0.9776 énuméré plus haut. Après avoir réalisé ces graphiques ci-dessous, nous n'avons pas entrepris d'améliorations supplémentaires.

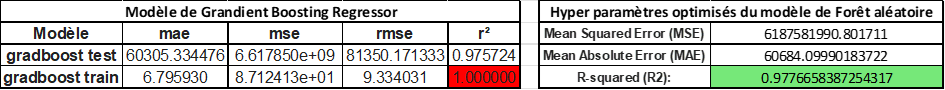


#### Avez-vous utilisé des techniques d’interprétabilité ?

En raison de notre familiarité limitée avec les diverses techniques et de la contrainte de temps restant, nous n'avons pas pu mettre en œuvre d'autres méthodes d'interprétabilité pour affiner nos tests au niveau national. Cependant, nous avons utilisé la technique d'interprétabilité de type [SHAP](#6zogq4kfqj0u) pour analyser la partie régionale de notre étude (voir page 39).

#### Qu’est ce qui a (ou non) engendré une amélioration de vos performances ?

Lorsque nous examinons les métriques de nos deux modèles, il est clair que le modèle de forêt aléatoire offre de meilleures performances, sans montrer de signes de sur-apprentissage, tandis que le modèle de Gradient Boosting Regressor présente des problèmes d'overfitting.



# 

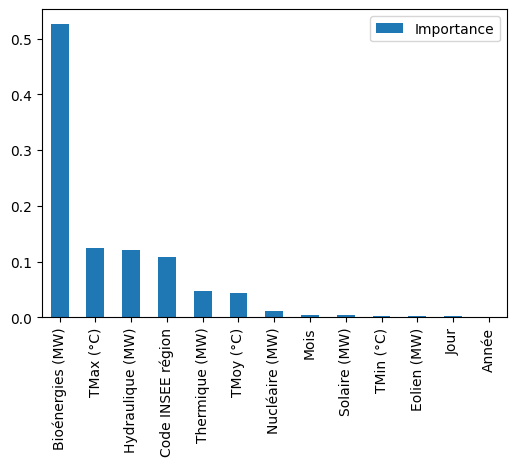
# La modélisation au niveau régional

## Choix du modèle et optimisation

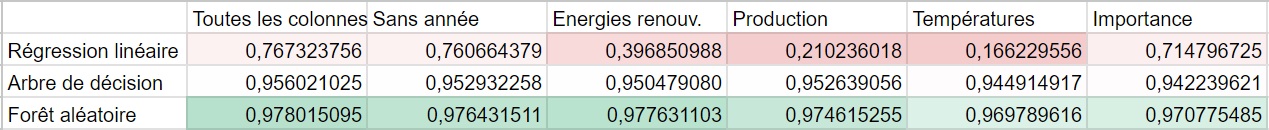
#### Quels algorithmes avez-vous essayés ?

Nous nous sommes concentrés sur des algorithmes de régression : LinearRegression, DecisionTreeRegressor et RandomForestRegressor. Et afin de trouver les meilleurs modèles, nous avons testé ces trois algorithmes sur différents nombres de colonnes : d’abord toutes les colonnes, sans la colonne année, avec une seule colonne énergies renouvelables (sans le détail : solaire, éolien, etc.), avec une seule colonne production, avec seulement les températures et enfin en sélectionnant les six premières colonnes selon les feature importances ci-dessous :

*Feature importances du modèle de forêt aléatoire*



Voici les résultats des coefficients de détermination (ou R²) des différents modèles de régression sur le jeu de test :



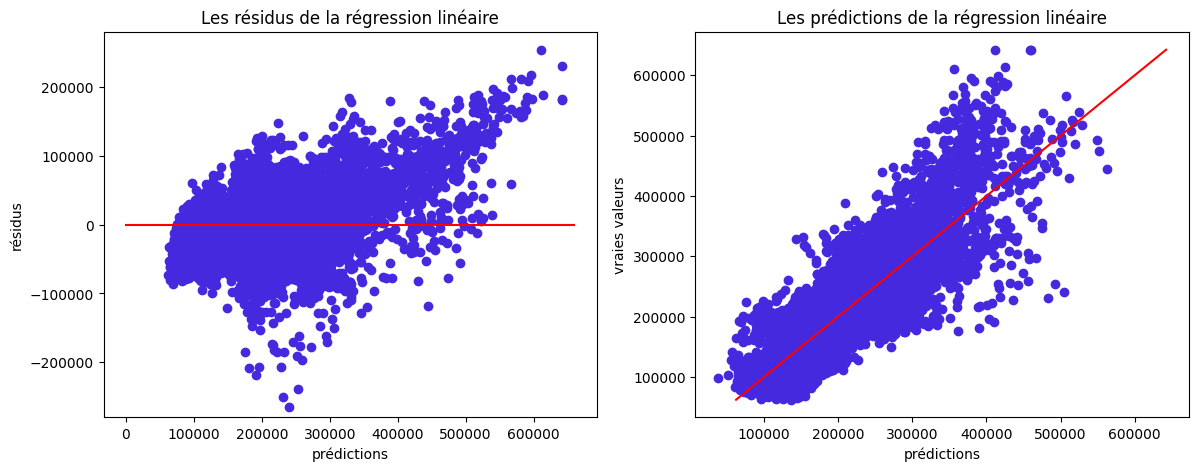
Comme on peut le voir, c’est le modèle avec toutes les colonnes qui performe pour les trois algorithmes. Nous allons donc nous concentrer sur cette configuration et l’étudier plus en détail à l’aide de graphiques.

Comme on peut le voir sur les graphiques ci-dessous, le modèle de forêt aléatoire est celui qui nous apporte le plus de précision, avec des résidus moins éparpillés que pour les deux autres modèles. C’est donc celui que nous avons choisi d’optimiser.

Le modèle de régression linéaire :

*Coefficient de détermination du modèle sur train: 0.7595525*

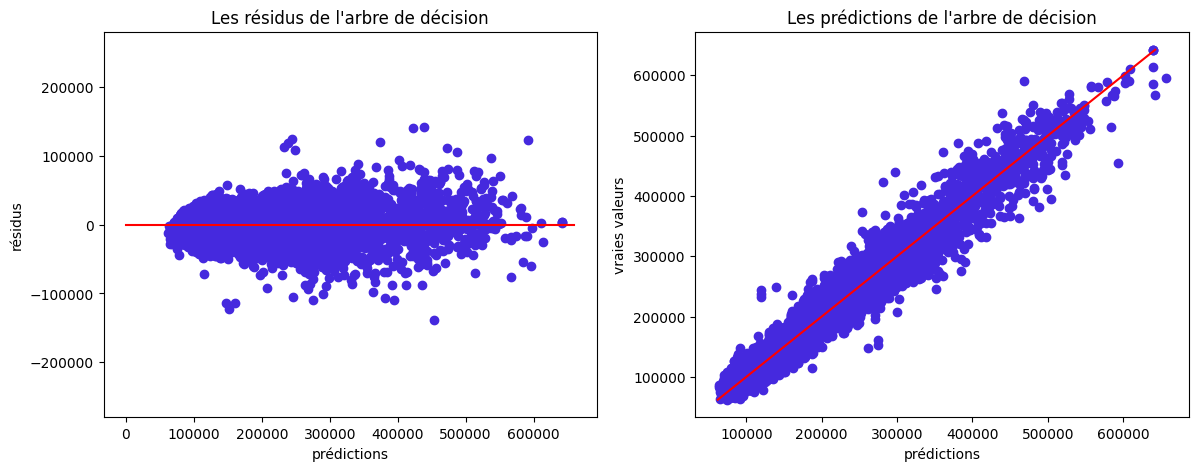
*Coefficient de détermination du modèle sur test: 0.7638239*



Le modèle d'arbre de décision :

*Coefficient de détermination du modèle sur train: 1.0*

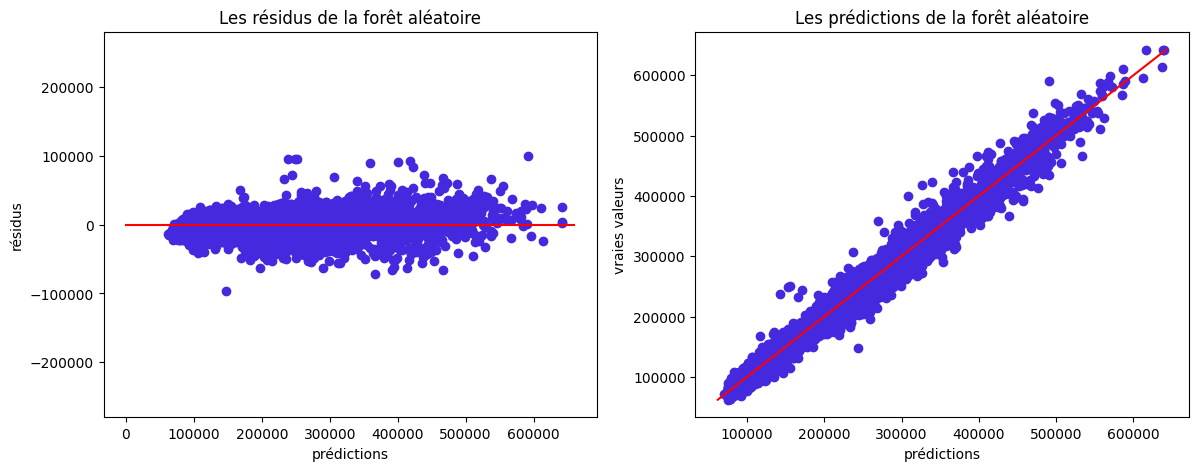
*Coefficient de détermination du modèle sur test: 0.9545123*



Le modèle de forêt aléatoire :

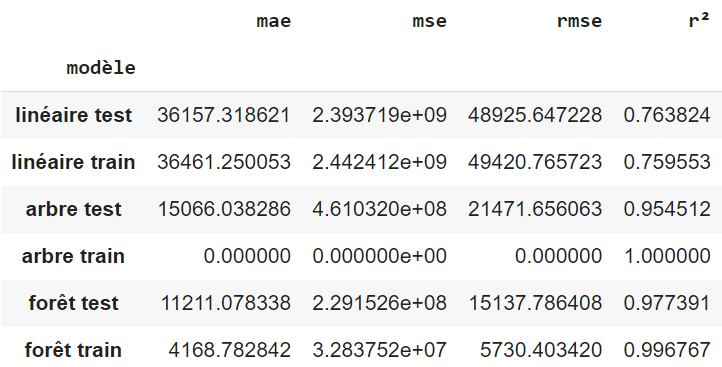
*Coefficient de détermination du modèle sur train: 0.9967673*

*Coefficient de détermination du modèle sur test: 0.9773907*



En regardant les métriques de performance dans leur ensemble, toujours sur ces trois mêmes modèles, on observe un problème de surapprentissage flagrant pour l’arbre (DecisionTreeRegressor). En effet, son coefficient de détermination pour le modèle d'entraînement est égal à 1. En modifiant la profondeur maximum du modèle, on peut améliorer les choses, sans pour autant obtenir un R² aussi intéressant que pour la forêt aléatoire.

La forêt aléatoire semble donc être la plus prometteuse. Elle ne se trompe que de 11 211 unités sur les prévisions de consommation d'électricité journalière. Ce qui est peu sachant que nos données varient de 59 000 à 657 000 MW consommés par jour.



#### Quel algorithme avez-vous retenu et pourquoi ?

D’après nos observations, le choix de la forêt aléatoire nous semble le plus évident. Elle présente une précision supérieure et des résidus moins dispersés que les autres modèles que nous avons testés. Cela indique que la forêt aléatoire est en mesure de mieux expliquer la variance des données et de fournir des prédictions plus précises sur notre problématique (prédire une consommation journalière au niveau régional). C’est le modèle qui donne les meilleurs résultats pour l’instant.

#### Avez-vous utilisé des techniques d’optimisation de paramètres ?

Après avoir identifié le modèle de régression de forêt aléatoire comme le plus pertinent, nous avons cherché les hyperparamètres de ce modèle grâce à la méthode Grid Search Cette méthode est très longue à s'exécuter. Elle met plus d’une heure. Avec cette technique d’optimisation, nous obtenons les hyperparamètres suivants :

*'max\_depth': 30 (profondeur maximal de l’arbre)*

*'max\_features': 'log2' (nombre maximal de caractéristiques)*

*'min\_samples\_leaf': 1 (nombre minimal d’échantillon pour former un noeud terminal)*

*‘min\_samples\_split': 2 (échantillon minimal requis pour effectuer une division dans l'arbre)*

*'n\_estimators': 413 (nombre d’arbres à utiliser pour l’ensemble de la forêt)*

A l’aide de ces hyperparamètres, notre forêt aléatoire optimisée obtient à présent les scores suivants :

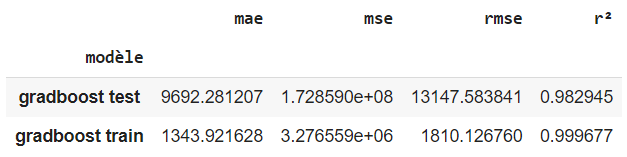
*MSE : 218 593 173*

*MAE : 11 211*

*r² : 0.9783*

#### Avez-vous testé des modèles avancés ?

Le Gradient Boosting Regressor est un modèle puissant pour la régression, mais il nécessite une attention particulière lors du réglage des paramètres pour éviter l’overfitting. Il est lent à entraîner, mais permet de traiter des données avec des valeurs extrêmes comme les nôtres. En manipulant les paramètres de façon assez empirique, vous les résultats que nous avons pu obtenir rapidement avec ce modèle :



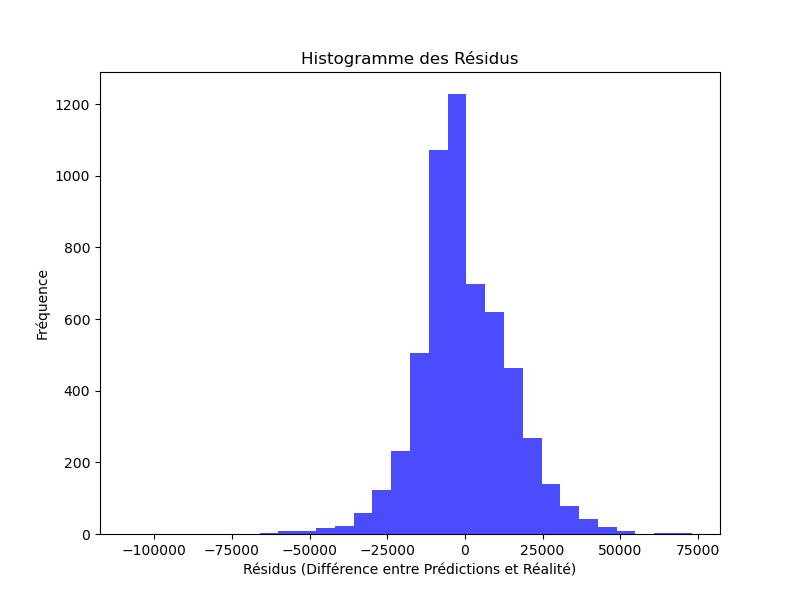
Ce modèle ne se trompe que de 9 692 unités sur les prévisions de consommation d'énergie journalière. Ce résultat a été obtenu avec un nombre de colonnes réduit, en supprimant les colonnes : Éolien (MW), Solaire (MW), Hydraulique (MW), Bioénergies (MW), et en ne conservant qu’une colonnes Énergies renouvelables (MW) qui les résume.

Il serait encore possible d’améliorer ce modèle en ajustant certains paramètres comme le n\_estimators, mais plus on augmente ce paramètre et plus le modèle met de temps à tourner, sachant que nous avons en plus choisi un learning\_rate assez bas. Le modèle actuel met plus de 4 minutes à s’exécuter.

## Interprétation des résultats

#### Avez-vous analysé les erreurs de votre modèle, avec quelles améliorations ?

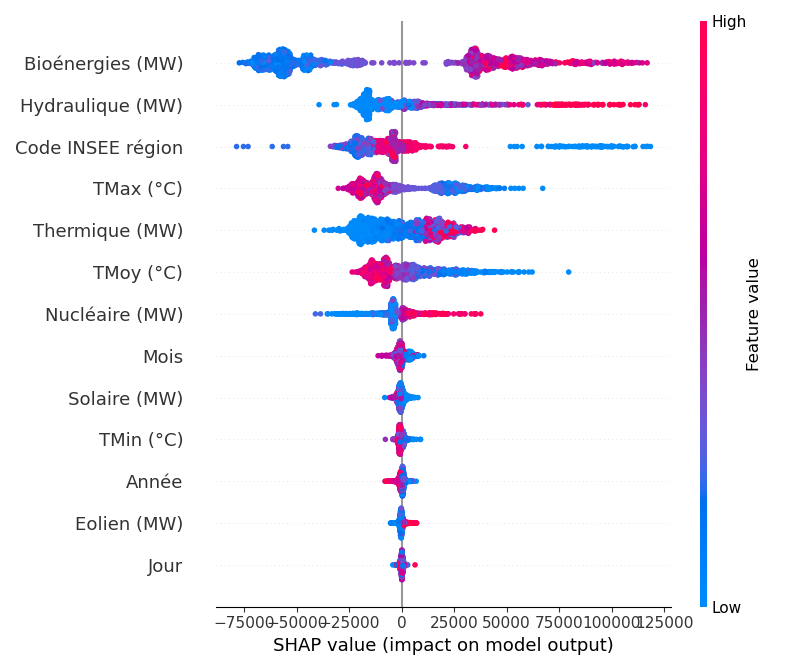
Pour analyser les résultats de notre forêt aléatoire optimisée, nous avons réalisé des visualisations de résidus et une comparaison entre valeurs prédites et valeurs réelles. Grâce à l’analyse des résidus en nuage de points, nous avons constaté l’absence de schéma de répartition montrant ainsi que le modèle capture correctement la structure des données. l’étude des résidus sous forme d'histogramme nous permet de confirmer cette hypothèse, en effet la répartition en forme de cloche avec une répartition équilibrée suggère que les résidus sont aléatoires. Nous n’avons pas effectué d’amélioration suite à la réalisation de ces graphiques.



|  |  |
| --- | --- |
|  |  |

#### Avez-vous utilisé des techniques d’interprétabilité ?

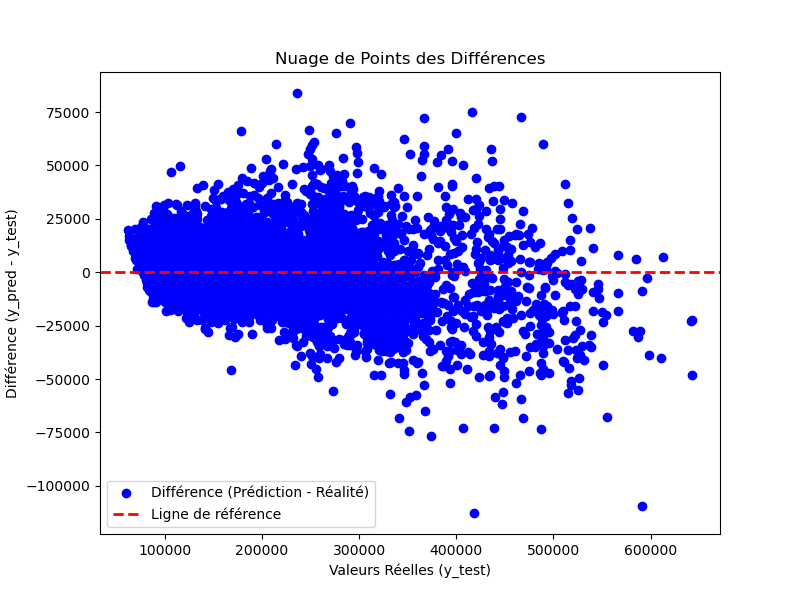
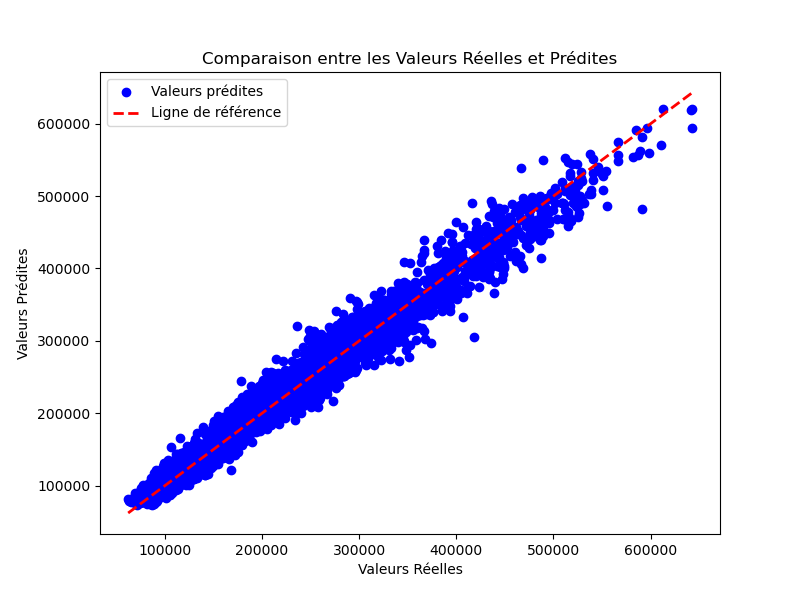
Nous avons réalisé un summary plot sur notre forêt aléatoire optimisée à l’aide de SHAP. Le but étant d’étudier les features importantes de notre dataframe. Voici les résultats obtenus :

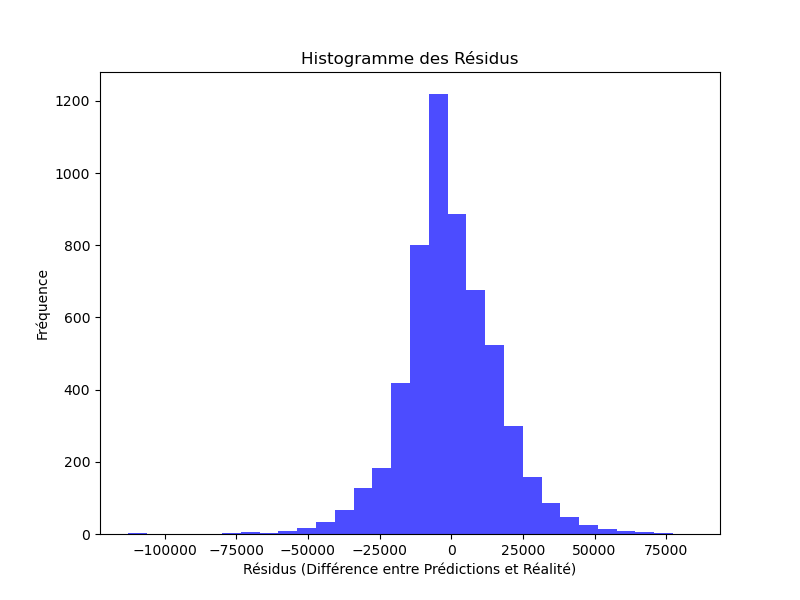


Nous avons ensuite réalisé une nouvelle étude en ne gardant que les 7 features les plus importantes (Hydraulique (MW), TMin (°C), Solaire (MW), Code INSEE région, Consommation (MW), TMAX (°C), Thermique (°C) ). Après cette étude avec seulement ces features, on remarque un r² , un MAE et une MISE sensiblement identique

*MSE : 279 560 649*

*MAE : 12 466*

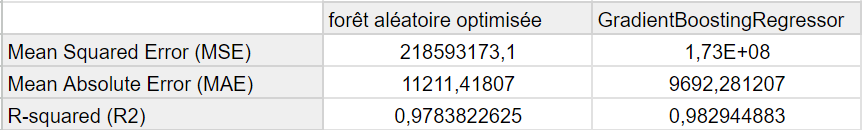
*r² : 0.9724* 



On remarque qu’avec seulement 7 variables. Les graphiques et métriques sont quasiment identiques. Notre modèle de forêt aléatoire regressor est donc bien influencé par ces 7 variables et que l’impact des autres variables est très léger. Même si le modèle est très légèrement moins précis, il est beaucoup plus rapide à se lancer.

#### Qu’est ce qui a (ou non) engendré une amélioration de vos performances ?

Si on compare nos deux meilleurs modèles avec les métriques, c’est finalement le Gradient Boosting Regressor qui obtient les meilleurs résultats, mais c’est également celui qui met le plus de temps à s'exécuter. Notre modèle de régression de forêt aléatoire optimisée ne met pas plus de 40 secondes à s’exécuter, tandis que le Gradient Boosting Regressor met plus de 4 minutes.

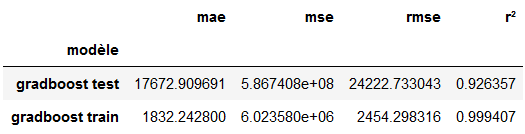


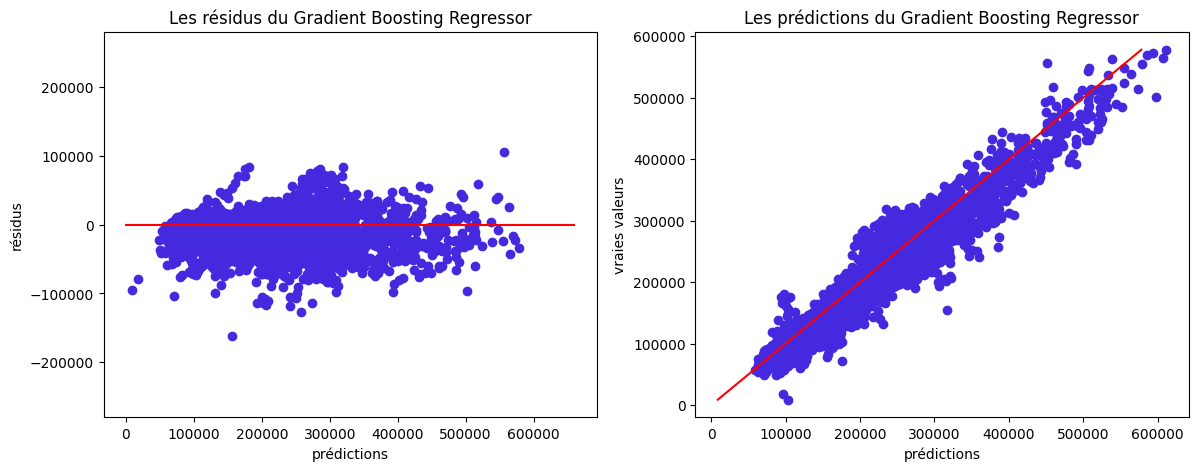
Tous nos tests pour trouver le nombre de colonnes optimal n’ont pas non plus engendré d’amélioration significative. Nous avions en effet déjà un nombre de colonnes assez restreint dès le départ.

#### Des prédictions pour 2022 et 2023

Nous avons voulu testé notre Gradient Boosting Regressor sur un autre jeu de données afin de voir s’il s’en sortait avec une source de données différente. Nous avons donc réuni le dataset [*éCO2mix régionales temps réel*](#_3bvr14gyefa1) (juin 2022 à septembre 2023) avec notre jeu de données des températures qui couvrent également la période de 2022 à 2023. Ce nouveau dataset a nécessité plusieurs étapes de nettoyage et un réajustement des valeurs qui étaient fournies au quart d’heure et non à la demi-heure. Nous avons cette fois entraîné notre modèle sur l’ensemble de nos données d’origine.

Avec un coefficient de détermination de 0.9264, cet test est plutôt concluant et nous permet d’obtenir les résultats suivants :





Nous avons ainsi pu obtenir des prévisions de consommation par jour, de juin 2022 à septembre 2023 pour toutes les régions de France métropolitaine. Si on regarde plus en détail les résidus (premier graphique), notre modèle semble avoir eu tendance à surestimer la consommation d’énergie sur cette période. Les points étant en majorité sous la barre du zéro, le modèle a prédit des consommations supérieures à la réalité pour un grand nombre de cas.

#### Des prédictions sans machine learning

Nous avons également essayé de créer un modèle de prédiction autre que machine learning en calculant simplement les moyennes des consommations journalières par région pour ensuite comparer ces moyennes calculées aux données déjà existantes. Afin de pouvoir comparer ce modèle aux autres, nous avons décidé de faire les prédictions de consommation par région et par jour.

Les graphiques ci-dessous montrent une dispersion uniforme. On constate aussi que les valeurs sont très élevées, le modèle est moins performant. Cela s’explique par le fait que les moyennes soient très sensibles aux valeurs extrêmes présentes dans nos dataframe.

# 

# 

Voici les résultats obtenus :

*MAE : 15 829*

*R² : 0.953*

# 

On remarque que la MAE est plus haute que le modèle choisi a une précision de 0.95. Ce modèle est légèrement moins précis que les modèles de machine learning mais est beaucoup plus rapide à se lancer. Nous aurions pu utiliser ce modèle avant de nous lancer dans le machine learning afin d’effectuer une première prédiction de consommation, le résultat étant très proche de la réalité.

# 

# Conclusion

L'objectif principal du projet était d'analyser le phasage entre la consommation et la production énergétique. Les objectifs spécifiques incluent l'analyse régionale de la consommation, l'examen des filières de production d'énergie. L'équipe n'avait pas de compétences spécialisées préalables dans le domaine de l'énergie, mais elle était motivée par la curiosité et l'intérêt pour le sujet. Cependant, nos compétences métier étant limitées, certains termes et certaines valeurs ne nous semblaient pas pertinents. Nous avons finalement obtenu une réponse floue et tardive d’une personne contactée par mail, ce qui ne nous a pas empêché d’avancer sur le projet.

Pour rappel, nous avons choisi d’effectuer des modélisations et de réaliser des prédictions à la fois au niveau national et au niveau régional dans le but de répondre au mieux aux objectifs de ce projet, c’est-à-dire effectuer des prévisions de consommation d’énergie en France. Nous avons également fait le choix d'ajouter des données climatiques, en particulier les températures, pour améliorer les prévisions. Ces deux approches nous ont permis d’entraîner un grand nombre de modèles différents et pour cela nous nous sommes séparés en deux équipes de deux. Chaque équipe avançant à son rythme et suivant ses compétences.

Pour la partie nationale, nous avons choisi d’étudier à la fois des modèles de classification et de régression. Il s’avère que les résultats de nos prédictions sur des modèles de classification ont été corrects dans l’ensemble. Cependant lorsque nous comparons les f1 score et les matrices de confusion au R², nous constatons de meilleurs résultats du côté du R² des modèles de régression. De plus, afin de réaliser une prédiction de la consommation journalière un modèle de régression est largement plus adapté à un modèle de classification qui nous permet seulement de déterminer une classe de consommation. C’est pour ces raisons que nous n'avons pas poursuivi sur les modèles de classification.

Nous avons testé trois modèles de régression et nous avons sélectionné le modèle de forêt aléatoire avec des R² sur l’ensemble des tests à 0,97 et d’excellents résultats sur les métriques. Dans l’optique d'améliorer la performance de notre modèle de forêt aléatoire, nous avons utilisé la méthode de Grid Search pour trouver les hyper-paramètres optimaux, ce qui a conduit à une augmentation de la précision de nos prédictions.

Pour approfondir et affiner davantage nos prédictions, nous avons opté pour l'utilisation de la méthode de Gradient Boosting. Cependant, malgré nos multiples ajustements de paramètres au cours de diverses phases de test, nous avons constaté que cette méthode n'a pas surpassé le modèle de forêt aléatoire. Le Gradient Boosting a présenté des signes de sur-apprentissage, avec un coefficient de détermination sur train (R²) de 1.

Pour conclure, nous gardons le modèle de forêt aléatoire optimisée qui a donné les résultats les plus performants et qui fait partie des plus rapides à exécuter. Nous avons utilisé le Gradient Boosting Regressor sans obtenir de bons résultats alors la forêt aléatoire optimisée reste notre meilleure option du fait de sa rapidité.

Pour la partie régionale, notre projet de modélisation de la consommation énergétique régionale a impliqué l'exploration de divers algorithmes de régression. Après avoir évalué ces modèles avec différentes configurations, nous avons identifié que le modèle de forêt aléatoire présentait les performances les plus prometteuses. Ce choix s'est basé sur des indicateurs tels que le coefficient de détermination (R²), mais également sur des graphiques qui nous ont permis de mieux comprendre les résultats obtenus.

Nous sommes allés au-delà de ce qui était proposé dans la formation en optimisant notre modèle de forêt aléatoire et en utilisant une méthode de Grid Search pour déterminer les hyper-paramètres optimaux. Cela a permis d'améliorer la précision de nos prédictions, même si cette étape a nécessité des temps d’exécution de plus d’une heure.

Nous avons également exploré un modèle avancé, le Gradient Boosting Regressor, qui a montré des résultats prometteurs, bien que beaucoup plus lent à s'exécuter. En recherchant les meilleures caractéristiques, nous avons obtenu des résultats encourageants, mais avec une perte de sens, ce type de modèle étant plus difficile à interpréter.

Enfin, nous avons réalisé des prédictions pour 2022 et 2023 en utilisant le Gradient Boosting Regressor sur un jeu de données différent, avec des résultats concluants. Nous avons également expérimenté un modèle de prédiction sans machine learning, en calculant simplement les moyennes de consommation par région. Si la précision de cette méthode est inférieure par rapport aux modèles de machine learning, elle est néanmoins rapide à mettre en place et à exécuter.

Au regard de toutes ces expériences, nous retenons le modèle de forêt aléatoire optimisée qui est parmi les plus performants de ceux que nous avons pu testé et qui est le plus rapide à s'exécuter. Bien que le modèle Gradient Boosting Regressor ait été exploré en raison de sa puissance en régression, la forêt aléatoire optimisée est demeurée le choix privilégié en raison de sa rapidité et de son efficacité.

Nous considérons que les objectifs principaux de ce projet ont été atteints puisque nous avons pu créer des modèles du niveau national et régional, que certains de ces modèles ont été optimisés et que des prévisions pour 2022 et 2023 ont été réalisées avec succès. Pour augmenter les performances de nos modèles, il serait possible d'explorer davantage les paramètres de modélisation, d'incorporer des données supplémentaires telles que les jours de la semaine et des données de population. De plus, la création de modèles régionaux dédiés pourrait améliorer la précision des prédictions pour chaque région de France.

Le projet a contribué à démontrer comment l'utilisation de données climatiques peut améliorer la précision des prévisions de consommation énergétique. Il a également montré l'impact de différentes approches de modélisation sur la qualité des prédictions, ce qui peut avoir des implications importantes pour le secteur de l'énergie et la prise de décision stratégique.

# 

# Bibliographie

Site web de l'ORDRE - Opendata Réseaux-Energies : Explorez, comprenez, et appropriez vous les données d’énergie, en direct des transporteurs d’énergie et de leurs partenaires : <https://opendata.reseaux-energies.fr>

Jeux de données issus de l'ORDRE :

* Données éCO2mix régionales consolidées et définitives (janvier 2013 à mai 2022) : <https://odre.opendatasoft.com/explore/dataset/eco2mix-regional-cons-def/information/>
* Température quotidienne régionale (janvier 2016 à aujourd’hui) : <https://odre.opendatasoft.com/explore/dataset/temperature-quotidienne-regionale/information/>
* Pic journalier de la consommation brute d'électricité (avec températures au niveau national, janvier 2012 à aujourd’hui) : <https://odre.opendatasoft.com/explore/dataset/pic-journalier-consommation-brute/information/>
* Données éCO2mix régionales temps réel (juin 2022 à aujourd’hui) : <https://odre.opendatasoft.com/explore/dataset/eco2mix-regional-tr/information/>

Site web de RTE - Réseau de Transport d'Électricité français : RTE, le gestionnaire du réseau de transport d'électricité français : <https://www.rte-france.com>

Pages web du site RTE :

* éCO2mix - Toutes les données de l'électricité en temps réel : <https://www.rte-france.com/eco2mix>
* Le Wiki de l'énergie : <https://www.rte-france.com/wiki-energie>
* Bilans électriques nationaux et régionaux : <https://www.rte-france.com/analyses-tendances-et-prospectives/bilans-electriques-nationaux-et-regionaux>

# 

1. Soutirage : lorsque de l'énergie est soutirée (retirée) des installations pour diverses raisons : maintenance, problème réseau, stockage, échanges, etc. [↑](#footnote-ref-1)