MASTER EN SCIENCES DU NUMERIQUE ET INTELLIGENCE ARTIFICIELLE

Université Mohammed V
Faculté des Sciences
Rabat

PRÉDICTION DE CONSOMMATION ÉLECTRIQUE

Réalisé par :

AIT MOULAY ERRADI YOUSSEF

BALHAN AMINE

Présenté devant :

M. LAANAYA Hicham

PLAN

Introduction

Ce projet vise à développer un modèle prédictif fiable de la consommation électrique en s'appuyant sur des techniques avancées de Machine Learning (ML) appliquées à des données réelles.

Il s'inscrit dans une approche d'apprentissage supervisé, plus précisément de régression, pour modéliser les relations entre la consommation électrique et différentes variables explicatives.

Trois algorithmes ont été mis en œuvre à cet effet :

- La régression polynomiale
- Le Random Forest
- Le XGBoost

Exploration des Données Utilisées

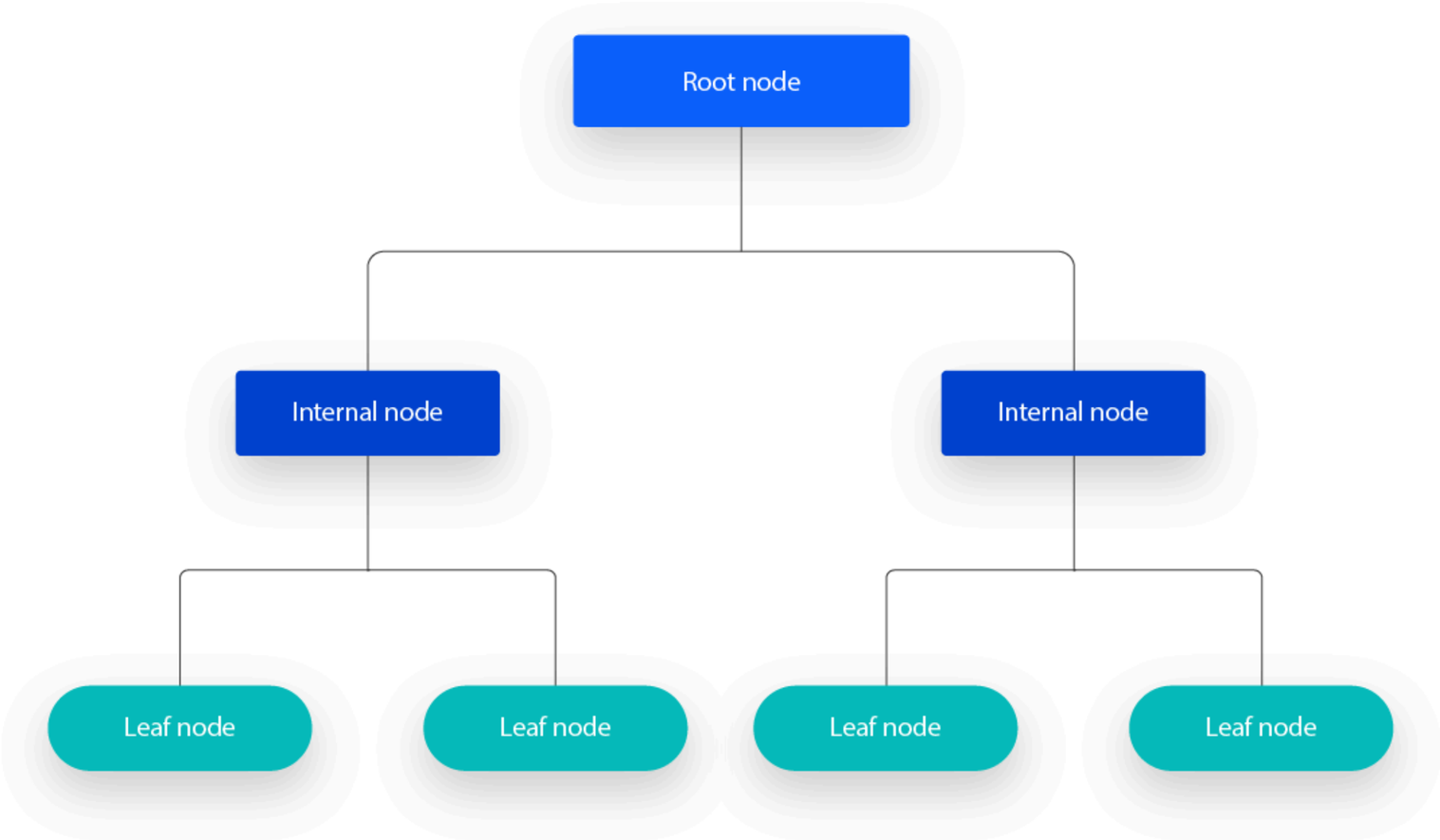
Le jeu de données utilisé provient d'une source publique disponible sur la plateforme **Kaggle**. Il comprend environ **26 000 lignes et 21 colonnes**, combinant des mesures horaires de consommation électrique avec des variables météorologiques et temporelles.

Les principales variables retenues pour la modélisation sont :

- Consommation électrique, variable cible à prédire.
- Température et humidité, facteurs majeurs influençant la demande énergétique.
- Pression atmosphérique et tendance barométrique, reflétant les conditions météorologiques.
- Vitesse et direction du vent, susceptibles d'affecter la consommation.
- Précipitations, témoins des conditions climatiques extrêmes.
- Variables temporelles (heure, mois), cruciales pour modéliser les cycles journaliers et saisonniers.

Arbre de décision

Un arbre de décision est un outil utilisé pour modéliser des décisions et leurs conséquences possibles sous forme d'un arbre. qui permet de prendre des décisions basées sur une série de conditions ou de règles.



Types d'arbres de décision

1. **Arbres de classification :**
   - Utilisés pour des problèmes où la variable cible est catégorielle.
   - L'objectif est de classer les entrées dans l'une des plusieurs catégories prédéfinies.
2. **Arbres de régression :**
   - Utilisés pour des problèmes où la variable cible est continue.
   - L'objectif est de prédire une valeur numérique.

Construction de l'arbre et prédiction

Lors de la construction d'un arbre de régression, le choix des caractéristiques et des points de division est essentiel et se fait de manière stratégique. Pour la régression, cette sélection vise à minimiser l'erreur quadratique résiduelle (RSS) ou l'erreur quadratique moyenne (MSE). L'objectif est de choisir les divisions qui réduisent au maximum ces erreurs. L'arbre segmente ainsi l'espace des données en plusieurs régions. Chaque observation appartenant à une région spécifique se voit attribuer la même prédiction, généralement la moyenne des valeurs cibles des points de données dans cette région.

L'algorithme cherche le meilleur point de division et la meilleure caractéristique qui réduisent le plus le RSS ou le MSE. La somme des RSS des régions résultant d'une division doit être inférieure au RSS de la région parente avant la division. Ce processus de division continue jusqu'à ce qu'un critère d'arrêt soit atteint.

Gestion du surapprentissage (Overfitting)

Les arbres de décision peuvent souffrir de surapprentissage, ce qui signifie qu'ils mémorisent les données d'entraînement et généralisent mal aux nouvelles données. Pour contrer cela, plusieurs techniques peuvent être utilisées :

Élagage (Pruning) : Supprimer les branches ou les nœuds de l'arbre qui ne contribuent pas significativement à des prédictions précises.

Méthodes d'ensemble (Ensemble Methods) : Combiner plusieurs arbres de décision pour améliorer la robustesse et réduire la variance. Des exemples incluent le Bagging et le Boosting.

Réglage des hyperparamètres (Hyperparameter Tuning) : Définir des contraintes sur la croissance de l'arbre.

- Max Depth : Le plus long chemin du nœud racine à un nœud feuille.
- Min Samples Split : Le nombre minimum d'observations requises dans un nœud pour qu'une division soit tentée.
- Min Samples Leaf : Le nombre minimum d'observations qui doivent se trouver dans un nœud feuille.
- Max Features : Détermine le nombre de caractéristiques à considérer lors de la recherche de la meilleure division à chaque nœud.

Nettoyage et Préparation des Données

Pour garantir la qualité des données utilisées dans nos modèles, plusieurs étapes de nettoyage et de préparation ont été réalisées :

Chargement du jeu de données 01

Les données ont été importées depuis un fichier CSV contenant environ 26 000 observations réparties sur 21 colonnes.

Suppression des colonnes non pertinentes 02

Certaines colonnes, telles que index, datehour et datemonth, ont été supprimées car elles étaient soit redondantes, soit inutiles pour la modélisation.

Extraction des variables temporelles 03

À partir de la colonne Date_Heure, nous avons extrait trois variables utiles :
hour (heure), day (jour) et month (mois), afin d'intégrer la dimension temporelle dans la modélisation.

Traitement des valeurs manquantes 04

La présence de quelques valeurs manquantes a été détectée dans certaines colonnes. Pour garantir la cohérence des données et éviter les erreurs lors de l'apprentissage, ces valeurs ont été remplacées par la moyenne de chaque colonne concernée.

Sélection des variables explicatives 05

Un ensemble de 11 caractéristiques a été retenu (pression, température, humidité, vent, etc.) pour prédire la variable cible : la consommation électrique.

Séparation du Jeu de Données

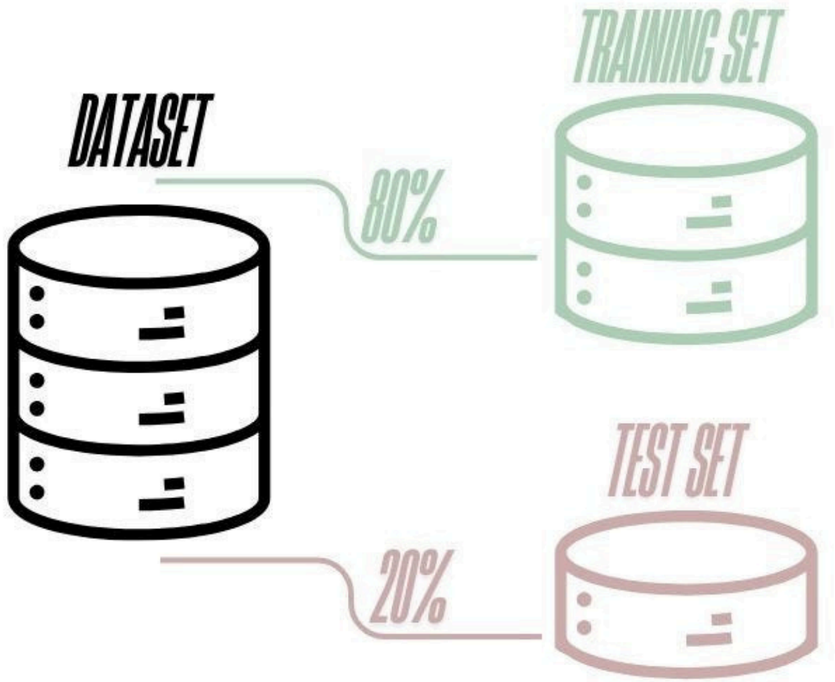
Nous avons d'abord défini :

- **X** comme l'ensemble des variables explicatives sélectionnées (features)
- **y** comme la variable cible (consommation)

Ensuite, nous avons séparé les données en deux sous-ensembles :

- **X_train / y_train** : 80 % des données utilisées pour entraîner les modèles
- **X_test / y_test** : 20 % des données utilisées pour évaluer les performances

Cette séparation permet de mesurer l'efficacité des modèles sur des données qu'ils n'ont jamais vues.



Random Forest

Random Forest est un algorithme supervisé basé sur un ensemble d’arbres de décision. Il combine plusieurs arbres formés sur des sous-échantillons aléatoires pour améliorer la précision et réduire le surapprentissage.

Types de Random Forest

peut résoudre deux types de problèmes :

- **Classification**
- **Régression**

Étapes clés de la Random

- **Bootstrap**
- **Sélection aléatoire de variables**

Modèle Random Forest en Régression

Dans ce projet, la Random Forest est utilisée pour modéliser une variable quantitative continue, la consommation électrique. Ce modèle repose sur la construction d'une multitude d'arbres de décision indépendants dont les prédictions sont agrégées par moyenne afin d'améliorer la robustesse et la précision.

L'évaluation des performances s'appuie sur des métriques standards :

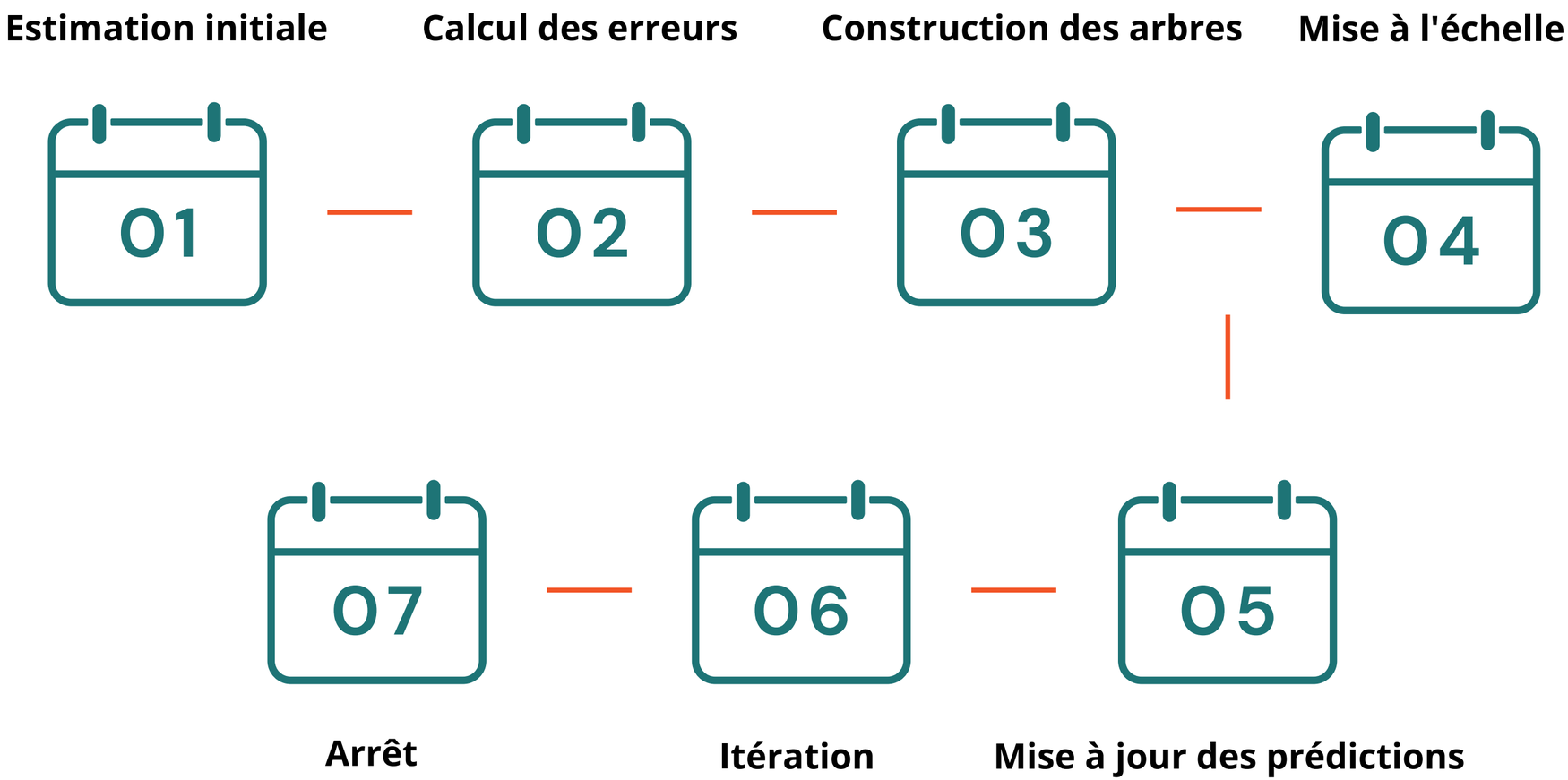
- MAE (Erreur Absolue Moyenne)
- RMSE (Erreur Quadratique Moyenne)
- R^2 (Coefficient de Détermination)

Afin d'illustrer ces résultats, des graphiques ont été tracés, notamment :

- Comparaison entre valeurs réelles et valeurs prédites
- Distribution des erreurs
- Visualisation de l'importance des variables explicatives
- Ces représentations graphiques facilitent l'interprétation et la validation visuelle de la performance du modèle.

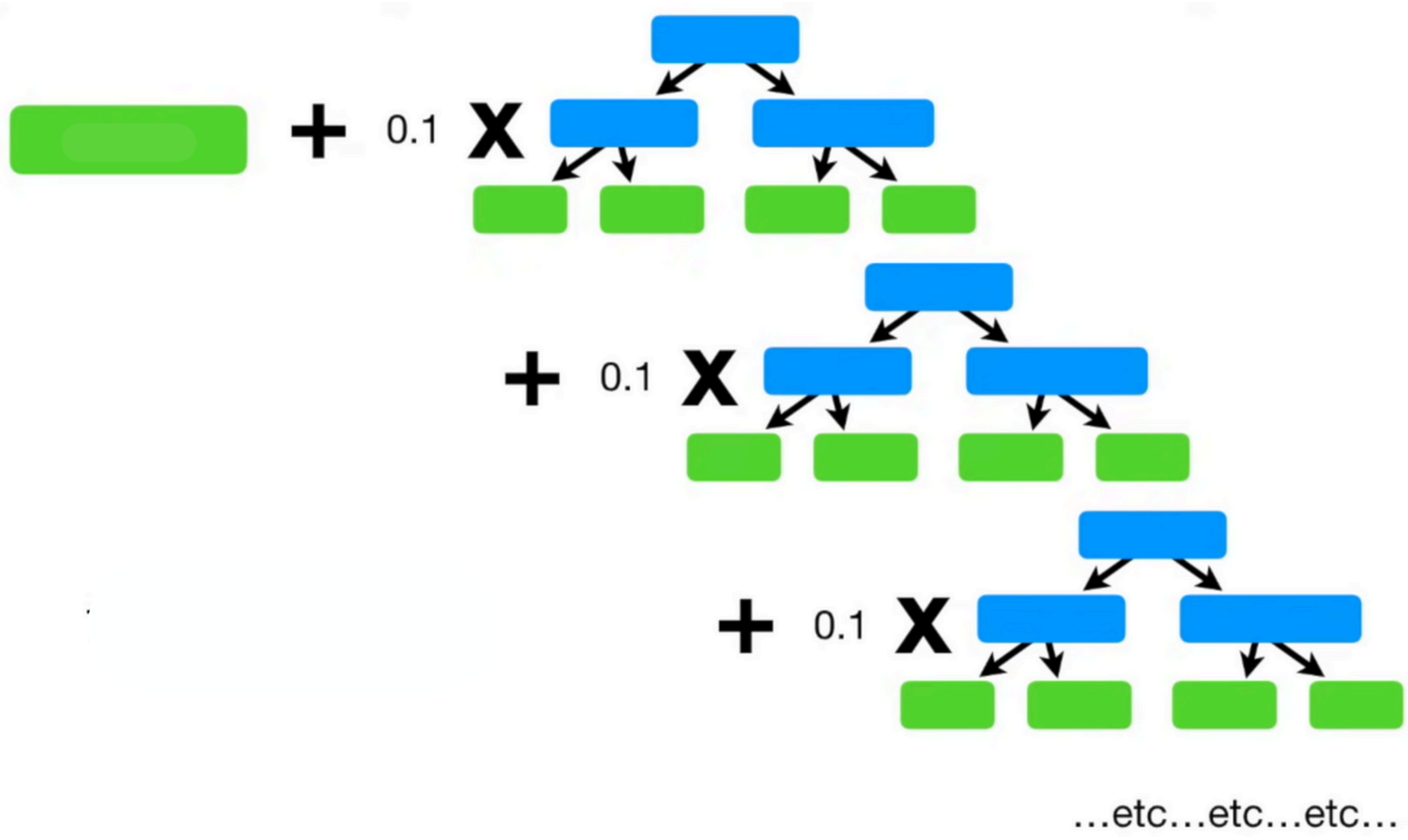
Concepts du Gradient Boosting

Le Gradient Boosting construit des arbres de décision de manière séquentielle, chaque nouvel arbre visant à corriger les erreurs commises par les arbres précédents.



Prédiction finale

Pour une nouvelle observation, la prédiction finale est obtenue en additionnant l'estimation initiale et les contributions mises à l'échelle de tous les arbres construits.



Concepts du Régression Polynomiale

La régression polynomiale est une forme de régression dans laquelle la relation entre la variable indépendante x et la variable dépendante y est modélisée comme un polynôme de degré n. Cela permet de capturer des relations non linéaires entre les variables, offrant ainsi une plus grande flexibilité que la régression linéaire simple.

Modèle Polynomial :

- Dans la régression polynomiale, le modèle est représenté par un polynôme. Par exemple, un polynôme de degré nn peut être écrit comme suit :

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \varepsilon$$

Processus de régression polynomiale

Transformation des variables :

- On génère de nouvelles variables à partir des variables originales en incluant leurs puissances et interactions jusqu'au degré polynomial choisi. Par exemple, pour un polynôme de degré 2, on inclut des termes comme x_1^2 et x_1x_2. Cela crée une matrice de variables transformées, également appelées caractéristiques étendues.

Calcul des coefficients :

- On résout l'équation matricielle suivante pour trouver les coefficients β:

$$\beta = (X^T X)^{-1} X^T y$$

Prédiction :

- Pour une nouvelle observation transformée x_new, on calcule la prédiction par :

$$\hat{y} = x_{\text{new}} \cdot \beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \cdots$$

Cas pratique

Comparaison des Performances des Modèles

R^2 : Mesure la proportion de variance expliquée par le modèle, variant de 0 à 1, où 1 indique un ajustement parfait.
MSE: Le MSE (Mean Squared Error) est la moyenne des carrés des erreurs. Sa racine carrée (RMSE) mesure l'écart moyen entre les prédictions et les valeurs réelles.
MAE : Moyenne des valeurs absolues des erreurs, indiquant l'erreur moyenne des prédictions.

Critère	XGBoost	Random Forest	Polynomial
Précision (MAE)	1627.96	1706.07	2569.1
Robustesse (MSE)	2098.85	4940132.43	3492.38
Explication (R^2)	0.881	0.866	0.67

XGBoost se distingue comme le meilleur compromis avec la plus haute précision, la meilleure robustesse aux valeurs extrêmes et le pouvoir explicatif le plus élevé. La régression polynomiale semble inadaptée à la complexité de ces données spécifiques.

Conclusion

Ce projet a permis de comparer trois modèles de régression supervisée pour la prédiction de la consommation électrique à partir de données météorologiques. Les résultats ont montré que les approches par ensemble, en particulier XGBoost et Random Forest, offrent une meilleure précision que la régression polynomiale. XGBoost s'est distingué comme le modèle le plus performant. Cette étude met en évidence l'importance du prétraitement des données, du choix des variables et de l'adéquation du modèle au problème.

MERCI

POUR VOTRE ATTENTION