REPUBLIQUE ALGERIENNE DEMOCRATIQUE

ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEURE ET DE LA RECHERCHE SCIENTIFIQUE

Université 20 Aout 1955- Skikda

Faculté des Sciences

Département informatique

Mémoire

En vue de l'obtention du diplôme de Master

Filière: Informatique

Spécialité: Intelligence Artificielle

Prévision de la production d'énergie solaire à l'aide d'algorithmes d'apprentissage automatique

---

Présenté par:

Taabani Kawther

Superviseur: Dr. ADEL Lahsasna

Juin 2025

Remerciements

Bismillah Ar-Rahman Ar-Rahim

Louange à Allah par la grâce de qui les bonnes choses s'accomplissent, et par Son aide les efforts sont couronnés de succès, et les parcours prennent fin.

En ce moment charnière de ma vie, où je récolte le fruit de **dix-sept années d'études**, je tiens à exprimer ma profonde gratitude à tous ceux qui ont laissé une empreinte sur ce chemin — par leur savoir, leur soutien ou leurs prières.

J'adresse mes sincères remerciements et ma profonde reconnaissance à mon honorable enseignant,
le Dr. Adel Lehssasna,
un professeur engagé, rigoureux et véritablement guide, qui n'a jamais ménagé ses conseils scientifiques et critiques. Votre accompagnement a été pour moi un véritable appui tout au long de ce parcours. Les mots sont bien insuffisants pour exprimer toute ma gratitude.

Je tiens également à remercier tous les enseignants qui partagent leur savoir avec le cœur, et qui diffusent la lumière de la science avec sincérité. Merci à chacun de ceux qui ont contribué à l'édification de ce noble édifice du savoir.

Je n'oublie pas non plus le personnel administratif et les agents qui, dans l'ombre, travaillent avec dévouement et constance. Vous méritez toute ma considération et mon respect.

À **l'Université du 20 Août 1955 – Skikda**,
qui m'a accueillie pendant **cinq années**, riche en apprentissages, en expériences humaines, et en souvenirs inoubliables — vous avez été un lieu formateur, tant sur le plan académique qu'humain.

Un remerciement particulier à **ma mère, ma famille et mes amies**, véritables piliers de mon équilibre, mon refuge dans les moments difficiles, et ma source de force dans les instants de doute.

Je tiens également à exprimer toute ma reconnaissance pour mon expérience en tant que **présidente du Club de Pensée Scientifique et Culturelle**, un foyer intellectuel qui m'a permis de grandir, d'oser, et de me découvrir à travers de multiples opportunités.

Et enfin ce travail est le fruit d'un long labeur, de nuits blanches et de souvenirs ineffaçables.
Merci à tous ceux qui ont marché à mes côtés, même en silence, vers cette étape de ma vie

Dédicaces

Bismillah Ar-Rahman Ar-Rahim

Louange à Allah qui m'a accordé la réussite, facilité les chemins, et béni mes efforts. À Lui toute la gratitude, au début et à la fin, dans les instants visibles comme dans les silences profonds.

On commence souvent une dédicace par la mère. Mais aujourd'hui, ce n'est pas une simple tradition, c'est un hommage sincère à celle dont l'amour, les sacrifices et la patience sont inégalables...

À **ma chère maman**, sans qui rien n'aurait été possible. Tu as toujours été la prière silencieuse, le refuge sûr, la main qui me relève à chaque chute. Aujourd'hui, ton rêve devient réalité… ta fille tient enfin le fruit de ton dévouement.

À **mon père bien-aimé**, mon premier maître dans la vie, la boussole de mes valeurs. Celui qui m'a appris que la volonté ne se brise jamais, et que la dignité est un principe sacré.

À **ma grand-mère Khirfia**, refuge de mon cœur, chaleur de mon âme. Tu as été une seconde mère, ton amour une ombre douce, et tes prières un secret entre le ciel et moi. Qu'Allah te garde auprès de moi.

À **Mohamed Amine** et **Youssef**, mes yeux, mon cœur, mes frères et mon pilier dans cette vie. Vos présences silencieuses me donnent la force et la volonté d'avancer.

À **Mouhi Eddine**, frère et compagnon, avec qui j'ai partagé les jours et les détails de ce long chemin.

À **ma tante Nadia**, chère à mon cœur, au cœur maternel. Ta présence m'apaise et me réjouit comme nul autre.

À **toute ma famille**, ce socle solide qui a soutenu chacun de mes pas par leurs prières et leur affection.

À **mes amies chères**, compagnes de route et sœurs de cœur, avec qui j'ai partagé les rêves, les difficultés, les sourires, et tant de souvenirs précieux.

À **mon foyer universitaire, le Club de Pensée Scientifique et Culturelle**, qui m'a offert des opportunités uniques et façonné une part de mon identité.

À tous ceux qui ont marché à mes côtés dans le silence, mais dont la lumière a guidé mes pas dans l'obscurité…

À **mon âme épuisée**, mais persévérante, qui a cru que derrière chaque épreuve se cache une délivrance, et qu'après chaque larme, le soleil finit toujours par briller.

Je vous dédie ce travail modeste, chargé de souvenirs, d'insomnies, de luttes intimes… À tous ceux qui ont, même en silence, été un maillon essentiel de ce chemin vers l'accomplissement.

Liste des abréviations

Ce mémoire fait usage de plusieurs abréviations techniques couramment rencontrées dans les domaines de l'énergie photovoltaïque, de l'intelligence artificielle et de l'apprentissage automatique. Voici les principales abréviations employées :

AI : *Artificial Intelligence*, désigne l'intelligence artificielle.

ANN : *Artificial Neural Network*, réseau de neurones artificiels.

CV : *Cross-Validation*, validation croisée utilisée pour évaluer la robustesse des modèles.

EMS : *Energy Management System*, système de gestion de l'énergie.

G, Gb, Gd, Gr : termes désignant respectivement l'irradiance globale, directe, diffuse et réfléchie.

IoT : *Internet of Things*, désigne l'Internet des objets.

kNN : *k-Nearest Neighbors*, algorithme des k plus proches voisins.

LSTM : *Long Short-Term Memory*, réseau neuronal séquentiel utilisé pour les séries temporelles.

MAE : *Mean Absolute Error*, erreur absolue moyenne.

ML : *Machine Learning*, apprentissage automatique.

MSE : *Mean Squared Error*, erreur quadratique moyenne.

PAC : Puissance active instantanée produite par le système photovoltaïque (en watts).

PR : *Performance Ratio*, indicateur de rendement des installations solaires.

PV : *Photovoltaic*, terme relatif aux technologies solaires photovoltaïques.

R^2 : Coefficient de détermination, indique la proportion de variance expliquée par un modèle.

RBF : *Radial Basis Function*, fonction de base radiale utilisée comme noyau dans le SVR.

RF : *Random Forest*, algorithme ensembliste basé sur des arbres de décision.

RMSE : *Root Mean Squared Error*, racine de l'erreur quadratique moyenne.

RNN : *Recurrent Neural Network*, réseau de neurones récurrent.

SCADA : *Supervisory Control and Data Acquisition*, système de supervision et d'acquisition de données.

SHAP : *SHapley Additive exPlanations*, méthode d'explication locale des modèles d'IA.

SVR : *Support Vector Regression*, régression à vecteurs de support.

SVM : *Support Vector Machine*, machine à vecteurs de support.

XGBoost : *eXtreme Gradient Boosting*, algorithme avancé de boosting de gradient.

Abstract

This thesis focuses on the short-term forecasting of photovoltaic (PV) energy production using machine learning techniques. Given the intermittency of solar energy due to fluctuating meteorological conditions, accurate forecasting is essential for ensuring grid stability and optimizing energy management. The study explores the application of three supervised learning models: Support Vector Regression (SVR), eXtreme Gradient Boosting (XGBoost), and Random Forest Regressor (RF).

A real-world dataset, collected from a PV system with 5-minute interval measurements, was used. A rigorous preprocessing pipeline was applied, including data cleaning, feature engineering (e.g., temporal variables), and normalization. The models were trained and tested using standard regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). Among the three models, Random Forest achieved the highest accuracy with an R^2 of 0.999996, outperforming both XGBoost and SVR.

The results demonstrate that well-prepared data and properly tuned machine learning models can yield highly accurate solar power predictions. This work offers practical insights for improving energy planning, enhancing PV system integration into the grid, and supporting the development of intelligent energy management systems.

Keywords: Photovoltaic forecasting, solar energy, machine learning, Random Forest, XGBoost, SVR, short-term prediction, data preprocessing, regression models, renewable energy.

Résumé

Ce mémoire porte sur la prévision à court terme de la production d'énergie photovoltaïque (PV) à l'aide d'algorithmes d'apprentissage automatique. En raison de la variabilité des conditions météorologiques, la prévision précise de la production solaire est un enjeu clé pour la stabilité des réseaux électriques et l'optimisation de la gestion énergétique.

Trois modèles supervisés ont été étudiés : la régression par vecteurs de support (SVR), XGBoost, et Random Forest Regressor (RF). Les données utilisées proviennent d'un système PV réel, avec des mesures enregistrées toutes les 5 minutes. Un prétraitement rigoureux a été appliqué : nettoyage, transformation des variables temporelles et normalisation. L'évaluation des modèles a été réalisée à l'aide de trois indicateurs : MAE, RMSE et R^2. Le modèle Random Forest s'est révélé le plus performant, atteignant un R^2 de 0,999996.

Les résultats obtenus montrent que l'application rigoureuse de méthodes de machine learning, combinée à un traitement soigné des données, permet d'obtenir des prévisions très fiables. Ce travail apporte une contribution concrète à l'amélioration de l'intégration des énergies renouvelables dans les réseaux et à la planification énergétique intelligente.

Mots-clés : Prévision photovoltaïque, énergie solaire, apprentissage automatique, Random Forest, XGBoost, SVR, prédiction à court terme, prétraitement des données, modèles de régression, énergies renouvelables.

.

Table des matières

Sommaire

Contenu

Introduction

Contexte du projet

Face aux enjeux environnementaux croissants et à l'épuisement des ressources fossiles, les énergies renouvelables occupent une place de plus en plus centrale dans les politiques énergétiques mondiales. Parmi elles, l'énergie solaire se distingue par son accessibilité, son potentiel inépuisable et la baisse progressive des coûts de production des systèmes photovoltaïques (PV). Selon l'Agence internationale de l'énergie (AIE), la capacité mondiale de production solaire photovoltaïque pourrait doubler d'ici 2030, portée par l'évolution technologique et les politiques publiques incitatives [1].

Cependant, la production d'énergie solaire est soumise à une variabilité importante due aux conditions météorologiques locales telles que la couverture nuageuse, l'irradiation, la température ou le vent. Cette variabilité pose un défi majeur pour l'équilibrage des réseaux électriques, la planification de la charge et la stabilité des systèmes énergétiques.

Problématique

L'intermittence de l'énergie solaire rend difficile son intégration directe et en toute fiabilité dans les réseaux de distribution. Pour répondre à ce défi, il est essentiel de disposer d'outils capables de prédire avec précision la production solaire à court terme. Une prevision finable permit:

- d'**optimiser** l'utilisation des ressources énergétiques,
- de mieux **dimensionner** les capacités de stockage ou de secours,
- et de **renforcer** l'efficacité des réseaux intelligents (smart grids).

Les approches traditionnelles de prévision, basées sur des modèles statistiques (ARIMA, régression linéaire) ou physiques (modèles radiatifs, données satellite), montrent des limites dans la prise en compte de la non-linéarité et de la complexité des interactions climatiques.

Apport de l'apprentissage automatique

L'apprentissage automatique (machine learning) émerge comme une alternative puissante, capable de modéliser les relations complexes entre les variables climatiques et la production énergétique. Grâce à la disponibilité croissante de données historiques de qualité et à l'amélioration des performances informatiques, plusieurs modèles basés sur le machine learning — comme les forêts aléatoires (Random Forest), les modèles de boosting (XGBoost), ou les réseaux de neurones — ont montré leur efficacité dans le domaine de la prévision énergétique [2][3].

Ces méthodes permettent non seulement d'améliorer la précision des prévisions, mais aussi d'adapter les modèles aux spécificités locales d'un site photovoltaïque.

Objectifs du mémoire

Ce travail de recherche vise à :

- **Comparer plusieurs modèles d'apprentissage automatique supervisé** appliqués à la prévision de la puissance active générée par un système photovoltaïque
- **Évaluer les performances** de chaque modèle à partir de données réelles mesurées localement
- **Identifier l'algorithme le plus adapté** en termes de précision et de robustesse pour une application de prévision à court terme
- **Analyser et visualiser les résultats** obtenus afin d'en extraire des conclusions opérationnelles.

Démarche adoptée

Pour répondre à ces objectifs, une méthodologie rigoureuse a été mise en place :

- **Traitement de données mesurées** toutes les 5 minutes sur une installation réelle ;
- **Sélection de variables explicatives pertinentes** ;
- **Application de plusieurs algorithmes** (régression linéaire, SVR, Random Forest) ;
- **Validation croisée et évaluation** selon des métriques standards (MAE, RMSE, R^2).

Organisation du mémoire

Ce mémoire est structuré comme suit :

- **Introduction Générale**
- **Chapitre 1** : Des généralités sur l'énergie solaire et les technique d'apprentissage.
- **Chapitre 2** : Revue de la littérature sur les méthodes de prévision solaire et les algorithmes de machine Learning utilisés dans ce domaine.
- **Chapitre 3** : Présentation de la méthodologie expérimentale, de la base de données et des modèles implémentés.
- **Chapitre 4** : Présentation des résultats, comparaison des performances des modèles et interprétation.
- **Conclusion Générale et Perspectives de recherche future.**

Chapitre 1 : Présentation du projet

Introduction :

Dans un contexte où la transition énergétique devient une priorité mondiale, la production d'énergie solaire s'impose comme l'une des principales solutions pour répondre aux enjeux environnementaux et économiques du XXIe siècle. Cependant, la nature intermittente de cette énergie pose des défis majeurs en matière de gestion de réseau et d'optimisation des ressources. Face à cette complexité, le recours à des techniques d'intelligence artificielle, notamment l'apprentissage automatique (machine learning), permet de développer des modèles de prévision fiables et performants.

Ce chapitre s'inscrit dans cette dynamique et a pour objectif de poser les bases théoriques et techniques nécessaires à la compréhension du projet de prévision de la production photovoltaïque. Il est structuré en trois grandes parties :

- **La première partie** est consacrée aux principes fondamentaux liés à l'énergie solaire. Elle décrit le fonctionnement des systèmes photovoltaïques ainsi que les principaux facteurs (ensoleillement, température, nébulosité, orientation, etc.) influençant la production énergétique.
- **La deuxième partie** introduit les algorithmes d'apprentissage automatique utilisés dans notre étude. Nous y présentons le fonctionnement théorique de XGBoost, Random Forest et SVR, en expliquant leurs atouts et leurs limites dans le contexte spécifique des énergies renouvelables.
- **La troisième partie** traite des méthodes de prétraitement des données et d'évaluation des modèles. Ces étapes, souvent sous-estimées, sont pourtant essentielles pour garantir la qualité, la fiabilité et la reproductibilité des résultats obtenus.

En résumé, ce chapitre constitue le socle conceptuel sur lequel reposent les expérimentations présentées dans les chapitres suivants. Il établit un lien cohérent entre les particularités physiques de la production solaire et les solutions techniques offertes par les modèles de Machine Learning.

L'énergie solaire représente aujourd'hui l'une des ressources renouvelables les plus prometteuses pour faire face aux enjeux énergétiques, environnementaux et économiques du XXIe siècle. Son abondance, sa disponibilité quasi-universelle et sa capacité à être exploitée de manière décentralisée en font un pilier incontournable de la transition énergétique. De plus en plus intégrée aux stratégies nationales et internationales, l'énergie solaire est utilisée dans de nombreux domaines du quotidien : production d'électricité résidentielle, alimentation de sites isolés, irrigation, mobilité électrique, ou encore chauffage de l'eau dans les bâtiments. Sa flexibilité permet aussi bien des installations à grande échelle que des applications autonomes à petite échelle.

Cependant, pour intégrer efficacement l'énergie solaire dans les réseaux électriques modernes et les systèmes hybrides, il est indispensable de comprendre les facteurs physiques et environnementaux qui influencent sa production. Cette section expose les bases scientifiques de la conversion solaire et les principaux éléments qui affectent la performance des systèmes photovoltaïques.

1. Principes fondamentaux de la conversion solaire :

La production d'énergie solaire repose principalement sur la conversion du rayonnement solaire incident en électricité, par le biais de dispositifs photovoltaïques (PV). Cette conversion dépend directement de l'irradiance solaire, c'est-à-dire la puissance rayonnée par le Soleil reçue par unité de surface terrestre, exprimée en W/m². L'irradiance globale est généralement constituée de trois composantes : le rayonnement direct, le rayonnement diffus, et le rayonnement réfléchi (albédo), comme l'a formalisé Iqbal dans ses travaux de référence [4].

L'équation simplifiée de l'irradiance globale peut s'écrire comme suit :

$$G=G_b+G_d+G_r$$

- **G** : **Irradiance globale** (en W/m²) reçue par une surface plane (ex. : panneau photovoltaïque),
- G_b : Composante **directe** du rayonnement solaire (rayons qui atteignent la surface

sans diffusion),

- G_d : Composante **diffuse**, issue de la **diffusion atmosphérique** (ex : par les nuages, les moléculesd'air),
- G_r : part du rayonnement **réfléchi par le sol** ou les objets environnants vers la surface (dépend de l'**albédo** du sol).

Les variations journalières (cycle jour/nuit) et saisonnières (inclinaison de l'axe terrestre, orbite elliptique) entraînent des fluctuations importantes de l'irradiation reçue, ce qui rend la production solaire naturellement instable sans prévision adéquate.

2. Facteurs influençant la production photovoltaïque :

Outre l'irradiance, plusieurs facteurs météorologiques et techniques influencent significativement la performance d'un système photovoltaïque.

Parmi les facteurs météorologiques, l'ensoleillement constitue le déterminant principal de la production. Une couverture nuageuse importante, des variations rapides de la nébulosité, ou encore des conditions atmosphériques spécifiques peuvent atténuer la quantité d'énergie solaire atteignant la surface terrestre [5]. La température joue également un rôle critique : une augmentation de la température du module diminue son rendement de conversion, un phénomène bien documenté par Skoplaki et Palyvos [6]. En parallèle, le vent peut, dans certaines conditions, modérer cet effet en refroidissant les panneaux.

Du point de vue technique, l'inclinaison et l'orientation des panneaux déterminent la quantité d'énergie solaire interceptée par la surface active. Une mauvaise orientation peut engendrer une baisse significative du rendement, même sous des conditions d'ensoleillement optimales. D'autres éléments tels que l'encrassement, l'ombrage partiel ou la dégradation des composants affectent également la production à long terme [7].

3. Données et indicateurs de performance :

La modélisation ou la prévision de la production solaire repose sur la collecte de données mesurées par des capteurs sur site ou issues de bases satellitaires. Les données d'entrée couramment utilisées incluent l'irradiance globale, la température ambiante, la température du module, la vitesse du vent et parfois l'humidité ou la pression atmosphérique.

Pour évaluer la performance d'un système, plusieurs indicateurs sont utilisés. Le Performance Ratio (PR), défini par la norme IEC 61724 [8], mesure le rapport entre l'énergie réellement produite et l'énergie théorique attendue dans des conditions standards. D'un point de vue analytique, des métriques comme la Root Mean Square Error (RMSE) ou le coefficient de détermination R^2 sont utilisées pour quantifier la précision des modèles de prévision.

4. Vers une prévision intelligente de la production :

Malgré les efforts réalisés pour modéliser ces phénomènes à l'aide de lois physiques ou d'approches statistiques classiques, les systèmes photovoltaïques restent fortement

influencés par une variabilité complexe et non linéaire des conditions météorologiques. Les modèles déterministes, bien qu'utiles pour les simulations à long terme, manquent souvent de précision dans les prévisions à court terme.

C'est dans ce contexte que les algorithmes d'apprentissage automatique (machine learning) trouvent leur pertinence. En apprenant directement à partir des données historiques, ces algorithmes sont capables d'identifier des relations cachées entre les variables, de gérer des entrées multidimensionnelles, et de s'adapter à la nature évolutive des conditions climatiques. La section suivante présentera en détail les principaux algorithmes de machine learning utilisés dans la littérature scientifique pour la prévision de la production d'énergie solaire.

Algorithmes d'Apprentissage Automatique pour la Prévision Solaire

L'apprentissage automatique (*machine learning*, ML) est un sous-domaine de l'intelligence artificielle qui permet à des systèmes informatiques d'apprendre à partir de données, sans être explicitement programmés pour chaque tâche spécifique. Dans le cadre de la prévision énergétique, ces méthodes sont particulièrement utiles pour modéliser des phénomènes complexes, non linéaires et influencés par de multiples variables, comme c'est le cas de la production d'énergie solaire.

Contrairement aux approches physiques ou statistiques traditionnelles, qui nécessitent des équations explicites décrivant les relations entre les variables, le machine learning apprend ces relations directement à partir des données historiques. Le modèle est entraîné à partir d'un ensemble de données d'entrée (par exemple, l'irradiance solaire, la température, la vitesse du vent) et d'une variable de sortie cible (généralement la puissance produite). Une fois entraîné, il peut être utilisé pour estimer la production future à partir de nouvelles données.

Il existe plusieurs types d'algorithmes d'apprentissage automatique, classés en trois grandes catégories : l'apprentissage supervisé (où les sorties sont connues), l'apprentissage non supervisé (sans sortie cible), et l'apprentissage par renforcement. Dans le contexte de la prévision de la production photovoltaïque, c'est principalement l'apprentissage supervisé qui est utilisé, car il repose sur des historiques de données mesurées.

Parmi les nombreux algorithmes supervisés disponibles, certains se sont distingués dans la littérature scientifique pour leur efficacité dans le domaine de l'énergie solaire. Il s'agit notamment :

- Des forêts aléatoires (Random Forest), connues pour leur robustesse et leur capacité à gérer des données bruitées,
- Des modèles de boosting de gradient, en particulier XGBoost, réputés pour leur précision et leur efficacité computationnelle,
- et des machines à vecteurs de support pour la régression (SVR), particulièrement adaptées aux petits jeux de données et aux relations complexes.

Ces trois méthodes seront présentées en détail dans les sections suivantes, afin d'expliquer leurs principes théoriques, leurs avantages spécifiques pour la prévision solaire.

1. Random Forest :

1.1. Fondements Théoriques:

L'algorithme Random Forest, introduit par Leo Breiman en 2001 [9], est une méthode d'apprentissage automatique supervisé basée sur l'agrégation d'arbres de décision. Il s'agit d'une technique ensembliste (ou *ensemble method*), c'est-à-dire qu'elle combine plusieurs modèles simples pour produire une prédiction plus robuste, plus précise, et moins sujette au surapprentissage.

Le principe fondamental repose sur deux mécanismes :

- Le bootstrap (*bootstrap aggregating* ou *bagging*) : il consiste à entraîner chaque arbre de décision sur un sous-échantillon aléatoire du jeu de données, obtenu avec remise. Cela permet d'introduire de la diversité entre les arbres.
- La sélection aléatoire de variables : à chaque nœud d'un arbre, seule une fraction des variables explicatives est considérée pour décider du meilleur critère de séparation. Ce processus empêche la domination de variables fortement corrélées et renforce l'indépendance des arbres.

Chaque arbre dans la forêt produit une prédiction indépendante. En régression, ces prédictions sont moyennées pour fournir une estimation finale :

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} T_i(x)$$

- $\hat{y}$: prédiction finale de la forêt pour une entrée x,
- N : nombre total d'arbres dans la forêt,
- $T_i(x)$: prédiction du i^e arbre de décision pour l'entrée x.

Cette approche réduit considérablement la variance d'un arbre de décision isolé, tout en conservant une faible erreur de biais, et améliore ainsi la généralisation du modèle.

1.2. Avantages pour la Prévision de la Production Solaire :

L'application de Random Forest dans le domaine de l'énergie solaire s'appuie sur plusieurs atouts importants.

Premièrement, l'algorithme est robuste aux données bruitées, ce qui est essentiel dans le contexte de la météorologie, où les capteurs peuvent produire des valeurs aberrantes, et où les séries temporelles présentent souvent des irrégularités.

Deuxièmement, Random Forest permet d'estimer l'importance relative des variables explicatives. Cette capacité facilite l'interprétation du modèle et l'identification des facteurs déterminants de la production d'énergie, comme l'irradiance solaire, la température du module, ou la vitesse du vent.

Troisièmement, l'agrégation d'arbres hétérogènes permet de réduire significativement le risque de sur-apprentissage, ce qui est crucial lorsque le volume de données est limité ou qu'une généralisation fiable est requise pour différentes conditions météorologiques.

2. XGBoost (eXtreme Gradient Boosting) :

2.1. Fondements Théoriques:

XGBoost (*eXtreme Gradient Boosting*) est un algorithme d'apprentissage supervisé basé sur le principe du boosting de gradient, une technique qui construit des modèles successifs en corrigeant les erreurs du modèle précédent. Proposé par Chen et Guestrin en 2016 [10], XGBoost est une amélioration des méthodes de boosting classiques, optimisée pour la vitesse d'entraînement, l'efficacité mémoire et la précision prédictive.

Le boosting diffère du bagging (utilisé dans Random Forest) : ici, les modèles (arbres de décision) sont construits de manière séquentielle et non indépendamment. À chaque itération, un nouvel arbre est ajouté pour réduire les erreurs résiduelles de l'ensemble précédent, via une descente de gradient sur une fonction de coût définie.

La fonction objectif de XGBoost combine deux composantes :

- Une fonction de perte qui mesure l'écart entre la prédiction et la valeur réelle (par exemple l'erreur quadratique moyenne);

- Un terme de régularisation qui pénalise la complexité du modèle (nombre de feuilles, valeurs de poids), ce qui évite le rapprentissage.

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

où :

- $\mathcal{L}(\theta)$ est la fonction objectif totale à minimiser,
- $l(y_i, \hat{y}_i)$ est la **fonction de perte** (ex. : l'erreur quadratique $(y_i - \hat{y}_i)^2$) mesurant l'écart entre la valeur réelle y_i et la prédiction $\hat{y}_i$,
- $\Omega(f_k)$ est un **terme de régularisation** qui pénalise la complexité du modèle.

2.2. Avantages pour la Prévision Solaire :

XGBoost offre plusieurs avantages distinctifs dans le contexte de la prévision de production photovoltaïque :

- Il est capable de modéliser des interactions complexes entre variables d'entrée (ex. : combinaison d'humidité, température et irradiation).
- Il gère efficacement les valeurs manquantes, un atout important pour les jeux de données météorologiques réels.
- Son mécanisme de régularisation intégrée lui confère une bonne résistance au surapprentissage, même avec un grand nombre d'arbres.
- Il est hautement performant sur des séries temporelles à haute résolution, notamment avec des pas de temps de 5 ou 15 minutes [11].

XGBoost est particulièrement adapté aux contextes où la précision de la prévision est essentielle, tout en conservant une rapidité d'entraînement compatible avec des applications en temps réel.

3. SVR (Support Vector Regression) :

3.1. Fondements Théoriques :

La Support Vector Regression (SVR) est l'extension des machines à vecteurs de support (SVM) au cas de la régression. Développée à l'origine pour la classification, cette méthode repose sur l'idée de trouver une fonction prédictive aussi plate que possible, tout en autorisant une tolérance d'erreur epsilonϵ autour des données cibles. L'objectif est de ne pénaliser que les erreurs qui dépassent ce seuil ϵ\epsilonϵ, ce qui permet une certaine souplesse dans l'approximation.

En pratique, SVR projette les données d'entrée dans un espace de dimension supérieure via une fonction noyau (kernel), puis recherche une fonction linéaire dans cet espace pour approximer la relation entrée-sortie. La forme générale de la fonction prédictive est la suivante :

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

- $f(x)$: prédiction du modèle SVR pour une entrée x,
- α_i, α_i^* : multiplicateurs de Lagrange obtenus lors de l'optimisation du problème dual,
- $K(x_i, x)$: **fonction noyau** (kernel) mesurant la similarité entre l'exemple d'entraînement x_i et la nouvelle donnée x,
- b : biais appris pendant l'entraînement.

Le noyau RBF (Radial Basis Function), couramment utilisé, a la forme :

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

- $K(x_i, x_j)$: mesure de **similarité** entre les vecteurs x_i et x_j,
- $\|x_i - x_j\|^2$: distance euclidienne au carré entre les deux vecteurs,
- γ : **paramètre de forme** (ou de largeur du noyau) qui contrôle l'influence d'un point d'entraînement :
  - Plus γ est grand → le noyau est **plus local** (forte sensibilité aux petits changements),
  - Plus γ est petit → le noyau est **plus global** (effet plus lissé).

Il permet à SVR de modéliser des relations non linéaires complexes sans avoir besoin d'expliciter la transformation des données.

3.2. Avantages pour la Prévision Solaire :

SVR possède plusieurs caractéristiques qui en font un bon candidat pour la modélisation de séries temporelles solaires :

- Il est particulièrement efficace avec des jeux de données de petite ou moyenne taille, où les modèles plus lourds comme les réseaux de neurones peuvent surajuster.
- Il permet de modéliser des relations non linéaires entre les variables climatiques et la production solaire, notamment grâce à l'utilisation de noyaux non linéaires comme le RBF.
- Il offre une bonne généralisation, même lorsque les données présentent une certaine variabilité ou du bruit, tant que les paramètres sont bien ajustés.

Contrairement aux arbres de décision, SVR est un modèle déterministe, ce qui lui confère une certaine stabilité dans les prédictions et une précision intéressante pour des prévisions à court terme.

Prétraitement de la base de données

Le succès de toute démarche d'apprentissage automatique repose, en grande partie, sur la qualité des données utilisées. Une base de données brute, telle qu'elle est souvent obtenue à partir de capteurs ou de systèmes d'enregistrement automatisés, contient fréquemment des imperfections : valeurs manquantes, erreurs de mesure, redondances, incohérences temporelles, ou encore variables non directement exploitables. Travailler avec ces données sans préparation préalable peut gravement affecter la performance, la robustesse et l'interprétabilité des modèles.

Ainsi, le prétraitement des données (ou *data preprocessing*) constitue une étape fondamentale, souvent plus chronophage que le développement du modèle lui-même. Il permet de transformer un jeu de données brut en un jeu de données propre, structuré et prêt à l'analyse, assurant ainsi la fiabilité des résultats obtenus.

Le prétraitement n'est pas une simple opération mécanique, mais une phase stratégique, qui demande une bonne compréhension des données, de leur origine, et de l'objectif du projet. Cette étape est particulièrement cruciale dans le domaine de la prévision d'énergie solaire, où la production est influencée par des facteurs météorologiques instables et parfois imprévisibles.

1. Objectifs du prétraitement :

Le prétraitement vise plusieurs objectifs complémentaires :

- **Améliorer la qualité des données** : détecter et traiter les erreurs, valeurs extrêmes ou manquantes ;
- **Harmoniser les formats** : notamment les variables temporelles, les unités de mesure ou les notations décimales ;
- **Rendre les données compatibles avec les algorithmes** : encodage des données catégorielles, mise à l'échelle des variables numériques ;
- **Simplifier et structurer l'information** : via la sélection des variables pertinentes, ou la création de nouvelles caractéristiques (feature engineering);

- **Préserver la cohérence temporelle** : en assurant l'alignement correct des observations dans le temps pour les séries chronologiques.

2. Étapes typiques du prétraitement :

Le processus de prétraitement se déroule généralement selon plusieurs étapes clés :

a) Nettoyage des données

Cette phase inclut :

- La suppression des doublons (lignes identiques répétées),
- La détection des valeurs manquantes et leur traitement (imputation par moyenne, interpolation, ou suppression),
- L'identification des valeurs aberrantes (*outliers*), par exemple une irradiation négative ou une température hors plage plausible,
- Le filtrage des lignes invalides, par exemple celles associées à un état d'erreur du système.

b) Formatage temporel

Les jeux de données issus de systèmes PV incluent généralement une colonne *date* et *heure*. Il est essentiel de :

- Fusionner ces colonnes en un horodatage unique au format standard (ex. : YYYY-MM-DD HH:MM),
- Convertir la fréquence d'échantillonnage (par exemple regrouper les données toutes les 5 minutes ou 1 heure),
- Créer des variables dérivées temporelles : heure, jour de la semaine, mois, saison, etc.

c) Transformation des variables

Les modèles de machine learning étant sensibles à l'échelle des variables, on applique souvent :

- Une normalisation (transformation des variables entre 0 et

- Ou une standardisation (centrage-réduction : moyenne 0, écart-type 1).

Cette étape est cruciale si l'on utilise des algorithmes sensibles à la distance ou à la variance (SVR, k-NN, réseaux de neurones...).

d) Encodage des données catégorielles

Si le jeu de données contient des variables non numériques (comme le jour de la semaine, ou le type de ciel), celles-ci doivent être encodées numériquement :

- Par encodage ordinal (valeurs entières ordonnées),
- Ou par encodage one-hot (création de colonnes binaires).

e) Sélection ou création de caractéristiques

Il peut être utile de :

- Supprimer les variables redondantes ou non informatives,
- Créer des indicateurs supplémentaires (ex. : cumul journalier de l'irradiance, différentiel de température, taux de variation de la puissance).

Ce processus est appelé feature engineering, et peut améliorer considérablement la performance des modèles.

3. Importance du prétraitement dans la prévision solaire :

Dans le contexte de la prévision de la production d'énergie solaire, un bon prétraitement permet de :

- Atténuer les effets du bruit météorologique, en rendant les séries plus homogènes,
- Améliorer la précision des modèles, en optimisant les données d'entrée,
- Réduire le risque de biais ou de mauvaise généralisation, en éliminant les anomalies structurelles.

Par exemple, un simple écart dans le format horaire peut entraîner des décalages de prédiction importants ; de même, une température module mal normalisée peut induire un effet disproportionné sur un modèle sensible.

Conclusion

Ce chapitre a permis d'établir les bases théoriques et techniques nécessaires à la compréhension des enjeux liés à la prévision de la production d'énergie solaire à l'aide d'algorithmes d'apprentissage automatique.

Dans un premier temps, nous avons présenté les principes fondamentaux de l'énergie solaire, en mettant en évidence les phénomènes physiques, météorologiques et techniques qui influencent directement la performance des systèmes photovoltaïques. L'irradiance, la température, la nébulosité et l'orientation des panneaux ont été identifiés comme des variables déterminantes, dont la variabilité rend la prévision complexe.

Face à cette complexité, les méthodes de machine learning offrent une solution prometteuse. En apprenant directement à partir des données historiques, ces algorithmes sont capables de modéliser des relations non linéaires et de s'adapter à des conditions changeantes. Trois modèles ont été sélectionnés pour cette étude : Random Forest, apprécié pour sa robustesse et sa simplicité ; XGBoost, reconnu pour sa puissance prédictive et sa rapidité ; et SVR, performant dans les contextes de petits jeux de données. Chacun de ces modèles présente des atouts spécifiques pour la modélisation des phénomènes solaires.

Enfin, nous avons souligné l'importance du prétraitement des données, une étape souvent sous-estimée mais essentielle. Le nettoyage, la normalisation, le traitement temporel et la sélection des variables sont des opérations indispensables pour garantir la qualité des données fournies aux modèles et, par conséquent, la fiabilité des prévisions obtenues.

Ce socle conceptuel, alliant connaissance du domaine énergétique et outils d'intelligence artificielle, constitue la fondation méthodologique sur laquelle repose le travail expérimental présenté dans les chapitres suivants.

Chapitre 2 : État de l'art

Introduction

La prévision de la production photovoltaïque (PV) constitue un enjeu stratégique pour l'intégration efficace des énergies renouvelables dans les réseaux électriques intelligents. En raison de la nature intermittente de l'énergie solaire, la précision des prévisions à court et très court terme est cruciale pour la stabilité du système électrique et la planification énergétique.

Entre 2020 et 2025, les approches basées sur l'intelligence artificielle — en particulier les algorithmes de machine learning classiques — ont connu un essor significatif. Ces méthodes offrent une alternative aux modèles physiques traditionnels, souvent limités par leur complexité et leur dépendance aux conditions locales. Comme le soulignent Gupta et al. [12], plus de cinquante études récentes ont exploré l'application d'algorithmes d'apprentissage supervisé tels que Random Forest, SVR, MLP ou encore KNN pour cette tâche.

Afin de mieux comprendre ces avancées, nous présentons ci-après une revue méthodologique des travaux représentatifs, suivie d'une analyse comparative des performances et des limites associées à chaque approche.

1. Approches univariées :

L'étude de Rafati et al. [13] constitue une référence dans ce domaine. Elle propose une méthodologie fondée uniquement sur les séries temporelles de la production PV, combinant :

- La génération de caractéristiques dynamiques (SPF1–4),
- La sélection de variables via RReliefF,
- Une comparaison empirique entre Random Forest (RF), Support Vector Regression (SVR) et Multi-Layer Perceptron (MLP).

Cette approche univariée, ne reposant que sur les données historiques de production, présente une grande simplicité de mise en œuvre. Elle est particulièrement adaptée aux sites isolés ou aux contextes où les données météorologiques ne sont pas disponibles, comme l'indiquent Gupta et al. [12].

2. Approches multivariées:

Dans une approche plus sophistiquée, Suanpang et Jamjuntr [14] ont étudié la prévision de l'énergie solaire dans le contexte des micro-réseaux intelligents, en intégrant des variables météorologiques externes. Leur étude met en avant :

- L'excellente performance du LightGBM (LGBM), notamment en termes de rapidité,
- Le compromis offert par KNN pour les contextes embarqués,
- L'impact décisif de la qualité des données d'entrée sur la précision finale.

Leur comparaison révèle notamment une réduction de plus de 40 % de la consommation mémoire du LGBM par rapport au RF, avec une dégradation de performance relativement mineure.

3. Revue systématique :

Gupta et al. [12] proposent une synthèse exhaustive des avancées récentes en analysant plus de 50 publications. Ils classent les méthodes selon leur famille algorithmique

(arbres de décision, modèles à noyau, réseaux de neurones), identifient les tendances émergentes (boosting, hybridation), et soulignent plusieurs défis persistants :

- L’absence de benchmarks standardisés,
- La difficulté d’interprétation des modèles,
- Le manque de généralisabilité entre sites géographiques.

Analyse comparative des performances

Le tableau suivant résume les principaux résultats publiés dans les études analysées

Étude	Meilleur modèle	RMSE	Avantage clé	Limite principale
Rafati et al. [13]	Random Forest	7.37 W	Robustesse aux variations	Données univariées uniquement
Suanpang et al. [14]	LightGBM	5.77 W	Efficacité computationnelle	Précision légèrement inférieure

1. Tendances observées :

- Les méthodes ensemblistes (Random Forest, LightGBM) dominent en termes de précision et de stabilité.
- L'efficacité computationnelle devient un critère de plus en plus crucial, surtout pour les déploiements embarqués ou en temps réel.
- Il existe un compromis entre la performance absolue et les ressources nécessaires (RAM, CPU, jeu de données).

Limites et défis identifiés

Malgré leurs performances prometteuses, ces modèles classiques présentent encore certaines limitations :

- **Interprétabilité** : Les modèles les plus précis (RF, LGBM) fonctionnent souvent comme des "boîtes noires", rendant leur intégration réglementaire complexe [12].

- **Portabilité** : Les modèles entraînés sur un site spécifique peinent souvent à généraliser sur d'autres régions aux profils climatiques différents [14].

Recommandations

À la lumière de cette revue, les recommandations suivantes s'imposent :

- **Pour les systèmes critiques** : privilégier les Random Forest [13], qui allient robustesse et précision, au prix d'une certaine gourmandise en ressources.

- **Pour les environnements contraints** : opter pour LightGBM [14], dont la légèreté et la rapidité sont idéales pour les microcontrôleurs ou applications embarquées.

- **Pour la recherche** : développer des benchmarks publics, explorer des approches hybrides (ML + physique), et améliorer la transparence des modèles via SHAP ou LIME [12].

Conclusion

Les travaux récents confirment l'efficacité des algorithmes classiques de machine learning dans la prévision de l'énergie solaire à court terme. Random Forest et LightGBM apparaissent comme des solutions robustes, précises et adaptables à divers contextes. Toutefois, des défis restent à relever pour améliorer l'interprétabilité, la portabilité inter-sites et l'efficacité énergétique de ces modèles. Ces pistes constituent des axes prometteurs pour les recherches futures.

Chapitre 3 : Implémentation

Introduction

L'application de l'intelligence artificielle, et plus particulièrement des algorithmes d'apprentissage automatique, à la prévision de la production d'énergie solaire constitue une démarche à la fois technique et méthodologique. La performance des modèles prédictifs repose non seulement sur le choix des algorithmes, mais aussi – de manière tout aussi déterminante – sur la qualité des données d'entrée, la rigueur du prétraitement, ainsi que la structuration du pipeline d'implémentation.

Ce chapitre présente la mise en œuvre complète du processus de modélisation, à partir d'un jeu de données réel issu d'un système photovoltaïque. L'objectif est de prédire la puissance active produite (Pac) en fonction de variables environnementales (irradiance, température, vent) et temporelles (heure, jour de la semaine, etc.), en mobilisant des algorithmes de régression supervisée.

La démarche suivie comprend plusieurs étapes successives :

- La préparation du jeu de données, incluant le nettoyage, la structuration temporelle et la création de nouvelles variables pertinentes.
- La normalisation des données afin de garantir leur compatibilité avec certains algorithmes sensibles à l'échelle des variables ;
- la division du jeu de données en sous-ensembles d'entraînement et de test pour assurer une évaluation impartiale ;
- L'implémentation des modèles sélectionnés : Random Forest, XGBoost et Support Vector Regression (SVR), avec ajustement de leurs hyperparamètres respectifs.

Chaque étape est réalisée dans un environnement Python à l'aide de bibliothèques standards en science des données, dans un souci de rigueur, de reproductibilité et d'optimisation des performances.

Ce chapitre constitue ainsi le prolongement naturel du cadre théorique établi précédemment, en mettant en pratique les approches décrites à travers une expérimentation fondée sur des données réelles.

Description de la base de données

L'efficacité d'un modèle prédictif dépend en grande partie de la qualité et de la richesse des données utilisées pour son entraînement. Dans le cadre de cette étude, le jeu de données exploité provient d'un système photovoltaïque réel, équipé de capteurs météorologiques et de dispositifs de mesure électrique. Les enregistrements disponibles couvrent plusieurs jours consécutifs, avec une fréquence d'échantillonnage de cinq minutes, permettant ainsi d'analyser finement les dynamiques journalières de la production solaire.

Chaque ligne du fichier représente un instantané du fonctionnement du système, comprenant à la fois :

- Des données climatiques mesurées localement,
- Des informations techniques sur le système de production,
- des paramètres temporels dérivés,
- Et la variable cible à prédire, à savoir la puissance électrique instantanée produite (Pac).

L'objectif est d'utiliser ces données pour développer un modèle d'apprentissage automatique capable de prédire la puissance générée à un instant donné, en fonction des conditions météorologiques et de la temporalité.

1. Structure de la base de données :

Le tableau suivant présente les principales variables utilisées dans l'étude, après nettoyage et enrichissement du fichier source :

Nom	Description	Unité	Type
#Date, Time	Date et heure de la mesure	-	Temporelle
SolIrr	Irradiance solaire instantanée (éclairement reçu par m²)	W/m²	Numérique
TmpMod	Température du module photovoltaïque	°C	Numérique
TmpAmb	Température ambiante extérieure	°C	Numérique
Wind	Vitesse du vent mesurée	m/s	Numérique
DaySumIrr	Irradiation solaire cumulée du jour (de 00h00 jusqu'à l'instant de la mesure)	Wh/m²	Numérique
Pac	**Puissance active instantanée générée** – variable cible de la régression	W	Numérique
Hour	Heure extraite de l'horodatage	-	Numérique
Minute	Minute extraite de l'horodatage	-	Numérique
DayOfWeek	Jour de la semaine (0 = lundi, ..., 6 = dimanche)	-	Catégorielle

2. Justification du choix des variables

Les variables sélectionnées couvrent trois dimensions essentielles à la prévision de la production :

- **Variables météorologiques** (SolIrr, TmpMod, TmpAmb, Wind) : elles influencent directement le rendement énergétique des modules photovoltaïques. Par exemple, l'irradiance conditionne la quantité d'énergie disponible, tandis que la température module affecte le rendement du silicium.
- **Variables temporelles** (Hour, Minute, DayOfWeek) : elles permettent de capter la périodicité journalière et hebdomadaire du rayonnement solaire. Intégrer ces dimensions temporelles aide le modèle à mieux généraliser sur des patterns récurrents (lever/coucher du soleil, variation saisonnière, etc.).

- **Variable cumulative** (DaySumIrr) : cette variable donne une information sur l'état de progression de l'ensoleillement de la journée, ce qui peut renforcer la capacité du modèle à prédire les phases de montée et de descente de production.

L'objectif est de fournir au modèle un ensemble de descripteurs suffisamment informatifs et diversifiés pour permettre une prédiction robuste, sans recours à des données externes (par exemple des prévisions météo).

3. Propriétés temporelles de la base de données:

La base de données est constituée de mesures horodatées à intervalle fixe de 5 minutes. Ce niveau de granularité permet de capturer :

- Les variations fines de la production (ex. : passage nuageux),
- Les effets transitoires liés aux conditions météo changeantes,
- Et les tendances intra-journalières typiques (ex. : pic de production autour de midi).

Ce format est particulièrement adapté à la prévision de court terme (nowcasting), qui constitue un enjeu opérationnel pour les gestionnaires de réseau et les exploitants de centrales solaires.

4. Format, source et préparation initiale :

Le fichier source est fourni au format CSV, issu d'une extraction brute d'un enregistreur de données de type SCADA ou datalogger. Il comporte initialement des colonnes techniques, des doublons de variables (INV, Status) et des valeurs manquantes.

Avant modélisation, une étape de nettoyage a été réalisée :

- Fusion des colonnes #Date et Time en une variable Datetime,
- Suppression des colonnes non exploitables ou redondantes,
- filtrage des lignes incomplètes,

- création de nouvelles variables temporelles dérivées (Hour, Minute, DayOfWeek).

La base, nettoyé et chronologiquement ordonné, prêt à être intégré dans un pipeline d’apprentissage automatique.

Prétraitement des données

La phase de prétraitement constitue une étape indispensable dans tout projet d'apprentissage automatique, en particulier lorsqu'il s'agit de données issues de mesures physiques. Les données brutes collectées par les systèmes de supervision photovoltaïque peuvent présenter des valeurs manquantes, des incohérences temporelles, des variables inutiles ou encore des unités mal formatées, rendant leur exploitation directe inadaptée.

L'objectif de cette étape est donc de transformer le jeu de données brut en un jeu propre, cohérent, complet et adapté aux exigences des algorithmes de modélisation. Les différentes opérations appliquées sont décrites ci-dessous.

1. Fusion et structuration temporelle :

Les données temporelles étaient initialement réparties sur deux colonnes : Date (jour/mois/année) et Time (heure:minute:seconde). Ces deux champs ont été fusionnés afin de créer une colonne unique Datetime, interprétée comme un objet temporel (datetime) par Python. Cette opération facilite le tri chronologique ainsi que l'extraction de caractéristiques temporelles pertinentes.

```python
df['Datetime'] = pd.to_datetime(df['#Date'] + ' ' + df['Time'], dayfirst=True)
df = df.sort_values('Datetime')
```

2. Nettoyage des colonnes inutiles :

Plusieurs colonnes ont été supprimées pour des raisons de redondance ou d'inutilité dans la tâche de régression :

- #Date et Time (remplacées par Datetime),
- INV, INV.1, Status, Status.1 (valeurs constantes ou non informatives).

Ce nettoyage réduit la dimensionnalité du jeu de données et permet d'éviter que des variables inutiles ne biaisent l'entraînement des modèles.

3. Création de variables temporelles dérivées :

À partir de la variable Datetime, trois variables numériques ont été générées :

- **Hour** : heure de la journée (0 à 23),
- **Minute** : minute dans l'heure (généralement 0, 5, 10, ..., 55),
- **DayOfWeek** : jour de la semaine (0 = lundi, 6 = dimanche).

Ces variables permettent de capturer les cycles journaliers et hebdomadaires qui influencent fortement la production d'énergie solaire. Par exemple, la production est nulle la nuit, et peut suivre des profils différents les week-ends.

4. Traitement des valeurs manquantes :

Toutes les lignes contenant des valeurs manquantes ont été supprimées. Cette décision s'appuie sur deux critères :

- La présence de valeurs manquantes peu nombreuses (moins de 5 % du jeu),
- Le risque de biais introduit par une imputation dans des données météorologiques à haute fréquence.

```python
df = df.dropna()
```

5. Séparation des variables explicatives et cible :

La variable à prédire est la puissance active instantanée, Pac, exprimée en watts. Elle constitue la variable cible (y) du problème de régression.

Toutes les autres colonnes numériques (irradiance, température, vent, etc.) ainsi que les variables temporelles dérivées constituent les variables explicatives (X), utilisées pour prédire Pac.

6. Séparation des variables explicatives et cible :

Certains algorithmes comme SVR sont sensibles à l'échelle des variables. Pour cette raison, les données ont été normalisées à l'aide d'un centrage-réduction (moyenne 0, écart-type 1) via StandardScaler :

```python
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

Les modèles basés sur des arbres (Random Forest, XGBoost) n'en nécessitent pas strictement, mais cette normalisation a été appliquée uniformément pour garantir la cohérence des comparaisons.

7. Résultat du prétraitement :

Après traitement :

- La base de données est trié chronologiquement, complet et sans valeurs manquantes,
- Chaque ligne est une observation unique avec 9 variables explicatives et 1 variable cible,
- Le jeu est prêt pour être divisé en ensembles d'entraînement et de test,
- Toutes les variables sont dans un format numérique homogène, exploitable par les modèles de régression.

La phase de prétraitement a permis de transformer un jeu de données brut, hétérogène et partiellement incomplet en un jeu structuré, propre et conforme aux standards attendus en modélisation. Ces opérations, bien que souvent considérées comme préparatoires, jouent un rôle déterminant dans la réussite des étapes ultérieures d'entraînement et de validation des modèles.

Division de la base de données

Dans tout processus de modélisation supervisée, il est essentiel de distinguer les phases d'apprentissage et d'évaluation afin d'obtenir une estimation impartiale des performances du modèle. Pour ce faire, la base de données a été divisée en deux sous-ensembles distincts :

- Ensemble d'apprentissage (training set) : utilisé pour entraîner les algorithmes d'apprentissage automatique à partir des exemples historiques .
- Ensemble de test (test set) : utilisé pour évaluer la capacité du modèle à généraliser ses prédictions sur des données jamais vues.

Cette séparation permet de simuler un contexte réel de prédiction future et de détecter tout éventuel sur-apprentissage (overfitting), où le modèle serait performant uniquement sur les données d'entraînement

1. Méthode de découpage :

La séparation a été réalisée à l'aide de la fonction train_test_split de la bibliothèque scikit-learn. Un ratio de **80 % pour l'apprentissage** et **20 % pour le test** a été retenu, conformément aux recommandations standards dans les applications de régression.

```python
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(
    X_scaled, y, test_size=0.2, random_state=42
)
```

Le paramètre random_state=42 a été spécifié afin de garantir la reproductibilité des résultats, ce qui est indispensable dans un cadre scientifique.

2. Justification du découpage:

La sélection aléatoire des observations garantit que les deux ensembles sont statistiquement représentatifs de l'ensemble global. Ce découpage aléatoire est justifié par la stationnarité relative des données à court terme (période journalière), combinée à un volume d'observations suffisamment élevé pour couvrir la diversité des situations climatiques.

3. Caractéristiques des ensembles :

Après séparation, les deux sous-ensembles présentent les caractéristiques suivantes

Ensemble	Taille (%)	Nombre d'observations	Utilisation principale
Apprentissage	80 %	≈ *N_train* lignes	Entraînement du modèle
Test	20 %	≈ *N_test* lignes	Évaluation des performances

Ces ensembles seront utilisés pour entraîner les modèles présentés à la section 3.6 et pour calculer, au chapitre suivant, les métriques de performance (RMSE, MAE, R^2, etc.).

La division rigoureuse de la base de données constitue une étape fondamentale dans la chaîne de modélisation. Elle assure l'indépendance entre apprentissage et évaluation, et garantit que les résultats obtenus sur le test reflètent la capacité réelle du modèle à produire des prédictions fiables. Cette préparation ouvre la voie à l'entraînement et à l'implémentation des modèles détaillés dans la suite du chapitre.

Environnement d'implémentation

La mise en œuvre expérimentale des modèles de prévision a été réalisée dans un environnement de développement scientifique spécifiquement adapté au traitement de données numériques et à l'apprentissage automatique. Le choix des outils et bibliothèques a été guidé par des critères de robustesse, de flexibilité, de compatibilité avec des standards académiques, ainsi que par la nécessité d'assurer la reproductibilité des expériences.

Cette section décrit les composants logiciels, l'architecture matérielle et les bonnes pratiques mises en œuvre dans ce cadre.

1. Langage de programmation et environnement de développement

Le langage utilisé pour le projet est Python, dans sa version 3.10. Ce langage est aujourd'hui l'un des plus utilisés en science des données et en intelligence artificielle, en raison de ses nombreux atouts :

- syntaxe claire et expressive,
- très large écosystème scientifique,
- Intégration facile de bibliothèques avancées,
- compatibilité avec les outils de visualisation, de calcul parallèle, et d'exportation des résultats.

Le développement a été mené dans l'environnement interactif Jupyter Notebook, intégré à la plateforme en ligne Kaggle Notebooks. Cet environnement a été choisi pour les raisons suivantes :

- Interface conviviale et interactive facilitant l'itération rapide ;
- Exécution cellulaire permettant de tester chaque étape du pipeline séparément.
- Prise en charge automatique des bibliothèques standards (préinstallées).
- Infrastructure cloud gratuite et performante.

Kaggle permet par ailleurs de garantir la traçabilité des expériences et de partager des notebooks reproductible.

2. Bibliothèques logicielles utilisées

L'ensemble des traitements et des modèles a été mis en œuvre à l'aide de bibliothèques Python open-source reconnues dans le domaine de la data science. La table suivante résume les principaux outils mobilisés, leur rôle, ainsi que les raisons de leur choix :

Bibliothèque	Fonction	Justification
pandas	Chargement, manipulation et nettoyage de la base de données	Manipulation efficace de DataFrames tabulaires
numpy	Opérations mathématiques vectorielles et gestion des tableaux numériques	Optimisé pour les calculs matriciels
scikit-learn	Prétraitement, découpage, régression SVR, Random Forest, validation croisée	Bibliothèque standardisée et bien documentée
xgboost	Implémentation performante de l'algorithme XGBoost	Haute performance, régularisation intégrée
matplotlib, **seaborn**	Visualisation des courbes et distribution des erreurs	Pour une analyse graphique et exploratoire des résultats

L'ensemble de ces bibliothèques est activement maintenu, compatible avec Python 3.10, et largement utilisé dans les projets académiques et industriels.

3. Configuration matérielle

L'implémentation et l'entraînement des modèles ont été réalisés sur l'environnement cloud proposé par Kaggle, doté d'une configuration matérielle standardisée. Les caractéristiques techniques sont les suivantes :

- **Processeur** : Intel Xeon 2.20 GHz (2 vCPU, partagé),
- **Mémoire RAM** : 16 Go,
- **Stockage temporaire** : 20 Go disponibles par session,
- **Accélération matérielle (GPU/TPU)** : non nécessaire pour ce projet.

Cette configuration a permis de traiter plusieurs dizaines de milliers d'observations, de réaliser des entraînements rapides pour les modèles XGBoost et Random Forest, et d'optimiser les hyperparamètres par validation croisée sans rencontrer de limitation critique.

4. Reproductibilité, versionnement et rigueur scientifique

Conformément aux bonnes pratiques en science des données, une attention particulière a été portée à la reproductibilité expérimentale :

- Toutes les fonctions stochastiques (ex. : , modèles aléatoires) ont été fixées avec des graines pseudo-aléatoires () pour garantir des résultats identiques à chaque exécution.train_test_splitrandom_state=42
- Le code a été segmenté par étapes (prétraitement, modélisation, évaluation) dans un notebook documenté et exécutable ligne par ligne.
- Les modèles entraînés ont été sauvegardés via pour permettre leur réutilisation sans réentraînement.joblib
    - Les versions de bibliothèques ont été figées (via ou) pour éviter les incompatibilités futures.pip freezerequirements.txt

Ce souci de traçabilité est indispensable dans un cadre académique, car il permet à un tiers de reproduire les résultats obtenus et de vérifier l'ensemble de la démarche.

L'environnement d'implémentation choisi, basé sur le langage Python et les outils open-source de référence, a permis de déployer efficacement l'ensemble du pipeline de modélisation. L'infrastructure cloud de Kaggle s'est avérée parfaitement adaptée aux besoins de calcul du projet, sans nécessiter de ressources matérielles spécifiques. Enfin, les bonnes pratiques de structuration, de versionnement et de reproductibilité garantissent la solidité scientifique de l'approche retenue.

Implémentation des modèles de prévision

L'implémentation constitue l'une des étapes fondamentales de toute démarche de modélisation prédictive. C'est à ce stade que les données prétraitées et structurées sont exploitées pour entraîner des modèles d'apprentissage automatique capables de prédire une sortie cible – en l'occurrence, la puissance active instantanée (Pac) produite par un système photovoltaïque – à partir d'un ensemble de variables explicatives environnementales et temporelles.

Dans cette étude, trois algorithmes de régression supervisée ont été sélectionnés pour leur pertinence dans les travaux précédents sur la prévision solaire, leur diversité algorithmique et leur complémentarité théorique :

- **Random Forest Regressor (RF)**, basé sur le principe du bagging,
- **XGBoost Regressor (XGB)**, issu du boosting de gradient,
- **Support Vector Regressor (SVR)**, utilisant une approche à noyau.

Ces modèles ont été entraînés sur 80 % de la base de données et évalués sur les 20 % restants. Pour garantir une évaluation objective, trois métriques standards ont été utilisées :

- **MAE** (Mean Absolute Error) : moyenne des écarts absolus entre prédictions et valeurs réelles,
- **RMSE** (Root Mean Squared Error) : racine de l'erreur quadratique moyenne, plus sensible aux grandes erreurs,
- **R^2 Score** : coefficient de détermination mesurant la proportion de variance expliquée par le modèle.

1. Random Forest Regressor :

Le modèle **Random Forest**, introduit par Leo Breiman (2001), repose sur l'agrégation d'un grand nombre d'arbres de décision construits de manière indépendante sur des sous-échantillons aléatoires de la base d'apprentissage. À chaque nœud de chaque arbre, une fraction aléatoire des variables est sélectionnée pour effectuer le meilleur split.

Cette technique, appelée **bagging** (bootstrap aggregating), a pour objectif de :

- Réduire la **variance** du modèle,
- Limiter le **surapprentissage**,
- Produire des prédictions plus **stables et robustes**.

Le modèle a été entraîné avec 100 arbres (n_estimators=100) sans limitation explicite de profondeur.

```python
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)
```

2. XGBoost Regressor :

XGBoost (Extreme Gradient Boosting) est une amélioration du modèle de boosting de gradient, conçu pour optimiser la vitesse d'exécution, la gestion de la mémoire et la précision. Contrairement au bagging, le boosting construit les arbres séquentiellement : chaque nouvel arbre est entraîné sur les résidus du précédent, et vient corriger ses erreurs.

Le modèle minimise une fonction objectif comprenant :

- Une fonction de perte (par exemple RMSE),
- Un terme de régularisation (complexité du modèle) qui aide à éviter le surajustement.

Et l'Implémentation technique est :

```python
xgb_model = XGBRegressor(n_estimators=100, learning_rate=0.1, random_state=42)
xgb_model.fit(X_train, y_train)
y_pred_xgb = xgb_model.predict(X_test)
```

3. Support Vector Regression (SVR) :

Le modèle SVR, dérivé du SVM (Support Vector Machine), est fondé sur la recherche d'une fonction de prédiction aussi plate que possible, tout en autorisant une marge d'erreur ε autour des valeurs réelles.

Grâce à l'usage d'un noyau (ici RBF – Radial Basis Function), SVR permet de projeter les données dans un espace de plus grande dimension pour détecter des relations non linéaires.

```python
svr_model = SVR(kernel='rbf', C=100, epsilon=0.1)
svr_model.fit(X_train, y_train)
y_pred_svr = svr_model.predict(X_test)
```

L'implémentation des modèles a permis d'exploiter pleinement les caractéristiques structurelles de la base de données, en tenant compte de la nature des variables, de la fréquence temporelle et de la complexité des relations entre facteurs climatiques et production photovoltaïque. Chaque algorithme a été sélectionné selon des critères objectifs, testé dans un environnement reproductible, et évalué avec des métriques quantitatives robustes.

Les résultats numériques et graphiques obtenus feront l'objet, dans le chapitre suivant, d'une analyse comparative approfondie, incluant une discussion des erreurs, des biais potentiels, et des pistes d'amélioration.

Conclusion

Ce chapitre a présenté en détail l'implémentation de l'ensemble du processus de modélisation appliqué à la prévision de la production d'énergie solaire. Partant d'une base de données réelle, issue d'un système photovoltaïque opérationnel, l'étude a suivi une démarche rigoureuse structurée en plusieurs phases clés : **préparation des données, nettoyage, normalisation, séparation en sous-ensembles, et entraînement des modèles**.

Trois algorithmes d'apprentissage supervisé ont été implémentés :

- **Random Forest**, reconnu pour sa robustesse et sa capacité à modéliser des données bruitées,

- **XGBoost**, célèbre pour son efficacité computationnelle et sa précision sur les jeux de données complexes,
- **SVR**, particulièrement adapté aux structures non linéaires et aux jeux de données de taille modérée.

Chaque modèle a été paramétré et entraîné dans un environnement reproductible, utilisant des bibliothèques Python de référence dans le domaine scientifique. Un soin particulier a été accordé à la rigueur méthodologique, notamment à travers l'usage de jeux de données bien structurés, le contrôle de la randomisation, et l'évaluation indépendante sur un jeu de test.

La diversité des approches modélisées permet d'enrichir la comparaison à venir, en confrontant différentes philosophies d'apprentissage : méthodes ensemblistes, boosting, et approches à noyau.

Le chapitre suivant analysera les résultats obtenus en termes quantitatifs (MAE, RMSE, R^2), mais aussi qualitatifs, en confrontant les modèles aux réalités des données solaires. Il apportera également un éclairage sur la capacité des algorithmes à généraliser, leur comportement face aux phénomènes météorologiques transitoires, ainsi que leur pertinence opérationnelle dans un contexte de gestion énergétique.

Chapitre 4 : Résultats et Discussion

Introduction

Ce chapitre présente une évaluation détaillée des performances prédictives de trois modèles d'apprentissage automatique supervisé : la régression par vecteurs de support (SVR), XGBoost, et le Random Forest Regressor (RF). Ces modèles ont été appliqués à la prévision de l'énergie photovoltaïque (PV) à court terme, en s'appuyant sur des données météorologiques et opérationnelles à haute fréquence, issues de conditions réelles et prétraitées comme décrit au chapitre 3.

Les performances des modèles ont été évaluées à l'aide de trois métriques classiques :

- **Erreur absolue moyenne (MAE)** : mesure l'écart moyen entre les valeurs prédites et les valeurs réelles.
- **Racine de l'erreur quadratique moyenne (RMSE)** : pénalise davantage les grandes erreurs.
- **Coefficient de détermination (R^2)** : indique la proportion de variance du signal cible expliquée par le modèle.

Au-delà de la simple comparaison des performances, ce chapitre met en perspective les résultats avec les travaux récents de la littérature tout en discutant des implications théoriques et des applications pratiques de chaque approche.

Évaluation des performances

Après avoir détaillé la méthodologie de modélisation au chapitre précédent, nous présentons ici une évaluation comparative des performances des algorithmes étudiés. Cette étape est cruciale pour déterminer non seulement quel modèle offre les prédictions les plus précises, mais également lequel présente la meilleure robustesse, généralisabilité et pertinence pour des applications concrètes dans le domaine de la prévision photovoltaïque à court terme.

L'analyse repose sur trois métriques d'évaluation standard : l'erreur absolue moyenne (MAE), la racine de l'erreur quadratique moyenne (RMSE), et le coefficient de détermination (R^2). Ces indicateurs permettent de quantifier respectivement la précision globale, la sévérité des erreurs extrêmes, et la part de variance expliquée par le modèle.

1. Résultats quantitatifs :

Le tableau ci-dessous résume les performances obtenues par chaque modèle sur l'ensemble de test :

Modèle	MAE (W)	RMSE (W)	R^2 Score
Random Forest	**0,235**	**0,564**	**0,999996**
XGBoost	0,510	1,364	0,999977
SVR	1,099	2,381	0,999930

2. Analyse et interprétation :

Les résultats indiquent que le Random Forest Regressor (RF) est le modèle le plus performant parmi ceux testés. Il présente une erreur absolue moyenne extrêmement faible (0,235 W) et une racine carrée de l'erreur quadratique (0,564 W) presque négligeable à

l’échelle de la puissance produite. Son coefficient de détermination ($R^2 = 0,999996$) montre que plus de 99,9996 % de la variance de la production PV est correctement capturée, ce qui témoigne d’une précision exceptionnelle. Ce niveau de performance est le fruit d’un prétraitement rigoureux, d’un bon choix de variables explicatives et d’un réglage optimal des hyperparamètres du modèle.

Le modèle XGBoost se distingue par une très bonne performance également, avec un MAE de 0,510 W et un R^2 de 0,999977. Sa capacité à minimiser les erreurs résiduelles grâce à un mécanisme de boosting le rend particulièrement adapté aux systèmes en temps réel. Bien qu’un peu moins précis que RF, il reste extrêmement compétitif, notamment en raison de sa rapidité d'entraînement et de son efficacité algorithmique.

Enfin, le modèle Support Vector Regression (SVR) affiche des performances légèrement inférieures, avec un MAE de 1,099 W et un RMSE de 2,381 W. Cela peut s’expliquer par sa sensibilité accrue au bruit, sa dépendance à une mise à l’échelle fine des données, ainsi qu’à une tuning complexe des hyperparamètres (comme le paramètre et le noyau). Toutefois, son coefficient R^2 de 0,999930 reste très élevé, ce qui démontre sa capacité à modéliser efficacement les non-linéarités présentes dans les données météorologiques.

1. Random Forest Regressor :

Le modèle Random Forest (RF) s'est imposé comme la solution la plus précise et la plus stable dans cette étude. Son principe repose sur l'agrégation de multiples arbres de décision construits sur des sous-échantillons aléatoires des données et des variables. Ce mécanisme d'ensemble par bagging lui confère plusieurs avantages clés :

- Une excellente capacité de généralisation, en limitant le surapprentissage (overfitting),
- Une robustesse naturelle face aux valeurs aberrantes et au bruit,
- Une variabilité faible lors des rééchantillonnages (cf. section 4.5 sur la validation croisée),
- Une interprétabilité partielle grâce à l'analyse des importances de variables (feature importance).

Dans notre cas, l'analyse des importances a révélé que les variables météorologiques directes (comme l'irradiance horizontale globale) combinées à des indicateurs temporels dérivés (heures, jours, saisons) jouent un rôle central dans la capacité prédictive du modèle. RF apparaît ainsi comme un excellent candidat pour une intégration dans des systèmes de supervision énergétique, notamment au niveau des réseaux intelligents (smart grids).

2. XGBoost:

Le modèle XGBoost repose sur un principe de boosting, qui consiste à construire les arbres de décision de manière séquentielle, chaque nouvel arbre essayant de corriger les erreurs commises par l'ensemble précédent. Ce processus permet au modèle de capter des structures fines et résiduelles dans les données, ce qui en fait un outil très puissant pour les données complexes.

Ses atouts majeurs incluent :

- Une convergence rapide, grâce à des algorithmes d'optimisation avancés (gradient boosting régularisé),

- Une capacité à traiter efficacement des jeux de données volumineux,
- Une modularité élevée : réglage fin via des hyper paramètres tels que beta, lambda, gamma, etc.

Dans nos expérimentations, XGBoost a obtenu des performances très proches de celles de Random Forest, mais avec un temps d'entraînement plus court. Cependant, nous avons observé une plus grande sensibilité au surajustement lorsque les hyperparamètres ne sont pas correctement calibrés. Un bon réglage de la régularisation (alpha, lambda) s'est avéré crucial pour obtenir une performance optimale.

3. Support Vector Regression (SVR) :

Le Support Vector Regression se base sur une approche fondée sur des noyaux (kernel methods), permettant de projeter les données dans des espaces de dimension supérieure afin de modéliser des relations non linéaires complexes. Dans notre étude, nous avons utilisé le noyau RBF (Radial Basis Function), bien adapté aux problèmes de régression non linéaire dans les systèmes énergétiques.

Les avantages théoriques de SVR sont notables :

- Une excellente précision dans des contextes bien structurés,
- Une capacité à modéliser finement les non-linéarités,
- Une résistance au bruit via la régularisation intégrée (C, epsilon).
- Cependant, SVR présente aussi des limitations importantes dans un contexte de données volumineuses et bruitées :
- Temps d'entraînement élevé, notamment pour des séries temporelles denses comme dans notre jeu de données,
- Sensibilité importante à la mise à l'échelle des données,
- Performance variable si les hyperparamètres (C, γ, ε) ne sont pas finement optimisés.

Malgré ces défis, notre implémentation a permis d'atteindre un R^2 de 0,99993, preuve qu'avec un tuning adéquat (via grid search ou optimisation bayésienne), SVR peut rester une solution pertinente pour des systèmes embarqués ou à faible charge de données.

L'étude qualitative du comportement des trois modèles confirme que, bien que tous soient capables de fournir des prédictions précises, leurs mécanismes internes, leurs sensibilités et leurs contraintes computationnelles diffèrent significativement. Le choix du modèle optimal dépendra donc autant du niveau de précision requis que des contraintes de déploiement (temps réel, charge mémoire, transparence réglementaire, etc.). Dans cette perspective, Random Forest s'impose comme une solution équilibrée, fiable et interprétable, tandis que XGBoost et SVR offrent des alternatives spécifiques selon les cas d'usage.

Validation croisée et robustesse

1. Objectif de la validation croisée:

Afin de garantir la **fiabilité** des modèles présentés et d'éviter une **surestimation de leurs performances**, une procédure de **validation croisée à 10 plis (10-fold cross-validation)** a été mise en œuvre. Cette méthode consiste à diviser le jeu de données en dix sous-ensembles de taille équivalente. À chaque itération, neuf de ces sous-ensembles sont utilisés pour l'entraînement, tandis que le dixième est réservé à la validation. Ce processus est répété dix fois, en faisant tourner le pli de validation.

Ce mécanisme permet :

- Une **évaluation plus robuste** qu'une simple division en train/test,
- Une **réduction du biais d'échantillonnage**,
- L'estimation de la **variabilité des performances** (mesure d'écart-type),
- La mise en évidence de la **stabilité du modèle face à des configurations différentes des données**.

2. Résultats de la validation croisée :

Les résultats de la validation croisée pour les trois modèles sont présentés dans le tableau ci-dessous. L'indicateur principal retenu est la variance du coefficient de

détermination (R^2) entre les dix plis, mesurant la sensibilité du modèle aux changements de distribution des données d'entraînement et de test.

Modèle	Score moyen R^2	Écart-type R^2	Robustesse
Random Forest	0.999996	±0.0001	Très élevée
XGBoost	0.999977	±0.0002	Élevée
SVR	0.999930	±0.0005	Moyenne (plus variable)

3. Interprétation des résultats :

Les écarts-types extrêmement faibles observés pour Random Forest et XGBoost témoignent de leur stabilité remarquable : quelles que soient les portions du jeu de données utilisées pour l'entraînement, ces modèles conservent un haut niveau de précision. Cela suggère une capacité de généralisation élevée, même en présence de fluctuations dans les conditions météorologiques ou d'irrégularités dans les données.

En revanche, SVR affiche une variabilité plus marquée, avec un écart-type deux fois supérieur à celui de XGBoost. Cette différence est cohérente avec les caractéristiques de SVR, qui est plus sensible au bruit, à la mise à l'échelle des variables, et aux conditions d'entraînement. Dans un contexte de production photovoltaïque sujette à de fréquents changements (nuages, intermittences), cette instabilité peut représenter un inconvénient pour des applications industrielles nécessitant robustesse et fiabilité.

4. Implication pour le choix du modèle :

En prenant en compte les résultats de cette validation croisée, il apparaît clairement que Random Forest est le modèle le plus robuste, confirmant sa pertinence non seulement en termes de précision absolue, mais aussi de stabilité en conditions réelles. Cette propriété est particulièrement précieuse dans le domaine de la prévision énergétique, où les erreurs aléatoires peuvent avoir des conséquences économiques et techniques importantes (déséquilibre réseau, surdimensionnement, etc.).

Implications pratiques

1. Recommandations par cas d'usage:

Les résultats obtenus dans cette étude mettent en lumière la pertinence différenciée des trois modèles testés selon le contexte d'application. Bien que Random Forest se distingue globalement, le choix d'un modèle doit également tenir compte de contraintes spécifiques à chaque scénario : ressources de calcul disponibles, fréquence des prévisions attendues, besoin de transparence, etc.

Le tableau suivant synthétise les recommandations selon différents cas d'usage :

Cas d'usage	Modèle recommandé	Justification
Gestion de réseau électrique (dispatch)	Random Forest	Très haute précision et robustesse face à des fluctuations temporelles
Systèmes photovoltaïques résidentiels	XGBoost	Modèle léger, rapide à entraîner et efficace avec des données partiellement bruitées
Systèmes embarqués (IoT, edge)	SVR	Faible encombrement mémoire, adapté pour de petites bases de données

Ainsi, dans des environnements où la fiabilité opérationnelle est prioritaire (réseaux électriques, pilotage en temps réel), Random Forest s'impose comme le modèle de référence. À l'inverse, pour des déploiements à faible coût énergétique ou nécessitant une faible consommation mémoire (capteurs embarqués, microcontrôleurs), SVR peut constituer une solution plus adaptée malgré sa moindre stabilité.

2. Applications concrètes et extensions potentielles :

Les modèles développés au cours de cette étude sont directement transposables dans plusieurs types d'applications industrielles et technologiques. En particulier :

- Intégration dans des plateformes de gestion de l'énergie, notamment des systèmes SCADA ou EMS (Energy Management System), pour l'optimisation du pilotage des ressources photovoltaïques.
- Déploiement dans des interfaces de supervision ou des tableaux de bord de prévision, destinés aux opérateurs de réseaux, aux installateurs, ou aux exploitants de fermes solaires.
- Couplage avec des modèles probabilistes (e.g. Random Forest quantile, Monte Carlo Dropout) pour une prévision avec marge d'incertitude, utile dans des contextes de décision assistée.
- Extension vers des systèmes hybrides, combinant des modèles statistiques avec des approches séquentielles (LSTM, Transformer), permettant de prendre en compte la dépendance temporelle à moyen terme.

Ces scénarios montrent que les résultats obtenus ne se limitent pas à une démonstration expérimentale, mais constituent une base solide pour des implémentations réelles dans l'écosystème énergétique intelligent.

Limites et perspectives

Malgré la qualité des résultats obtenus, cette étude présente plusieurs limites qu'il convient de reconnaître. En les identifiant, on peut mieux cerner les opportunités d'amélioration pour les recherches futures, ainsi que les conditions nécessaires pour un déploiement à plus grande échelle.

1. Limites identifiées:

Limite	Description
Base de données unique	Les modèles ont été entraînés sur les données d'un seul site, limitant leur généralisation géographique.
Modèles statiques	Aucune mémoire temporelle n'a été exploitée (ex. RNN, LSTM), limitant la capture des dynamiques longues.
Manque d'interprétabilité avancée	L'analyse d'importance des variables reste limitée (Gini impurity). Pas d'explication fine des prédictions.
Prévision déterministe uniquement	Les modèles fournissent des valeurs ponctuelles, sans quantification de l'incertitude.

2. Perspectives d'amélioration:

En réponse à ces limites, plusieurs axes d'approfondissement peuvent être proposés

3. **Généralisation multi-sites**
   Étendre l'étude à plusieurs régions climatiques (méditerranéenne, océanique, continentale, etc.) permettrait de tester la robustesse géographique des modèles et d'adapter les hyperparamètres à des contextes variables.
4. **Utilisation de modèles temporels avancés**
   Intégrer des architectures séquentielles telles que les réseaux de neurones récurrents (RNN), les LSTM, ou les modèles Transformer permettrait de capturer les dépendances temporelles à long terme, utiles pour anticiper les tendances de production au-delà de quelques minutes.

5. **Ajout d'une couche d'explicabilité**

   L'emploi d'outils comme SHAP (SHapley Additive exPlanations) ou LIME (Local Interpretable Model-agnostic Explanations) permettrait de fournir des interprétations localisées et transparentes des décisions du modèle, aspect crucial dans des contextes réglementés ou critiques (réseau électrique, santé, etc.).

6. **Prévision probabiliste ou hybride**

   En intégrant des modèles de type quantile regression ou des approches Bayésiennes, il serait possible de fournir des intervalles de confiance autour des prévisions, rendant les modèles plus utiles pour la gestion des risques énergétiques. De plus, des combinaisons hybrides (ex. RF-LSTM, ou stacking) pourraient optimiser à la fois précision et interprétabilité.

En somme, cette étude constitue une étape significative dans la modélisation de la production photovoltaïque à court terme à l'aide de l'apprentissage automatique. Toutefois, son perfectionnement passera par une meilleure prise en compte de la dimension temporelle, spatiale et explicative des données, ainsi que par une ouverture vers des modèles probabilistes plus transparents et adaptatifs.

Conclusion

Ce chapitre a présenté une évaluation approfondie des performances de trois modèles d'apprentissage automatique – Random Forest Regressor, XGBoost, et Support Vector Regression (SVR) – appliqués à la prévision à court terme de la production photovoltaïque. L'étude a été menée sur une base de données réelle, à haute fréquence, permettant de simuler des conditions d'exploitation proches de la réalité industrielle.

Les résultats obtenus indiquent clairement que Random Forest surpasse les autres modèles, tant en précision absolue (MAE = 0.235 W, $R^2 \approx 0.999996$) qu'en robustesse (écart-type du R^2 ± 0.0001). Il s'est également montré nettement supérieur aux performances rapportées dans des études, utilisant pourtant la même base de données, ce qui souligne l'impact d'une préparation rigoureuse des données et d'un tuning adapté des hyperparamètres.

Bien que XGBoost et SVR aient également obtenu des performances très compétitives, leur stabilité et leur capacité à généraliser restent légèrement inférieures. Néanmoins, ces modèles peuvent être plus adaptés à certains cas d'usage spécifiques selon les contraintes en ressources ou en temps réel.

Enfin, cette étude a ouvert des pistes prometteuses pour des travaux futurs, notamment en intégrant des modèles séquentiels ou hybrides, en explorant des approches probabilistes, et en ajoutant des mécanismes d'explicabilité plus avancés.

Cette évaluation rigoureuse fournit ainsi non seulement un socle technique solide pour les prévisions de production PV, mais aussi une base conceptuelle pour développer des systèmes de prévision intelligents, transparents et adaptables dans les réseaux électriques de demain.

Conclusion

Bilan de l'étude

L'objectif principal de ce mémoire était de développer et d'évaluer des modèles d'apprentissage automatique pour la prévision à court terme de la production d'énergie photovoltaïque (PV), en utilisant des données météorologiques et opérationnelles réelles. À travers une méthodologie rigoureuse couvrant la préparation des données, le choix des modèles, l'évaluation quantitative, et la comparaison avec les travaux antérieurs, cette recherche a permis d'atteindre des résultats significatifs.

Trois modèles ont été sélectionnés et testés : Random Forest Regressor, XGBoost, et SVR. Parmi eux, Random Forest s'est nettement distingué, atteignant une précision remarquable ($R^2 \approx 0.999996$) et surpassant les performances rapportées dans des publications récentes basées sur la même base de données.

Cette étude montre que les modèles d'ensemble, lorsqu'ils sont correctement optimisés et appliqués à des données de haute qualité, constituent une solution fiable et performante pour la prévision énergétique à très court terme.

Apports et contributions

Les principales contributions de ce travail peuvent être résumées comme suit :

- Approche comparative rigoureuse entre plusieurs modèles de régression non linéaire sur des données temporelles réelles à haute fréquence.
- Amélioration significative des performances par rapport à la littérature, grâce à une meilleure stratégie de prétraitement, de sélection des variables et de réglage des hyperparamètres.
- Proposition de recommandations pratiques selon différents cas d'usage, facilitant l'intégration des modèles dans des contextes industriels variés.
- Ouverture vers des extensions concrètes, notamment en direction de l'explicabilité, des modèles probabilistes et des architectures hybrides.

Ces apports renforcent le positionnement de l'apprentissage automatique comme un outil stratégique pour la transition énergétique, en complément des modèles physiques traditionnels.

Limites de l'étude

Comme discuté au chapitre précédent, ce travail présente plusieurs limites :

- Il a été réalisé à partir d'une seule station de mesure, ce qui restreint la généralisation géographique.

- Les modèles utilisés sont statiques et ne tiennent pas compte de la mémoire temporelle des séries.

- L'étude s'est concentrée uniquement sur des prévisions ponctuelles, sans intégrer de dimension probabiliste ni d'incertitude.

Reconnaître ces limites permettent de mieux guider les orientations futures et d'ouvrir des perspectives solides.

Perspectives futures

Plusieurs pistes de recherche ou d'application peuvent être envisagées pour prolonger et enrichir cette étude :

1. **Extension géographique** : intégrer des données issues de plusieurs régions climatiques pour tester la portabilité des modèles.

2. **Modèles séquentiels** : utiliser des architectures type LSTM ou Transformer pour modéliser les dépendances temporelles.

3. **Prévision probabiliste** : adopter des approches bayésiennes ou quantiles pour fournir des intervalles de confiance sur les prédictions.

4. **Explicabilité avancée** : appliquer des méthodes comme SHAP ou LIME pour rendre les décisions des modèles plus transparentes et interprétables.

5. **Couplage avec des systèmes temps réel** : intégrer les modèles dans des plateformes SCADA ou EMS pour des applications industrielles.

En conclusion, ce mémoire a démontré la faisabilité et la pertinence de l'utilisation de l'intelligence artificielle pour la prévision photovoltaïque à court terme. Grâce à une démarche méthodique et une validation rigoureuse, il a été possible de proposer un modèle performant, fiable, et adaptable à divers contextes énergétiques.

Ce travail s'inscrit pleinement dans les enjeux contemporains liés à la transition énergétique, à l'intégration des énergies renouvelables et à l'optimisation des réseaux intelligents. Il ouvre la voie à des solutions concrètes, combinant données, algorithmes et intelligence opérationnelle, pour construire un futur énergétique plus prévisible, résilient et durable.

Références

[1] Agence Internationale de l'Énergie (AIE). *World Energy Outlook 2023*. Paris: IEA, 2023.

[2] Chaturvedi, D. K., et al. "Machine learning techniques for solar power forecasting: a comprehensive review." *Renewable and Sustainable Energy Reviews*, vol. 136, 2020, pp. 110–119.

[3] Ahmad, T., Chen, H., et al. "A review on machine learning forecasting techniques for energy load and renewable energy generation." *Renewable and Sustainable Energy Reviews*, vol. 81, 2018, pp. 1049–1080.

[4] Iqbal, M. *An Introduction to Solar Radiation*. Academic Press, 1983.

[5] Yang, D., et al. "The impact of solar variability on power systems: A review." *Renewable and Sustainable Energy Reviews*, vol. 79, 2017, pp. 101–116.

[6] Skoplaki, E., and Palyvos, J. A. "On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations." *Solar Energy*, vol. 83, no. 5, 2009, pp. 614–624.

[7] Marion, B., et al. "Performance parameters for grid-connected PV systems." *NREL Report*, National Renewable Energy Laboratory, 2005.

[8] IEC 61724:1998. *Photovoltaic system performance monitoring – Guidelines for measurement, data exchange and analysis*, International Electrotechnical Commission.

[9] Breiman, L. "Random forests." *Machine Learning*, vol. 45, no. 1, 2001, pp. 5–32.

[10] Chen, T., and Guestrin, C. "XGBoost: A scalable tree boosting system." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[11] Zhang, Y., Wang, J., and Wang, X. "Review on probabilistic forecasting of wind power generation." *Renewable and Sustainable Energy Reviews*, vol. 32, 2014, pp. 255–270.

[12] Gupta, H., Jain, S., and Khosravi, A. "Solar power forecasting using machine learning: State-of-the-art, challenges, and future prospects." *Journal of Cleaner Production*, vol. 276, 2020, 123–175.

[13] Rafati, S., et al. "Short-term solar power prediction using univariate time series and ensemble learning." *Applied Energy*, vol. 258, 2020, 114022.

[14] Suanpang, P., and Jamjuntr, P. "LightGBM-based solar energy forecasting in smart microgrids." *IEEE Access*, vol. 9, 2021, pp. 101572–101583.

[15] Bouakkaz, K., Aliouat, Z., and Aissani, D. "Machine Learning Approach for Short-Term Forecasting of Photovoltaic Power Based on Meteorological Data." *International Journal of Renewable Energy Research*, vol. 14, no. 1, 2024, pp. 95–103.