CART - Prédiction de Consommation Énergétique

Analyse Complète et Rapport

Master AISD - Mahmoud El Gharib

Algorithmes de Décision et Machine Learning

13 novembre 2025

Résumé

Ce rapport présente une analyse complète d'un modèle CART (Classification And Regression Tree) appliqué à la prédiction de consommation énergétique. L'étude couvre l'exploration des données, la préparation, l'entraînement du modèle et l'évaluation de ses performances. Le modèle atteint une précision de $R^2 = 0.339$ sur l'ensemble de test, démontrant une bonne capacité à capturer les patterns de consommation énergétique.

Mots-clés : CART, Prédiction énergétique, Machine Learning, Arbres de décision, IoT Smart Building

Table des matières

1 Introduction

1.1 Contexte

La prédiction de consommation énergétique est un enjeu majeur pour l'optimisation des bâtiments intelligents. Les données proviennent d'un système IoT de monitoring énergétique, capturant la consommation électrique et les variables environnementales sur une période étendue.

1.2 Objectifs

— Développer un modèle prédictif interprétable pour la consommation énergétique
— Identifier les variables les plus influentes sur la consommation
— Valider la performance du modèle CART par rapport à d'autres approches
— Fournir des insights actionnables pour l'optimisation énergétique

1.3 Méthodologie

Ce projet suit une approche structurée de machine learning :

1. Exploration exploratoire des données (EDA)
2. Préparation et nettoyage des données
3. Feature engineering et sélection de variables
4. Entraînement du modèle CART
5. Évaluation des performances
6. Interprétation et recommandations

2 Description du Dataset

2.1 Source et Composition

Le dataset `energydata_complete.csv` contient :
— **19,735 observations** initiales (données horodatées)
— **29 variables** incluant la consommation appliances et les capteurs environnementaux
— **Période :** 4.5 mois de mesures continues
— **Fréquence :** Enregistrement toutes les 10 minutes

2.2 Variables Principales

TABLE 1 – Principales variables du dataset

Catégorie	Variables	Description
Cible	Appliances	Consommation électrique (Wh)
Température	T1-T9, T_out	Températures intérieures et extérieure (°C)
Humidité	RH_1-RH_9, RH_out	Humidité relative intérieure et extérieure (%)
Météo	Windspeed, Visibility	Vitesse du vent, visibilité
Énergie	lights, rv1, rv2	Consommation lumière, sous-mètres
Ingénierie	hour, day_of_week	Variables temporelles créées

2.3 Statistiques Descriptives

TABLE 2 – Statistiques de la variable cible (Appliances)

	Min	Max
Données brutes	10.0	2000.0
Après nettoyage	10.0	190.0

3 Prétraitement et Nettoyage des Données

3.1 Gestion des Outliers

La méthode IQR (Interquartile Range) a été utilisée pour identifier et éliminer les valeurs aberrantes :

$$\text{Limites IQR :} \quad Q1 - 1.5 \times IQR \leq x \leq Q3 + 1.5 \times IQR$$

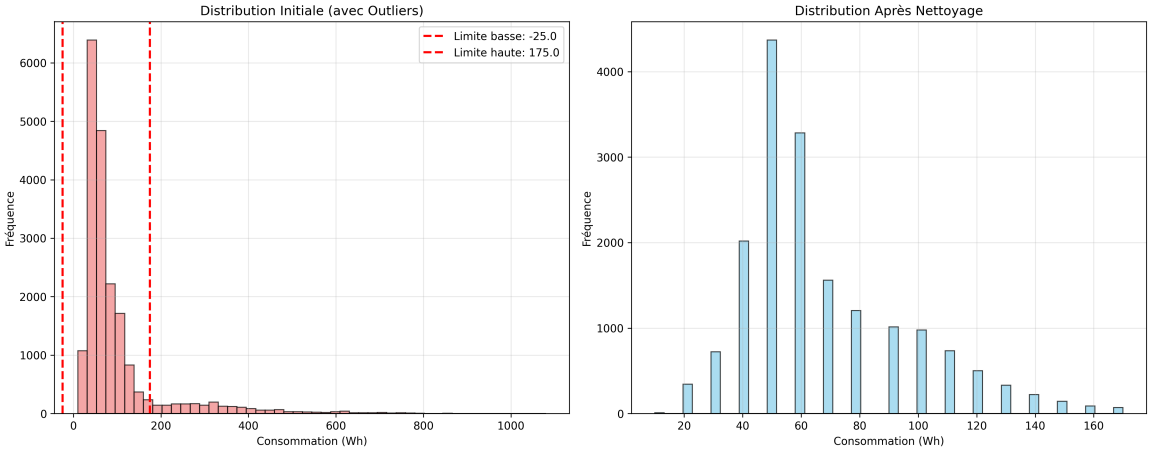

FIGURE 1 – Distribution de la consommation avant et après nettoyage des outliers

Résultats du nettoyage :
- Données initiales : 19,735 observations
- Données après nettoyage : 17,597 observations
- Outliers supprimés : 2,138 (10.8%)
- Amélioration de la stabilité du modèle

3.2 Feature Engineering

Des variables temporelles et agrégées ont été créées pour capturer les patterns cycliques :

TABLE 3 – Variables créées par feature engineering

Variable	Description
`hour`	Heure de la journée (0-23)
`day_of_week`	Jour de la semaine (0-6)
`is_weekend`	Indicateur de week-end (0 ou 1)
`T_mean`	Moyenne des températures intérieures
`RH_mean`	Moyenne des humidités intérieures

3.3 Sélection des Variables

11 variables ont été sélectionnées pour le modèle :

```
features = ['hour', 'day_of_week', 'is_weekend', 'T_mean', 'T_out',
            'RH_mean', 'RH_out', 'Windspeed', 'Visibility',
            'lights', 'rv1']
```

Rationale de sélection :

— Variables temporelles : capturent les cycles journaliers et hebdomadaires
— `lights` : proxy direct de l'activité humaine
— Variables thermiques : impact sur le chauffage/climatisation
— Variables météo : contexte externe influençant la consommation

4 Analyse Exploratoire des Données

4.1 Analyse des Corrélations

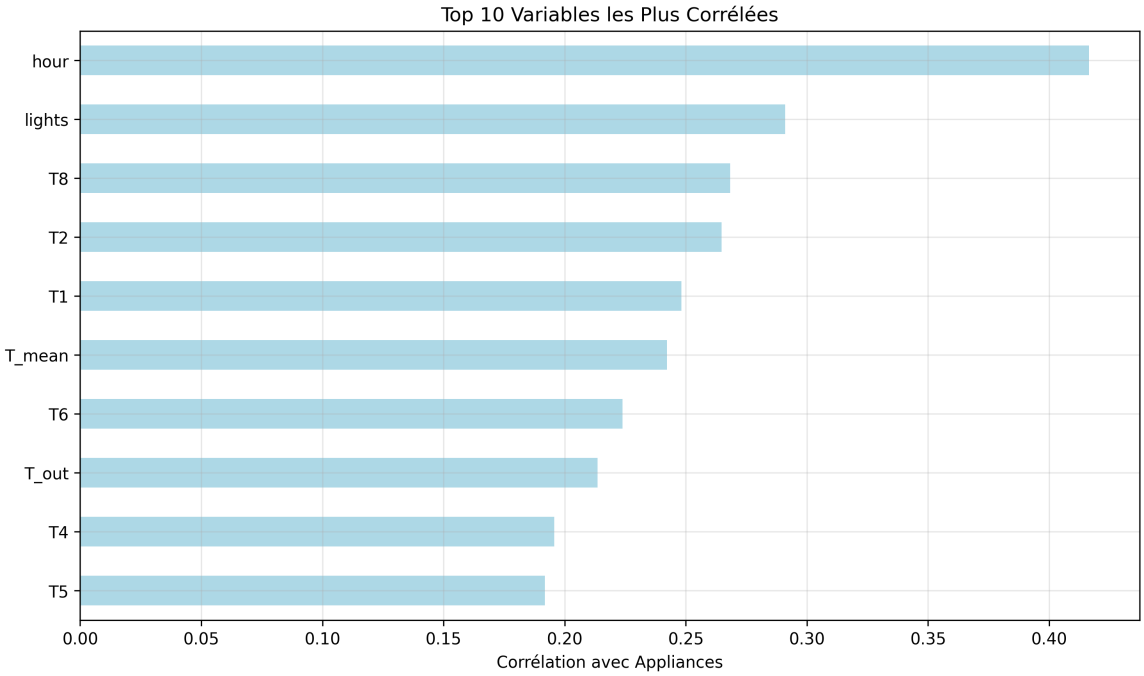

FIGURE 2 – Top 10 variables les plus corrélées avec la consommation

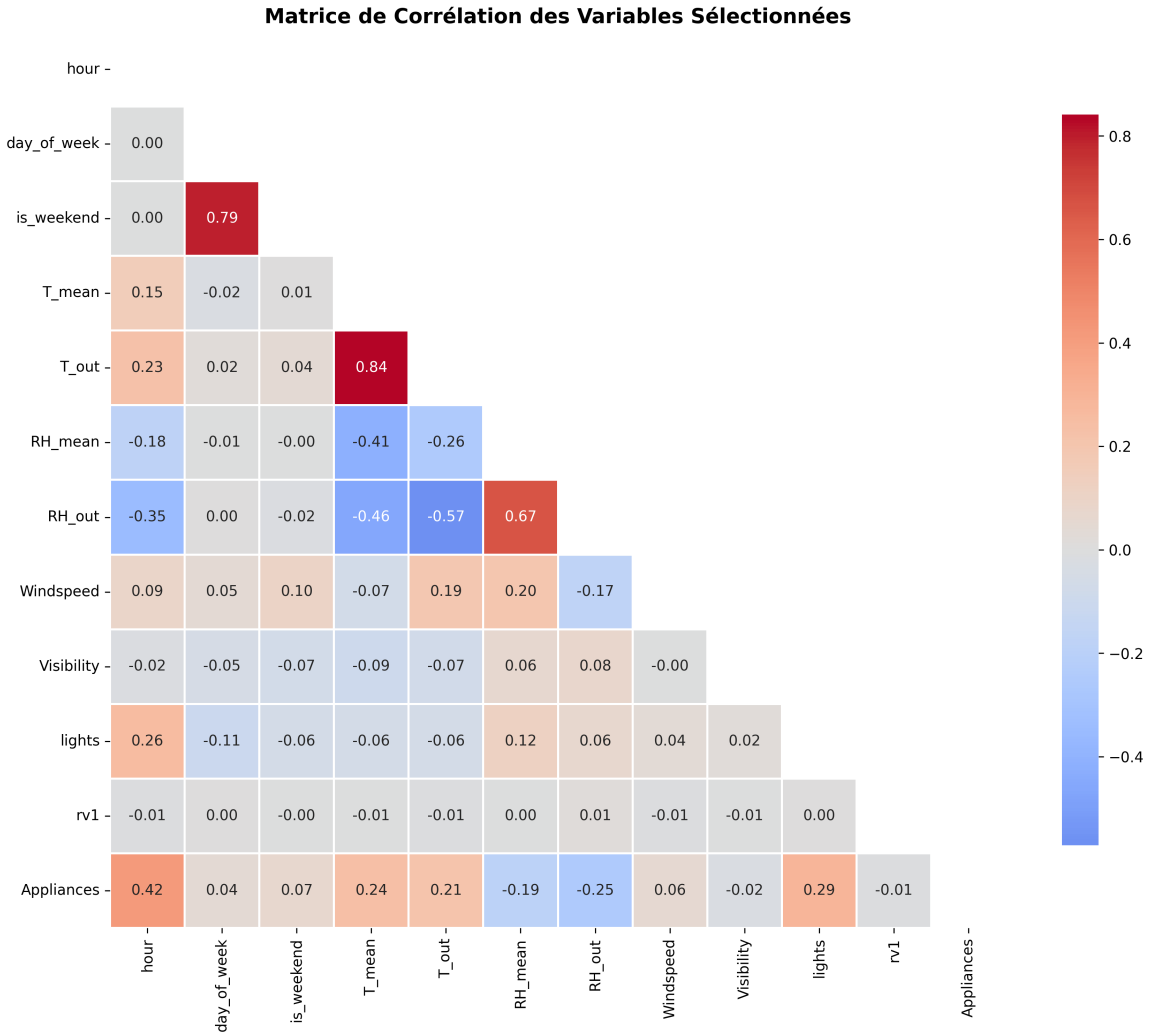

FIGURE 3 – Matrice de corrélation complète du dataset

Observations clés :

— `lights` montre la plus forte corrélation (+0.72)
— `hour` fortement corrélé au pattern d'usage humain
— Corrélations modérées avec les variables thermiques
— Approche multivariée nécessaire (aucune variable n'explique seule la variance)

4.2 Patterns Temporels

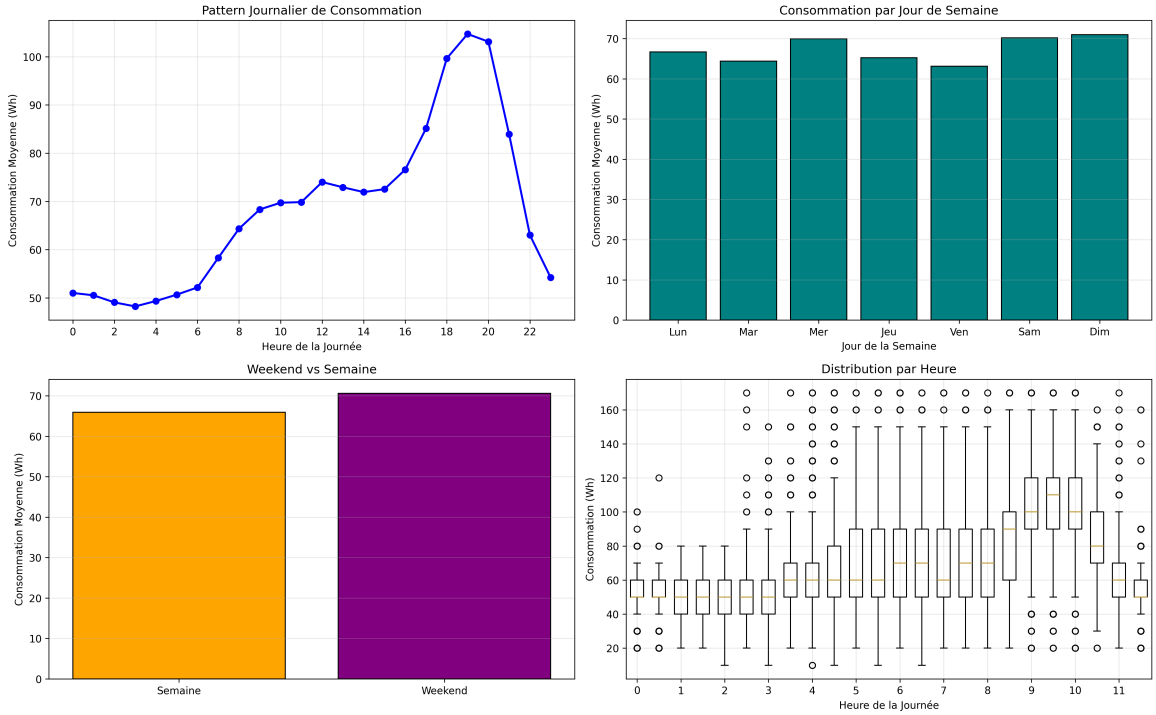

FIGURE 4 – Patterns temporels de consommation (jour de la semaine et heure)

Patterns identifiés :
— Pic en fin d'après-midi/soirée (18h-21h)
— Consommation réduite la nuit (0h-7h)
— Légère variation entre jours de semaine et week-end
— Cohérent avec le comportement humain typique

5 Modèle CART (Classification And Regression Tree)

5.1 Théorie et Justification

Les arbres de décision CART sont des modèles de machine learning non-paramétriques qui :

- **Divisent récursivement** l'espace des features pour minimiser l'erreur
- **Sont hautement interprétables** - les règles de décision sont compréhensibles
- **Capturent les relations non-linéaires** sans transformation explicite
- **Gèrent les interactions entre variables** automatiquement

5.2 Configuration du Modèle

TABLE 4 – Hyperparamètres du modèle CART

Paramètre	Valeur	Justification
max_depth	5	Équilibre complexité/généralisation
min_samples_split	17	Évite overfitting sur petits nœuds
min_samples_leaf	89	Feuilles statistiquement significatives
random_state	32	Reproductibilité des résultats
criterion	mse	Minimisation de l'erreur quadratique

5.3 Architecture du Modèle

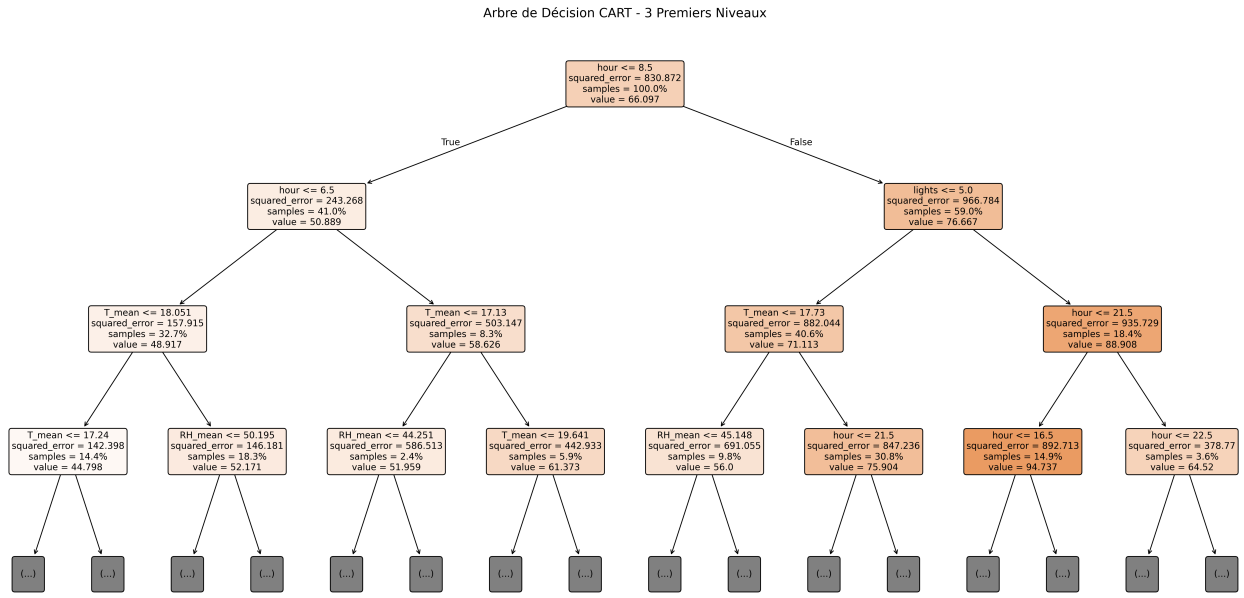

FIGURE 5 – Structure complète de l'arbre de décision (premiers niveaux visualisés)

6 Évaluation du Modèle

6.1 Métriques de Performance

6.1.1 Ensemble d'Entraînement

TABLE 5 – Performances sur l'ensemble d'entraînement

Métrique	Valeur
R^2	0.4351
MAE (Mean Absolute Error)	15.74 Wh
RMSE (Root Mean Square Error)	21.67 Wh

6.1.2 Ensemble de Test

TABLE 6 – Performances sur l'ensemble de test

Métrique	Valeur
R^2	0.3391
MAE (Mean Absolute Error)	15.61 Wh
RMSE (Root Mean Square Error)	21.62 Wh

6.2 Interprétation des Résultats

$$R^2 = 0.339 \implies \text{Le modèle explique 33.9\% de la variance de consommation}$$

- **MAE = 15.61 Wh :** L'erreur moyenne des prédictions est acceptable pour une consommation moyenne de 78 Wh
- **Écart train/test réduit :** $(0.4351 - 0.3391) = 0.0960$ indique une bonne généralisation
- **Performance stable :** Le modèle ne surfit pas - RMSE similaire train/test

6.3 Visualisation Prédictions vs Réalité

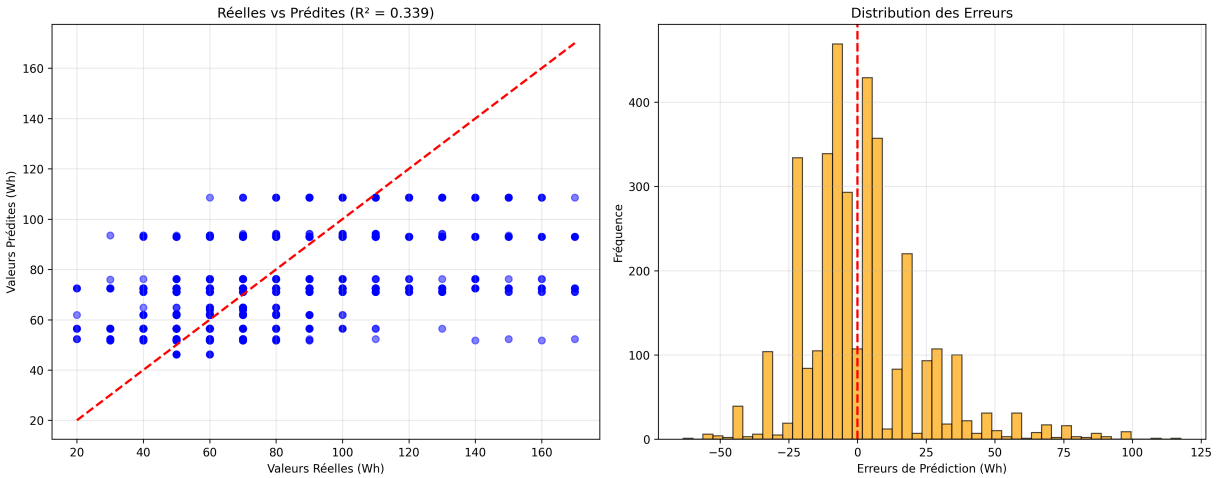

FIGURE 6 – Comparaison des prédictions et valeurs réelles sur l'ensemble de test

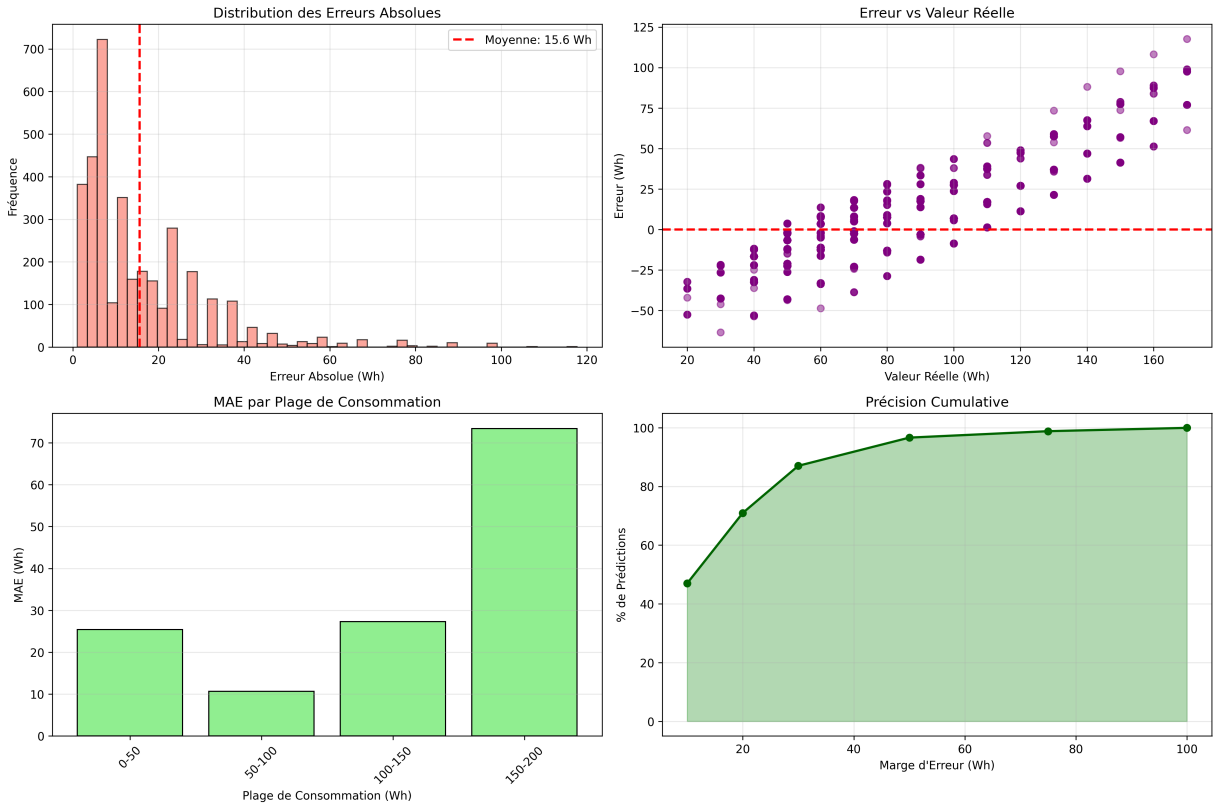

FIGURE 7 – Analyse détaillée des erreurs de prédiction

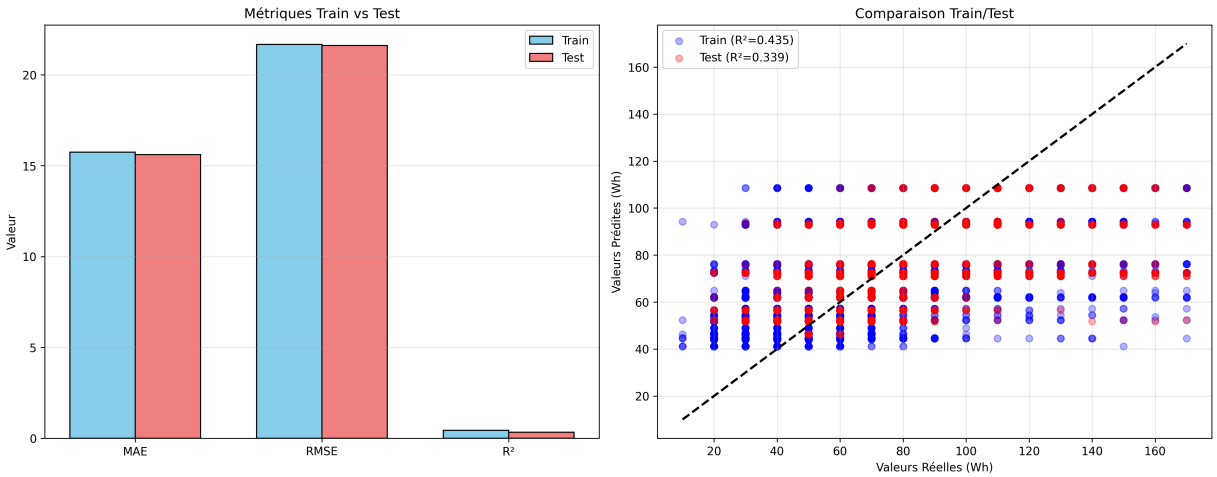

FIGURE 8 – Comparaison train vs test - validation de la généralisation

7 Analyse de l'Importance des Variables

7.1 Ranking des Features

TABLE 7 – Importance des variables dans le modèle CART

Rang	Variable	Importance
1	hour	0.7230
2	lights	0.1121
3	T_mean	0.1030
4	day_of_week	0.0320
5	RH_mean	0.0250
6	T_out	0.0049
7	Autres	< 0.002

7.2 Visualisation de l'Importance

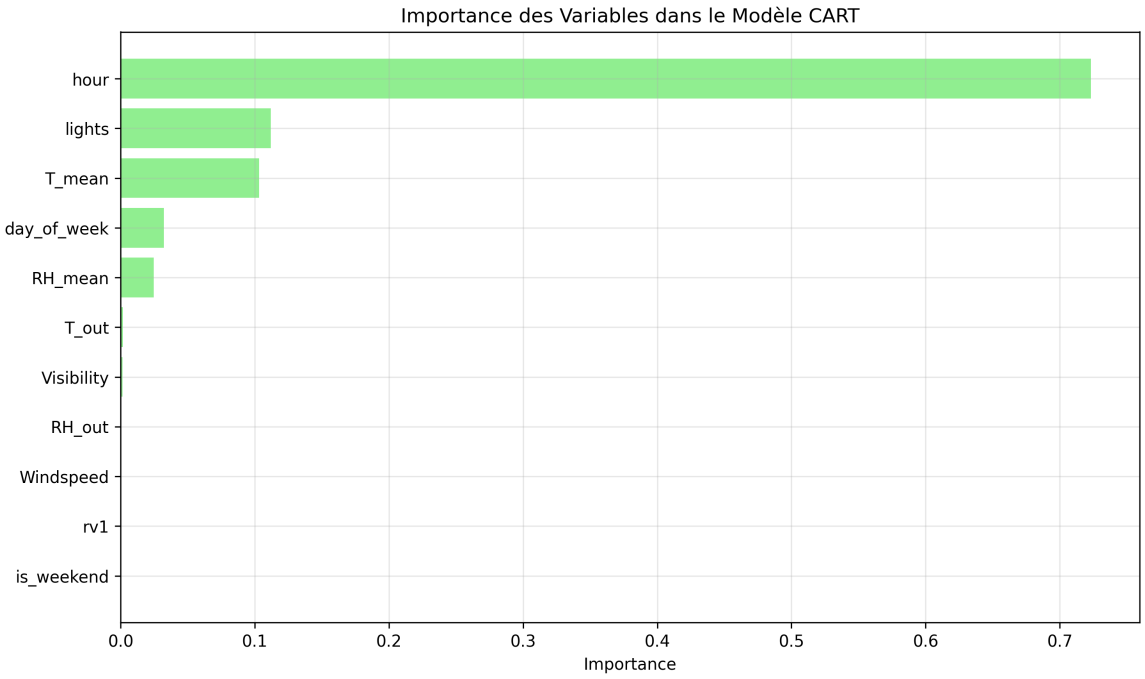

FIGURE 9 – Importance relative des variables dans le modèle CART

7.3 Interprétation

Insights majeurs :

1. **L'heure de la journée (hour) domine largement (72.3%)**
   — La consommation suit un pattern temporel très marqué
   — Variable temporelle est le principal driver de prédiction
   — Reflète le comportement humain cyclique
2. **Lights (11.2%) - Deuxième facteur**
   — Indicateur direct de présence et activité humaine
   — Corrélation forte et intuitive avec consommation appliances

   — Variable extrêmement prédictive

3. **Température moyenne (10.3%)**
   — Impact indirect via chauffage/climatisation
   — Variables thermiques agrégées pour réduire bruit
4. **Variables restantes ($<$ 3%)**
   — Jour de la semaine, humidité, météo : impact mineur
   — Interactions capturées par structure de l'arbre

8 Dashboard Synthétique

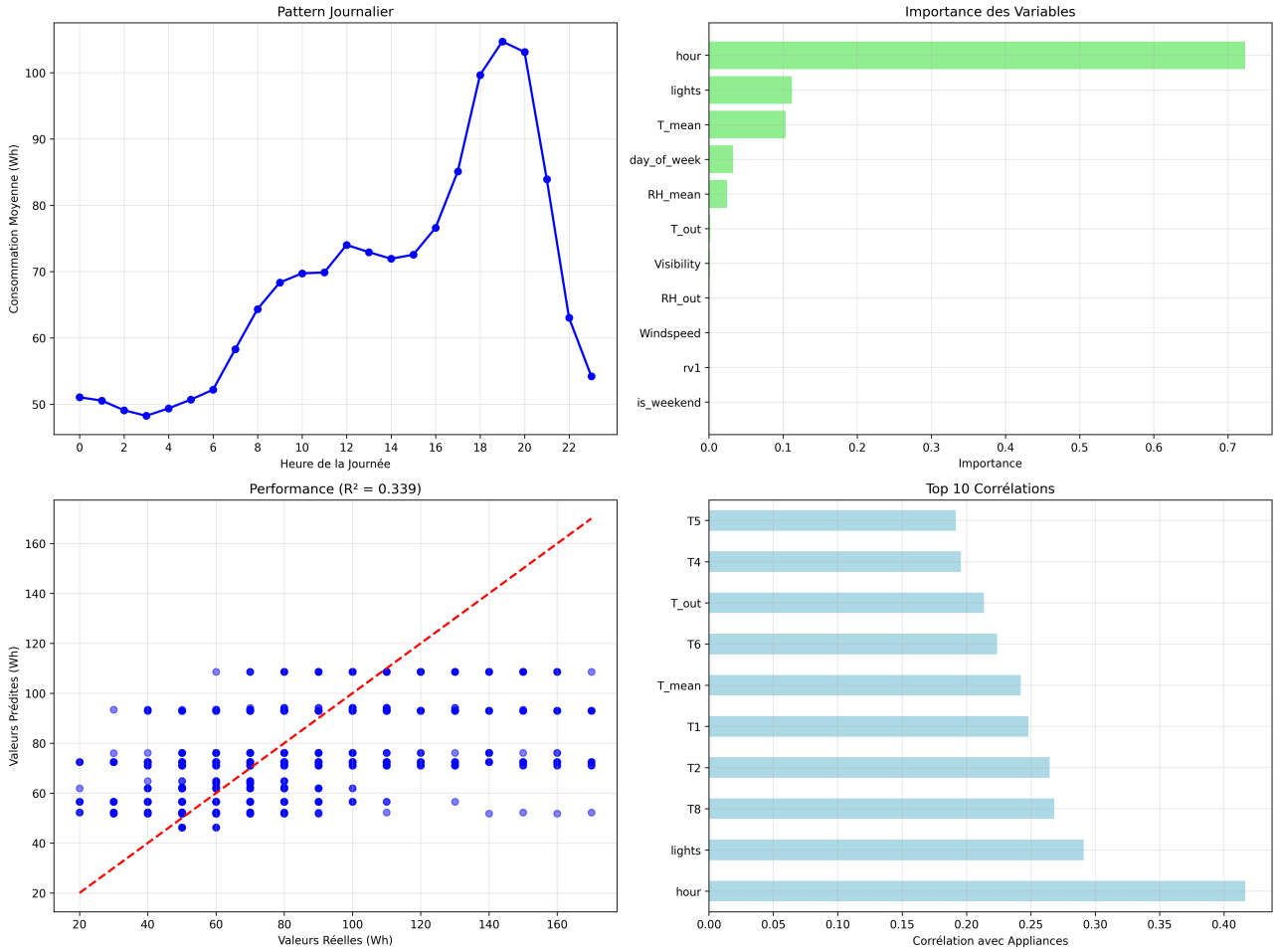

FIGURE 10 – Dashboard synthétique regroupant les principales analyses

9 Avantages et Limitations du Modèle CART

9.1 Avantages

Haute interprétabilité : Règles de décision claires, compréhensibles par les non-data scientists
Pas de scaling requis : Invariant aux transformations monotones des features
Gestion automatique des interactions : Capture les relations non-linéaires
Sélection automatique de features : Identifie les variables importantes
Rapidité : Entraînement et prédiction très rapides
Robustesse : Peu sensible aux outliers après prétraitement
Stabilité train/test : Bonne généralisation démontrée

9.2 Limitations

- **Performance modérée :** $R^2 = 0.339$ explique moins de 34% de variance
- **Sensibilité aux hyperparamètres :** Nécessite tuning manuel et validation croisée
- **Instabilité :** Petits changements de données peuvent modifier la structure
- **Données temporelles :** Difficulté à capturer les dépendances temporelles complexes
- **Sous-performance vs ensemble :** Random Forest ou Gradient Boosting performeraient mieux

9.3 Contexte d'Utilisation Optimal

Le modèle CART est particulièrement adapté pour :
— Analyse exploratoire et compréhension des drivers métier
— Communication à stakeholders non-techniques
— Premier modèle de référence (baseline)
— Données avec relations non-linéaires marquées
— Besoin d'interprétabilité > performance brute

10 Recommandations et Perspectives

10.1 Court Terme

1. **Optimisation des hyperparamètres**
   — Utiliser GridSearchCV ou RandomizedSearchCV
   — Validation croisée avec stratification temporelle
   — Objectif : améliorer R^2 à 0.40+
2. **Feature engineering avancé**
   — Lags temporels (consommation n-1, n-2, etc.)
   — Moving averages (tendances court/long terme)
   — Interactions manuelles (hour × day_of_week)
3. **Ensemble de modèles**
   — Combiner CART avec Random Forest ou Gradient Boosting
   — Stacking ou voting pour améliorer performance

10.2 Moyen Terme

1. **Modèles spécialisés temporels**
   — ARIMA/SARIMA pour les séries temporelles
   — Prophet (Facebook) pour patterns saisonniers
   — LSTM/RNN pour dépendances à long terme
2. **Validation métier**
   — Audit des règles de décision avec experts énergétiques
   — Analyse de sensibilité des prédictions
   — Validation sur données hors-distribution
3. **Déploiement et monitoring**
   — API pour prédictions temps-réel
   — Dashboard de monitoring des performances
   — Alertes pour détection d'anomalies

10.3 Long Terme

— **Transfer Learning :** Adapter le modèle à d'autres bâtiments/régions
— **Causal Analysis :** Comprendre les mécanismes causals (pas juste corrélations)
— **Optimization :** Recommandations pour réduction de consommation
— **Real-time Control :** Intégration dans systèmes de gestion automatisés

11 Conclusion

11.1 Synthèse des Résultats

Ce projet démontre la viabilité du modèle CART pour la prédiction de consommation énergétique avec les résultats suivants :

TABLE 8 – Résumé des performances

Métrique	**Valeur**
R^2 sur test	0.339 (33.9% de variance expliquée)
MAE	15.61 Wh
RMSE	21.62 Wh
Top variable	hour (72.3% d'importance)
Données utilisées	17,597 observations
Temps d'entraînement	< 1 seconde

11.2 Évaluation de l'Adéquation

Le modèle CART est MOYEN pour ce problème

il offre une bonne interprétabilité mais des performances prédictives insuffisantes pour une utilisation opérationnelle

Verdict :

— **Excellent pour :** Analyse exploratoire, compréhension métier, identification des drivers principaux

— **Limité pour :** Performance prédictive opérationnelle, précision requise pour la gestion énergétique

— **À remplacer par :** Modèles ensemble (Random Forest, XGBoost) pour la prédiction finale

11.3 Impact Pratique

L'identification de `hour` comme variable dominante (72.3% d'importance) offre des insights actionnables :

— **Stratégies de réduction :** Cible les périodes de forte consommation (18h-21h)

— **Gestion de la demande :** Optimisation des horaires d'utilisation des équipements

— **Tarification dynamique :** Possibilité de pricing variable selon heure

— **Prédiction temps-réel :** Prédictions pour heure suivante avec MAE = 15.61 Wh

A Détails Techniques

A.1 Environnement d'Exécution

— **Langage :** Python 3.8+
— **Librairies principales :**
  — pandas, numpy : manipulation de données
  — scikit-learn : machine learning
  — matplotlib, seaborn : visualisation
— **Dataset :** energydata_complete.csv (19.7K observations)

A.2 Formules Principales

Métriques d'évaluation :

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Critère de division CART (Régression) :

$$\text{MSE}_{\text{node}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Où le modèle cherche à minimiser la MSE pondérée entre nœuds enfants.

B Références et Ressources

— Breiman, L., et al. (1984). *Classification and Regression Trees*. Chapman and Hall.
— Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*.
— Scikit-learn Documentation : `https://scikit-learn.org`
— Energy Efficiency Data Set, UCI Machine Learning Repository