



Rapport de stage

Validation des mesures d'irradiance solaire par Deep Learning



Jean-Pierre MANSOUR – Auteur

Gireg BACHELOT – Encadrant

ENGIE GREEN, MONTPELLIER – Lieu du stage

UNIVERSITÉ DE TOULOUSE – Établissement délivrant le Master

Résumé

Afin de mesurer la quantité d'irradiance solaire reçue par les panneaux photovoltaïques, ENGIE Green installe des capteurs d'irradiance sur chacun de ses parcs solaires. Parmi la vaste famille de capteurs disponibles, ceux qui se sont avérés les plus adaptés dans ce contexte sont les pyranomètres.

Actuellement, il n'existe pas de méthode déterministe permettant de caractériser la validité des mesures d'irradiance, c'est-à-dire de déterminer à chaque instant si une mesure est valide ou non. En effet, cette validité dépend elle-même d'une multitude de facteurs aléatoires : conditions météorologiques, perturbations logistiques ou problèmes techniques liés au fonctionnement d'un parc solaire.

Par conséquent, l'intervention physique d'un technicien devient nécessaire à chaque fois qu'une mesure est suspecte pour valider ou non cette mesure. Avec l'augmentation du nombre de parcs photovoltaïques, cette validation humaine manuelle devient rapidement fastidieuse.

Pour cela, l'équipe Data et Modélisation d'ENGIE Green décide d'implémenter une approche statistique basée sur le machine learning. Cette approche consiste à récupérer la base de données d'irradiance accompagnée des validations existantes provenant de l'ensemble des parcs solaires, puis d'entraîner un modèle d'intelligence artificielle sur ce jeu de données, afin de pouvoir, à terme, valider automatiquement les mesures à venir.

A priori, certains obstacles rendent difficile l'apprentissage du modèle sur la base de données. En effet, les campagnes sont très différentes en termes d'instrumentation (nombre et nature des capteurs installés), de positionnement géographique, de topographie, ainsi que d'inclinaison et d'orientation des capteurs.

Ainsi, le travail mené dans ce stage consiste précisément à tenir compte de ces difficultés et à permettre au modèle de les intégrer efficacement, en lui fournissant des variables explicatives pertinentes qui apportent le maximum d'information possible.

Remerciements

Je tiens à remercier très chaleureusement mon encadrant de stage, Gireg Bachelot, en particulier pour son engagement constant tout au long du stage, sa disponibilité pour discuter de sujets variés, liés ou non au stage. J’apprécie également sa bienveillance et l’attention particulière qu’il a portée à mon intégration dans les équipes du Pôle Ressource et Productible. Il n’a jamais hésité à prendre du temps pour m’aider, quel que soit le sujet ou la charge de travail, et n’a jamais refusé d’échanger, même lorsque cela demandait beaucoup de lui.

Je tiens également à remercier, au sein de l’équipe Data et Modélisation, Paul Mazoyer pour le suivi constant de nos travaux et de mon avenir professionnel, Émilien Duverger pour sa disponibilité à répondre à mes questions et pour ses idées qui se sont révélées très utiles, ainsi qu’Ismaël Hennou pour son accompagnement au début du stage.

Par ailleurs, au sein de l’équipe Mesures, je dois une grande partie de ma culture et de mes connaissances sur le fonctionnement des parcs solaires à Alexandre Le Lay. Je remercie également Aurélien Corvazier de m’avoir fourni toute la base de données nécessaire et de m’avoir initié au langage SQL.

Enfin, durant ce stage, j’ai créé des souvenirs marquants et rencontré des collègues que je continuerai à valoriser. Parmi les moments inoubliables : la chasse au trésor à Strasbourg et la sortie en char à voile à Caen, qui ont été de belles occasions de s’amuser et de mieux se connaître.

Je remercie enfin ENGIE Green, ainsi que tout le Pôle Ressource et Productible, à travers la personne de Benoît Buffard, un responsable très bienveillant et proche de chacun.

Table des matières

1 Présentation de ENGIE Green	8
1.1 Implantations et capacités installées	8
1.2 Organisation de mon travail	9
2 État des lieux : campagnes solaires	10
2.1 Campagnes centralisées et décentralisées	10
2.2 Instruments installés	10
2.2.1 Mesures d'irradiance	11
2.2.2 Mesures météo	12
3 Rayonnement solaire	12
3.1 Composantes du rayonnement solaire	13
4 Présentation DataWin	15
4.1 Structure de la base de données sur DataWin	15
4.1.1 Des structures de données hétérogènes entre campagnes	16
4.1.2 Variabilité des instruments GTI au sein d'une même campagne	17
4.1.3 Manque de cohérence entre campagnes dans les capteurs disponibles	17
4.2 Marqueurs d'invalidité	18
4.3 Base de données des états de validité des mesures	19
4.3.1 Gestion des mesures non marquées	19
4.3.2 Restriction des périodes d'apprentissage aux mesures explicitement validées	20
4.3.3 Comprendre les invalidations en termes statistiques	20
5 Création de features avancées pour améliorer l'apprentissage	22
5.1 Azimut et élévation solaire	23
5.1.1 Présentation de la bibliothèque <code>pvlib</code>	24
5.1.2 Calculer l'azimut et l'élévation solaire sous <code>pvlib</code>	25
5.2 Utilisation de <code>pvlib</code> pour estimer le rayonnement solaire théorique	26
5.2.1 Indices beau-temps	27
5.3 Écarts absolus entre les pyranomètres	27
5.4 Données satellitaires (SolEye)	28
5.4.1 Fraction diffuse	28
5.5 Cumuls d'irradiance	29

5.5.1	Cumul journalier	29
5.5.2	Cumul décalé	29
5.6	Uniformisation des mesures GTI	30
5.6.1	Algorithme de transposition inverse	30
6	Méthodes et métriques	32
6.1	Algorithme Random Forest	33
6.1.1	Arbre de décision	33
6.1.2	Random Forest	34
6.1.3	XGBoost	35
6.2	Entraînement par validation croisée	36
6.3	Scores considérés pour évaluer le modèle	36
6.3.1	Précision, rappel et F-score	37
7	Bilan des résultats	38
7.1	Première approche non retenue	38
7.1.1	Empilement des colonnes d'irradiance et de météo	38
7.2	Approche retenue	39
7.3	Deuxième approche non retenue	40
7.4	Cas d'application réel	41
7.4.1	Analyse par journée - Campagne 619	42
7.4.2	Analyse par journée - Campagne 480	44
7.4.3	Analyse des prédictions du modèle par colonne GHI/GTI	45
7.5	Test du modèle sur des mois plus anciens	48
7.5.1	Mois de novembre 2021	48
7.5.2	Mois de février 2022	51
8	Conclusion	53

Table des figures

1	Répartition des sites de production renouvelable du groupe ENGIE Green en France (données de fin 2017)	8
2	Nombre de campagnes par combinaison (GHI , GTI)	12
3	Profils des quatre composantes de l'irradiance solaire par temps clair	14
4	Invalidation automatique d'une mesure GHI	18
5	Invalidation utilisateur suite à un marqueur automatique	19
6	Distribution des invalidations sur l'année	21
7	Distribution des taux d'invalidations en fonction des mesures GHI	21
8	Distribution des taux d'invalidations en fonction des mesures GTI	22
9	Angles d'azimut et élévation solaire	23
10	Distribution des invalidations par azimut et élévation	24
11	Logo pvlib	25
12	Visualisation géographique des sites	25
13	Tracé de GHI clearsky	26
14	Écarts calculés entre trois mesures GTI	28
15	Cumul décalé 1 jour, 2 jours, 3 jours.	30
16	Diagramme illustrant l'algorithme de transposition inverse	31
17	Estimation des mesures GHI à partir des mesures GTI en utilisant la transposition inverse	31
18	Disponibilité brute et disponibilité valide par campagne	32
19	Exemple de validation croisée sur 5 folds	36
20	Matrice de confusion	37
21	Exemple d'empilement des colonnes d'irradiance et de météo	39
22	Scores de chaque test	40
23	Ratios VN et FN sur le nombre de zéros par campagne	41
24	Ratios VN et FN sur le nombre de zéros sur toutes les campagnes	42
25	Invalidations de 11h40 à 13h10 bien prédites par le modèle.	43
26	Invalidations de 8h00 à 16h00 mal prédites par le modèle.	43
27	Nombre d'invalidations par type de mesure	48
28	Précision et rappel par colonne (classe 0)	50
29	Invalidations effectuées sur plusieurs jours consécutifs.	50
30	Précision et rappel par colonne (classe 0)	52

Liste des tableaux

1	Extrait du jeu de données	16
2	Exemple de base de données des états de mesures	19
3	Exemples d'élévations et d'azimuts solaires	26
4	Matrice de confusion par colonne	46
5	Précisions par colonne pour les classes 0 et 1	46
6	Rappels par colonne pour les classes 0 et 1	46
7	Nombre d'invalidations et de validations par colonne	47
8	Précisions par colonne pour les classes 0 et 1	49
9	Rappels par colonne pour les classes 0 et 1	49
10	Précisions par colonne pour les classes 0 et 1	51
11	Rappels par colonne pour les classes 0 et 1	51

Liste des notations

GHI : Global Horizontal Irradiance (irradiance horizontale globale), en W/m²

GTI : Global Tilted Irradiance (irradiance sur plan incliné), en W/m²

DHI : Diffuse Horizontal Irradiance (rayonnement horizontale diffus), en W/m²

DNI : Direct Normal Irradiance (rayonnement normale directe), en W/m²

RHI : Rear Horizontal Irradiance (rayonnement réfléchi mesuré à l'horizontale à l'arrière du capteur), en W/m².

RTI : Rear Tilted Irradiance (rayonnement réfléchi mesuré à un plan incliné à l'arrière du capteur), en W/m².

Albédo Mesure de la réflectivité d'une surface à l'irradiance, exprimée ratio.

GHI_clearsky : Valeur théorique de l'Irradiance Globale Horizontale, en W/m²

GTI_to_GHI : Valeur de GTI convertie en GHI à l'aide du modèle de transposition inverse

Mesures SolEye : Modèle numérique de Engie Green permettant d'évaluer l'irradiance des sites de production d'après données satellites.

XGBoost : eXtreme Gradient Boosting algorithm

1 Présentation de ENGIE Green

Engie Green est une entreprise française spécialisée dans les énergies renouvelables, l'éolien et le solaire, et fait partie du groupe ENGIE. Présente sur tout le territoire avec plusieurs agences régionales, elle développe, construit et exploite des parcs de production d'électricité verte, en prenant en charge toutes les étapes d'un projet : développement, construction, exploitation et maintenance.

L'entreprise dispose aujourd'hui de plus de 2,5 GW d'éolien et 1,9 GW de solaire, permettant d'alimenter plusieurs millions de foyers. Elle mène également des projets de modernisation des parcs existants afin d'améliorer leur performance en réutilisant les infrastructures déjà en place.

Au sein de l'entreprise, le pôle *Ressources et Productible* analyse la ressource et prédit la production future en modélisant le comportement des parcs. C'est dans ce pôle que s'est déroulé mon stage.

1.1 Implantations et capacités installées



FIGURE 1 – Répartition des sites de production renouvelable du groupe ENGIE Green en France (données de fin 2017).

La figure 1 illustre la répartition des sites de production renouvelable d'**ENGIE Green** en France. On observe :

- **Éolien terrestre** : 1 333 MW installés et exploités, répartis sur 91 parcs comprenant 701 éoliennes, avec une forte concentration dans le nord et l'est du pays.
- **Solaire photovoltaïque** : 862 MW installés et exploités, au travers de 101 centrales, principalement situées dans le sud et le sud-ouest.
- **Éolien en développement** : plus de 3 000 MW supplémentaires sont prévus, dont le projet pilote d'éolienne flottante en Méditerranée (mise en service prévue vers 2020).
- **Production totale** : une capacité équivalente à la consommation de près de 1,7 million d'habitants en électricité verte.

1.2 Organisation de mon travail

Mon stage s'est déroulé à Montpellier, au sein de l'équipe Data Science et Modélisation du pôle Ressource et Productible, qui s'occupe principalement des projets éoliens et solaires. Mon travail portait spécifiquement sur la partie solaire.

Il ne s'est pas limité à l'équipe Data Science et Modélisation : j'ai également collaboré régulièrement avec l'équipe Mesures, car le sujet du stage était directement lié à leurs activités. Concrètement, nous avons utilisé leur base de données de mesures afin de tester l'approche de Machine Learning développée dans ce travail de stage.

2 État des lieux : campagnes solaires

2.1 Campagnes centralisées et décentralisées

Sur l'ensemble du territoire français, ENGIE Green compte 110 campagnes solaires centralisées et 111 campagnes décentralisées. Au total, nous disposons de 6104 instruments répartis sur ces deux types de campagnes. Parmi ces instruments, les plus importants sont les capteurs météorologiques.

Ces deux types de campagnes se distinguent fondamentalement par la manière dont les instruments sont installés spatialement :

- **Campagnes centralisées** : tous les instruments sont regroupés au même endroit.
- **Campagnes décentralisées** : les instruments sont répartis à différents emplacements.

D'un point de vue technique, les campagnes centralisées sont d'expérience plus faciles à gérer, car il est plus simple de vérifier si elles ne fonctionnent pas : il suffit de comparer les instruments entre eux. Si un instrument présente un comportement anormal par rapport aux autres, cela peut alerter sur un problème à investiguer.

De l'autre côté, il faut être bien plus vigilant avec les campagnes décentralisées, dans le sens où un écart de mesure entre deux instruments ne signifie pas nécessairement qu'un des deux est erroné, car ils ne sont pas installés au même endroit et ne mesurent donc pas exactement la même chose. Un exemple simple : si, sur une période donnée, on observe que les mesures d'un instrument s'écartent de celles des autres instruments similaires, il est possible que celui-ci soit simplement à l'ombre alors que les autres ne le sont pas.

Un autre exemple : pour des instruments au sein d'une même campagne placés selon une orientation et une inclinaison identiques, il se peut que, dans une campagne décentralisée, un de ces instruments ait bougé alors qu'un autre, situé plus loin, n'ait pas bougé. Cela va modifier de manière importante la mesure d'irradiance observée pour celui qui a changé de position. Dans ce cas, l'écart détecté entre les deux instruments déclenchera bien une alerte ; toutefois, la correction à apporter ne sera pas une recalibration de l'appareil de mesure, mais une vérification de son installation. Quoi qu'il en soit, il s'agit bien d'un problème à traiter, même si l'instrument en lui-même fonctionne normalement.

Cependant, si un instrument a été décalé, cela reste un problème sérieux : ce n'est pas une panne de l'instrument, mais cela fausse quand même la qualité des mesures.

2.2 Instruments installés

Sur l'ensemble des campagnes solaires, deux types de mesures sont installés : des capteurs pour mesurer l'irradiance solaire, et des capteurs météo pour mesurer la température, la pression, l'humidité et la pluie.

2.2.1 Mesures d'irradiance

Afin de mesurer le rayonnement solaire global, des pyranomètres, qui sont des capteurs d'irradiance, sont installés. Un même capteur peut mesurer différents types d'irradiance selon la manière dont il est monté. Nous distinguons notamment :

- **GHI** : irradiance globale mesurée sur un plan horizontal.
- **GTI** : irradiance globale mesurée sur un plan incliné.

Selon le modèle et l'installation, un pyranomètre peut également mesurer les rayonnements solaires diffus et réfléchis :

- **DHI** : irradiance diffuse mesurée sur un plan horizontal.
- **RHI** : irradiance réfléchie mesurée sur un plan horizontal.
- **RTI** : irradiance réfléchie mesurée sur un plan incliné.

Ces pyranomètres sont placés dans le plan des panneaux photovoltaïques afin de recevoir la même quantité d'irradiance que ces derniers. L'irradiance solaire est mesurée en $W \cdot h/m^2$. Ils sont caractérisés par une surface parfaitement sphérique, ce qui permet de limiter les réflexions des rayons solaires sur la surface. De plus, la géométrie de la zone qui entoure cette surface sphérique est conçue de manière à empêcher les rayons réfléchis de revenir vers la surface du capteur, car cela fausserait les mesures que l'on souhaite obtenir.

Egalement, pour mesurer le rayonnement solaire direct des pyrhéliomètres sont utilisés. Ces outils mesurent :

- **DNI** : irradiance normale directe

Il est important de noter que le choix du type de mesure dépend aussi du type de panneaux installés. Les panneaux photovoltaïques classiques, tels que ceux majoritairement présents dans la flotte d'Engie, sont inclinés et donc plus sensibles à l'irradiance globale sur plan incliné (**GTI**). À l'inverse, des panneaux à concentration, plus rares, sont beaucoup plus sensibles au rayonnement direct, ce qui rend alors la mesure du **DNI** particulièrement pertinente.

Dans le cadre de ce stage, nous nous intéressons uniquement aux mesures **GHI** et **GTI**, en particulier à la prédiction de leur validité à partir d'une base de données d'irradiance complète déjà disponible. Les autres types de mesures sont quant à elles rarement disponibles sur l'ensemble des campagnes.

Enfin, terminons cette section par une statistique sur la distribution du nombre de campagnes par combinaison (**GHI**, **GTI**) :

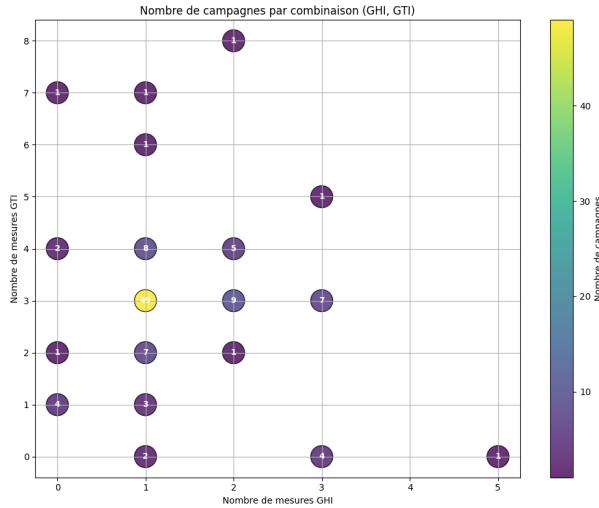


FIGURE 2 – Nombre de campagnes par combinaison (**GHI,GTI**)

Nous constatons d’après ce graphe que la très grande majorité des campagnes dispose de 1 **GHI** et 3 **GTI** mais il y a également une diversité importante de configurations qui sont parfois uniques.

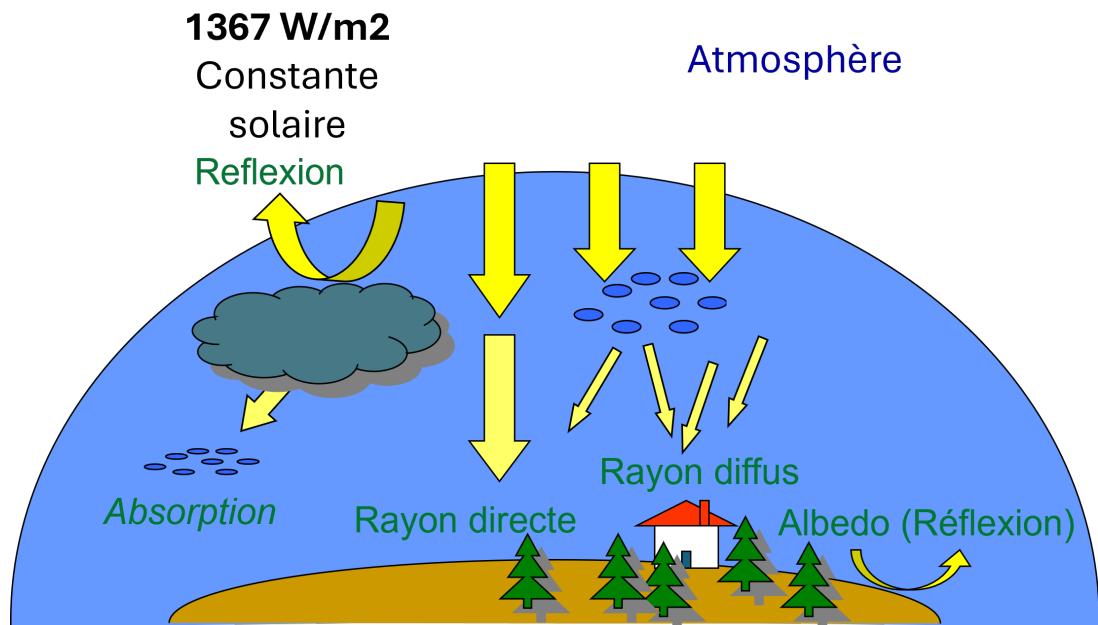
2.2.2 Mesures météo

Les mesures météorologiques sont des éléments clés et peuvent servir à formuler des hypothèses décisives pour évaluer la validité d'une mesure d'irradiance. Sur l'ensemble des campagnes solaires, nous disposons d'un grand nombre de mesures météo. Parmi elles, celles que nous jugeons les plus importantes et contenant le plus d'informations pertinentes dans notre contexte sont :

- Température ambiante, température du module PV
- Intensité de pluie, cumul instantané de pluie (calculé toutes les 10 minutes)
- Humidité relative
- Pression atmosphérique

3 Rayonnement solaire

Avant d'entrer dans les détails des travaux réalisés pendant ce stage, il est utile de rappeler comment le rayonnement solaire arrive jusqu'à nous et interagit avec l'atmosphère et la surface terrestre. Le schéma ci-dessous présente de manière simple et claire les différents phénomènes qui entrent en jeu : la réflexion, l'absorption et la diffusion. Même si comprendre ces phénomènes n'est pas strictement indispensable, cela m'a été utile pour mieux me situer dans le cadre des travaux réalisés par la suite au cours de ce stage, en particulier pour comprendre ce que le modèle doit apprendre à « deviner » à partir des données disponibles.



La constante solaire, c'est-à-dire la puissance du rayonnement solaire à l'entrée de l'atmosphère terrestre, est d'environ 1367 W/m^2 . Lorsque cette énergie traverse l'atmosphère :

- Une partie est directement **réfléchie** vers l'espace par les nuages et les particules présentes dans l'air
- Une autre partie est **absorbée** par l'atmosphère elle-même, ce qui diminue la quantité d'énergie arrivant au sol
- Ce qui atteint directement la Terre, sans être dispersé, s'appelle le **rayonnement direct**
- Ce qui est dispersé dans toutes les directions par l'air et les particules s'appelle le **rayonnement diffus**

À la surface terrestre :

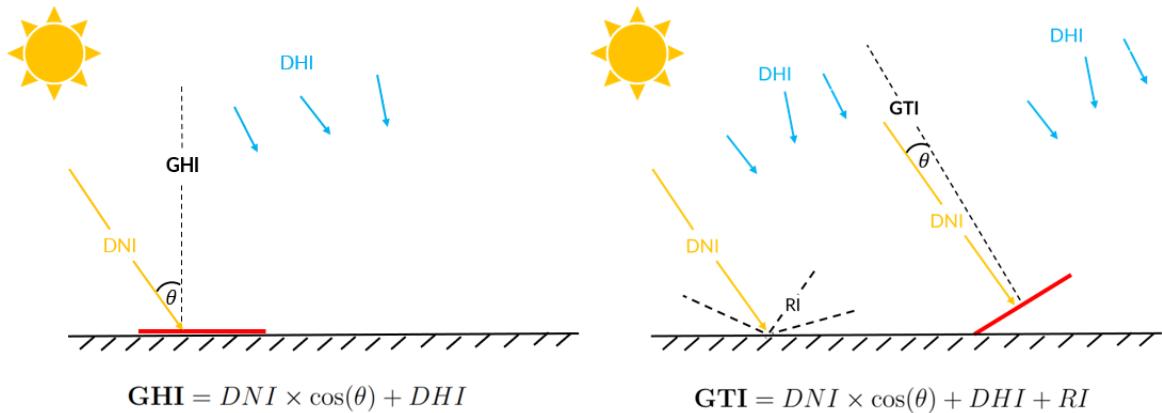
- Une partie du rayonnement est absorbée par le sol
- L'autre partie est réfléchie vers l'espace, selon la nature de la surface, plus précisément son *albédo*, une quantité qui mesure la réflectivité d'une surface

3.1 Composantes du rayonnement solaire

Cette section est dédiée à la compréhension des composantes du rayonnement solaire que nous mesurons. Il s'agit de mesures d'irradiance essentielles, que nous enregistrons et stockons dans notre base de données.

- Pour une surface parfaitement horizontale, nous nous intéressons à l'irradiance totale reçue, appelée *Global Horizontal Irradiance (GHI)*. Elle est obtenue en **sommant** la projection de l'*irradiance normale directe (DNI)* sur la surface horizontale, selon l'*angle zénithal θ* , ainsi que le *rayonnement diffus horizontal (DHI)*.
- Pour une surface inclinée, nous nous intéressons à l'irradiance totale reçue par cette surface, appelée *Global Tilted Irradiance (GTI)*. Elle est obtenue en **sommant**

trois composantes : la projection de l'*irradiance normale directe (DNI)* sur le plan incliné selon l'*angle d'incidence* θ , le *rayonnement diffus horizontal (DHI)*, ainsi que l'*irradiance réfléchie par le sol (RI)*.



La figure ci-dessus montre l'évolution des différentes composantes du rayonnement solaire enregistrées lors d'une journée de beau temps au cours d'une campagne de mesure. On observe notamment un pic d'irradiance autour de midi, le moment où le soleil est au plus haut dans le ciel. Ces courbes représentant le **GHI**, le **DNI**, le **DHI** et le **GTI** permettent de mieux visualiser l'impact de la position du Soleil et des conditions atmosphériques sur la quantité d'énergie reçue par les surfaces au sol.

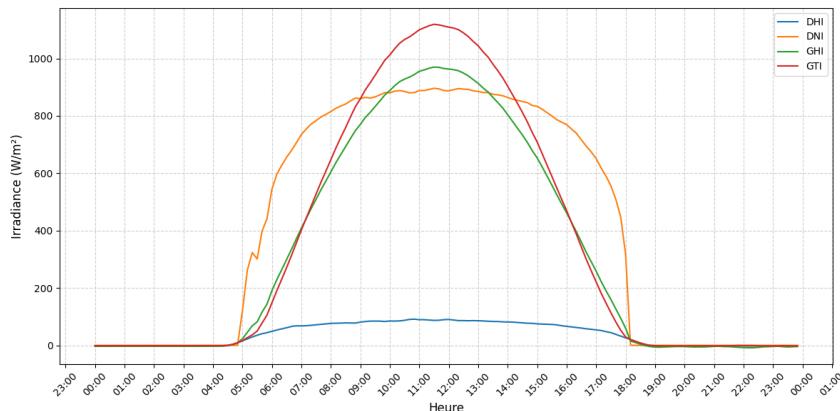


FIGURE 3 – Profils des quatre composantes de l'irradiance solaire par temps clair

4 Présentation DataWin

DataWin



DataWin est un outil interne d'ENGIE Green, développé pour récupérer automatiquement des mesures météorologiques issues de nombreuses campagnes, et pour estimer le productible de projets éoliens et photovoltaïques.

Comme le montre la figure ci-dessous, cet outil repose sur deux grands volets :



- D'abord, la gestion des campagnes de mesures au quotidien : cela passe par l'import et l'export des données, la visualisation et validation des données (notamment en cas de panne d'un instrument), ainsi que la gestion des pièces et des interventions sur le terrain.
- Ensuite, l'estimation du productible, c'est-à-dire d'estimer quantité d'énergie qu'un site peut produire. Pour cela, des modèles physiques et des outils de data science sont utilisés.

Même si DataWin est un outil plus vaste et plus polyvalent que le cadre de mon travail, et que je n'en ai pas découvert toutes les fonctionnalités, il m'a servi de référence pour la visualisation des données d'irradiance. Il m'a également aidé à mieux comprendre l'invalidation de certaines mesures dans différents scénarios, ainsi que le processus d'exportation des données.

4.1 Structure de la base de données sur DataWin

Sur l'ensemble des campagnes solaires, des mesures d'irradiance et de météo sont enregistrées dans la base de données toutes les 10 minutes.

Nous disposons actuellement de 110 campagnes solaires centralisées. La durée d'une campagne dépend de l'âge de la centrale photovoltaïque, mais en moyenne, chaque campagne couvre une période de 4 ans, entre 2021 et 2025. Avec un pas de temps de 10 minutes, cela représente 144 mesures par jour. Chaque campagne contient en moyenne 8 variables liées à l'irradiance et à la météo, soit environ 1 682 688 enregistrements par campagne (4 ans \times 365,25 jours \times 144 mesures/jour \times 8 variables). Au total, cela représente plus de 185 millions de mesures.

Typiquement, pour chaque campagne, on retrouve :

- des colonnes correspondant aux mesures de **GHI**. Il peut y avoir plusieurs colonnes GHI, car plusieurs instruments différents mesurent parfois la même grandeur.
- des colonnes correspondant aux mesures de **GTI**. Comme pour le GHI, il peut y avoir plusieurs colonnes GTI car plusieurs capteurs mesurent cette grandeur, installés avec des inclinaisons et orientations parfois différentes. L'**orientation** indique la direction vers laquelle un capteur est tourné, exprimée en degrés par rapport au nord : par exemple, une orientation de 0° signifie que le capteur est dirigé vers le nord, 90° vers l'est, 180° vers le sud, et 270° vers l'ouest. En pratique, on cherche généralement à orienter les capteurs plein sud (180°), mais selon la configuration de la parcelle et l'optimisation du site, cela n'est pas toujours possible.

Ce choix d'orientation n'est pas fait au hasard : en France, située dans l'hémisphère nord, le soleil se lève à l'est, se couche à l'ouest, et atteint son point le plus haut dans le ciel lorsqu'il est situé au sud. C'est pourquoi, pour capter un maximum d'énergie solaire, les capteurs sont en général orientés vers le **sud**.

- des colonnes correspondant aux mesures météorologiques telles que la température, la pression, l'humidité et la pluviométrie.

TABLE 1 – Extrait du jeu de données

Date	GHI A	GHI B	GTI 180° 20 F	GTI 180° 20 G	GTI 180° 20 H	GTI 180° 20 I	Pluie cumul 3m	Pluie intensité 3m	Press 3m	Temp ext 3m	HR 3m	Temp_ext 2.5m
12/03/2025 06 :20	2.723	2.951	2.629	2.843	2.527	2.897	9.47	2.965	975.0	6.974	99.8	6.955
12/03/2025 06 :30	4.054	4.261	3.946	4.092	3.750	4.173	9.96	2.733	975.0	6.768	99.5	6.747
12/03/2025 06 :40	6.364	6.535	6.115	6.173	5.801	6.338	10.27	1.072	975.0	6.400	98.7	6.411
12/03/2025 06 :50	9.730	9.810	9.360	9.470	9.100	9.590	10.52	2.183	975.0	6.108	98.2	6.135
12/03/2025 07 :00	12.230	12.220	11.610	11.790	11.440	11.900	10.95	2.448	976.0	5.922	97.9	5.957
12/03/2025 07 :10	17.220	17.150	16.260	16.480	16.140	16.540	11.43	3.453	976.0	5.846	97.5	5.884

Dans ce qui suit, nous présenterons un aperçu des difficultés liées au traitement des données, et surtout à l'apprentissage statistique.

4.1.1 Des structures de données hétérogènes entre campagnes

Ce point n'est pas un problème en soi, mais plutôt une remarque : les bases de données exportées depuis DataWin n'ont souvent pas la même structure d'une campagne de mesure

à l'autre, notamment en termes de nombre de colonnes. Cela s'explique par les différences d'instrumentation, notamment le nombre de capteurs **GHI**, **GTI** et météo installés. (cf. figure 2)

Sur l'ensemble des campagnes solaires centralisées, nous disposons au maximum de 5 mesures **GHI**, 7 mesures **GTI**, 4 mesures de température, 2 mesures d'humidité, 2 mesures d'intensité de pluie, 1 mesure de cumul de pluie et 2 mesures de pression.

D'autre part, en pratique, les mesures d'irradiance et de météo ne sont pas toujours disponibles en continu. Il arrive que certains instruments tombent en panne, que des problèmes de communication surviennent, ou que les dataloggers cessent temporairement de fonctionner. Lorsqu'une mesure est manquante, nous la remplaçons par -1000 au lieu de **NaN**, afin que le modèle puisse clairement identifier qu'il ne s'agit pas d'une vraie mesure et éviter de la traiter comme une donnée valide.

4.1.2 Variabilité des instruments GTI au sein d'une même campagne

En revanche, ce qui constitue un véritable obstacle dans le traitement des données — et qui impacte directement la manière dont on entraîne le modèle —, c'est le fait que, sur une même campagne, les instruments de mesure de **GTI** peuvent être installés avec des inclinaisons et des orientations différentes. Il est alors difficile de traiter de manière homogène les mesures issues de ces configurations variées.

Un arbre de décision, qui traite les colonnes indépendamment, ne pourra pas faire cette distinction. Une solution possible serait d'ajouter des colonnes supplémentaires indiquant les inclinaisons et les orientations associées à chaque mesure **GTI**. Mais cela pose un autre problème : lors de l'apprentissage, le modèle risque de s'appuyer principalement sur certaines inclinaisons et orientations particulières présentes dans les données, ce qui limiterait sa capacité à généraliser. Or, notre objectif est justement qu'il puisse apprendre sur un ensemble le plus diversifié possible de situations.

4.1.3 Manque de cohérence entre campagnes dans les capteurs disponibles

D'un autre côté, les inclinaisons et orientations varient aussi très souvent d'une campagne à l'autre. Sur certaines grandes campagnes incluses dans l'ensemble d'apprentissage, il est possible que le modèle finisse par surapprendre sur une portion restreinte de ces configurations.

Par ailleurs, comme les campagnes sont équipées différemment, il peut arriver qu'une campagne ne contienne aucune mesure de **GHI**, ou au contraire aucune mesure de **GTI**. Ce type de cas complique l'apprentissage, car le modèle risque d'avoir du mal à généraliser ou à adopter un comportement cohérent face à ces situations.

Maintenant que la structure de la base de données d'irradiance a été détaillée, intéressons-nous aux étapes de validation des mesures. Pour cela, il est essentiel de se familiariser avec l'outil central de ce processus : les marqueurs d'invalidité.

4.2 Marqueurs d'invalidité

Les marqueurs d'invalidité sont des outils algorithmiques conçus pour alerter les utilisateurs sur d'éventuelles mesures suspectes. En pratique, ils permettent de guider l'utilisateur vers les zones à surveiller en priorité, en attribuant à chaque mesure un état (valide ou invalide).

En résumé, les marqueurs sont des outils permettant de détecter des pannes et d'invalider des mesures liées à des capteurs défaillants.

Il existe en général deux types de marqueurs, les marqueurs automatiques et les marqueurs utilisateurs :

Les marqueurs automatiques sont conçus pour détecter automatiquement des mesures défaillantes, sans intervention de l'utilisateur. Par exemple, ils peuvent signaler (ou "flager") des mesures d'irradiance qui dépassent largement les seuils considérés comme physiquement acceptables.

Prenons un exemple : une mesure de **GHI** enregistrée lors d'une journée nuageuse ne devrait jamais dépasser le **GHI clearsky**, c'est-à-dire la valeur maximale théorique en ciel clair. Pourtant, dans certains cas, cela peut se produire. Sur l'illustration ci-dessous, on observe un dépassement du **GHI clearsky**, causé par la présence de givre au sol. Le givre agit comme un miroir et réfléchit l'irradiance vers le capteur **GHI**, faussant ainsi la mesure.



FIGURE 4 – Invalidation automatique d'une mesure **GHI**

D'autre part, on retrouve les marqueurs utilisateurs, créés par l'utilisateur à partir des marqueurs automatiques. En effet, ces marqueurs signalent des mesures potentiellement suspectes à l'aide d'algorithmes. C'est ensuite au technicien de Mesures de confirmer ou non la validité de ces mesures, en s'appuyant sur son expertise et les éléments à sa disposition.

Un exemple de ce type de marqueurs est présenté ci-dessous : un marqueur automatique a invalidé une mesure de **GTI** en raison d'un écart relatif supérieur à 4% par rapport aux

deux autres mesures **GTI** disponibles, tandis qu'une autre a été validée car son écart avec la mesure de référence restait inférieur à ce seuil. Ce seuil de 4% correspond à la tolérance cumulée des instruments utilisés : en effet, chaque capteur présente une incertitude d'environ 2%, ce qui implique qu'entre deux instruments identiques, l'écart maximal attendu est de l'ordre de 4%.

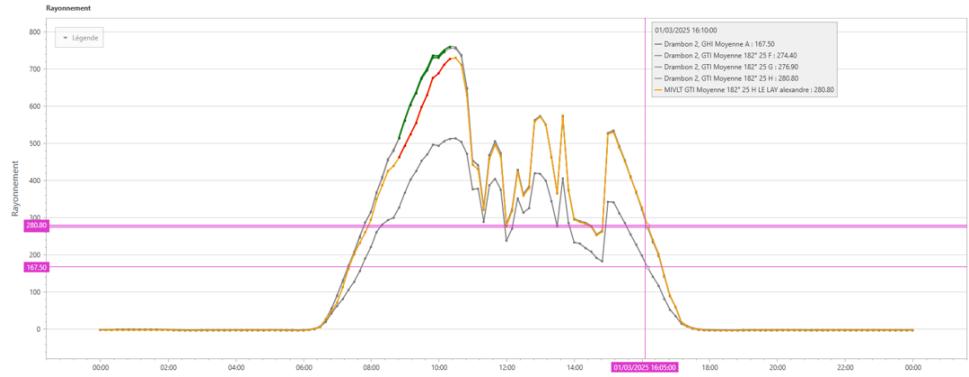


FIGURE 5 – Invalidation utilisateur suite à un marqueur automatique

4.3 Base de données des états de validité des mesures

En pratique, les périodes pendant lesquelles les marqueurs automatiques/utilisateurs valident ou invalident les mesures sont enregistrées dans la base de données, avec un label associé : 1 pour une mesure valide, 0 pour une mesure invalide. Les données sont stockées sous un format similaire au tableau suivant :

TABLE 2 – Exemple de base de données des états de mesures

Début	Fin	Etat
2024-02-15 00 :00	2024-02-15 00 :00	1
2018-06-04 11 :20	2018-06-04 11 :40	0
2018-08-25 11 :10	2018-08-25 11 :40	1
2019-05-25 12 :00	2019-05-25 12 :20	1
2020-04-23 11 :10	2020-04-23 11 :30	0
2020-06-02 11 :10	2020-06-02 11 :30	1

Dans tout cela, seules les états de validité liées aux mesures d'irradiance nous intéressent ; nous ne prenons pas en compte les états des mesures météo.

4.3.1 Gestion des mesures non marquées

Un choix important que nous avons fait concerne les périodes où aucune validation ou invalidation explicite n'a été enregistrée. Dans ces cas-là, nous avons décidé de faire confiance au fait qu'aucun marqueur n'ait été déclenché, et donc de considérer la mesure comme

valide par défaut.

C'était un choix incontournable, mais pas évident, car les marqueurs automatiques comme les validations utilisateurs, ne sont pas toujours fiables. Il reste donc un certain risque, mais c'était la solution la plus raisonnable pour pouvoir avancer dans le traitement des données. (voir exp -10)

4.3.2 Restriction des périodes d'apprentissage aux mesures explicitement validées

Toujours concernant la base de données des états, avant l'entraînement d'un modèle de machine learning sur les mesures d'irradiance, nous avons choisi de restreindre la plage temporelle des mesures aux seules périodes pour lesquelles nous disposons d'états de validité.

La raison de ce choix est la suivante : en dehors de ces périodes, nous ne savons pas exactement ce qui s'est passé ni pourquoi aucun marqueur ne s'est déclenché. Cela pourrait être lié à un problème dans les marqueurs eux-mêmes. En particulier, l'historique n'est pas toujours fiable : sur certaines anciennes campagnes, il n'est pas garanti que les marqueurs aient bien fonctionné. Les marqueurs solaires n'ont été créés qu'à partir de 2020, alors qu'une partie des campagnes est bien plus ancienne. Théoriquement, on pourrait relancer ces marqueurs sur l'historique, mais comme ils nécessitent une validation manuelle, cela n'est pas possible. Ce choix confère également deux avantages importants : il permet à la fois de réduire l'incertitude liée aux données d'irradiance utilisées pour l'apprentissage et de diminuer substantiellement la taille de la base de données.

4.3.3 Comprendre les invalidations en termes statistiques

Pour mieux comprendre comment le modèle de machine learning pourrait se comporter, il est utile de savoir quand les mesures sont invalidées. Dans le graphique ci-dessous, on observe la distribution des invalidations sur l'année. Ce graphe permet de repérer les périodes plus problématiques de l'année, comme l'hiver par exemple. Ces informations peuvent aider à expliquer certaines erreurs du modèle et à améliorer son fonctionnement dans le futur.

Ci-dessous, un graphique illustrant la distribution des mesures invalides au cours de l'année :

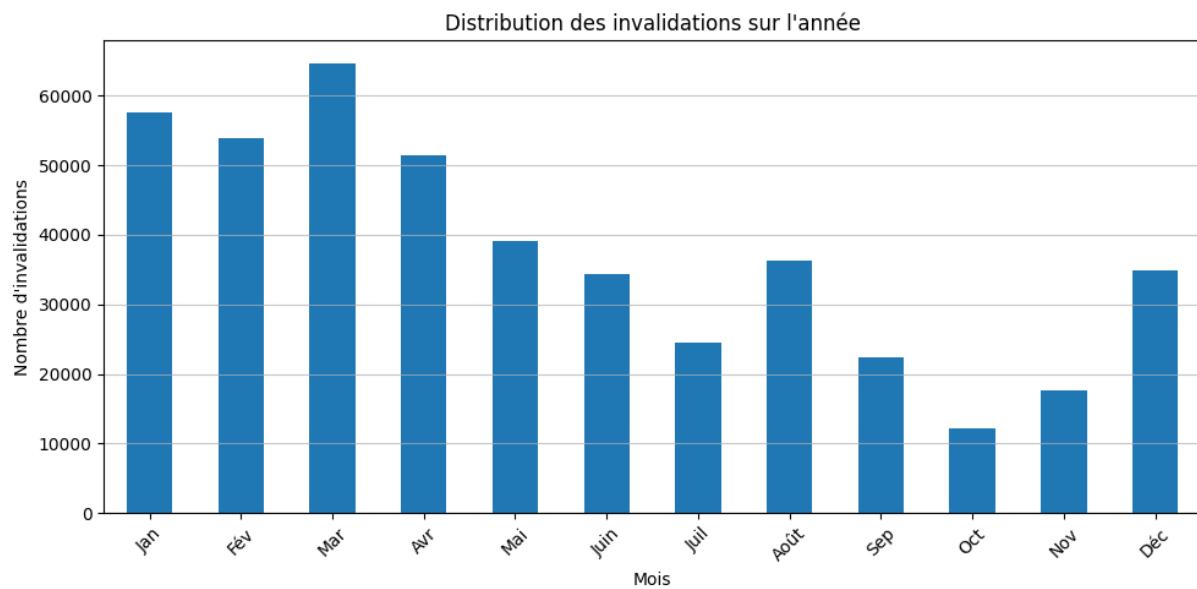


FIGURE 6 – Distribution des invalidations sur l'année

D'après la figure 6, nous nous apercevons que nous avons à peu près le même nombre d'invalidations en hiver (octobre - mars) qu'en été (avril - septembre). En revanche, un grand nombre d'invalidations est concentré entre janvier et mars. Il est difficile d'en donner une explication unique et certaine.

Enfin, nous nous intéressons également à la distribution des taux d'invalidation en fonction des mesures d'irradiance **GHI** et **GTI** :

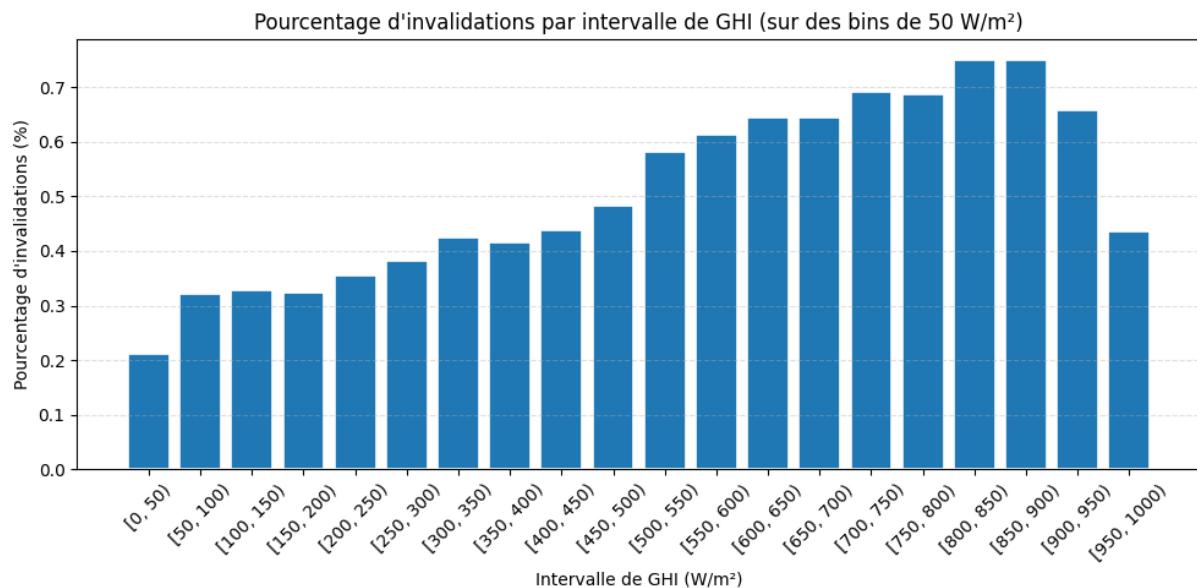


FIGURE 7 – Distribution des taux d'invalidations en fonction des mesures **GHI**

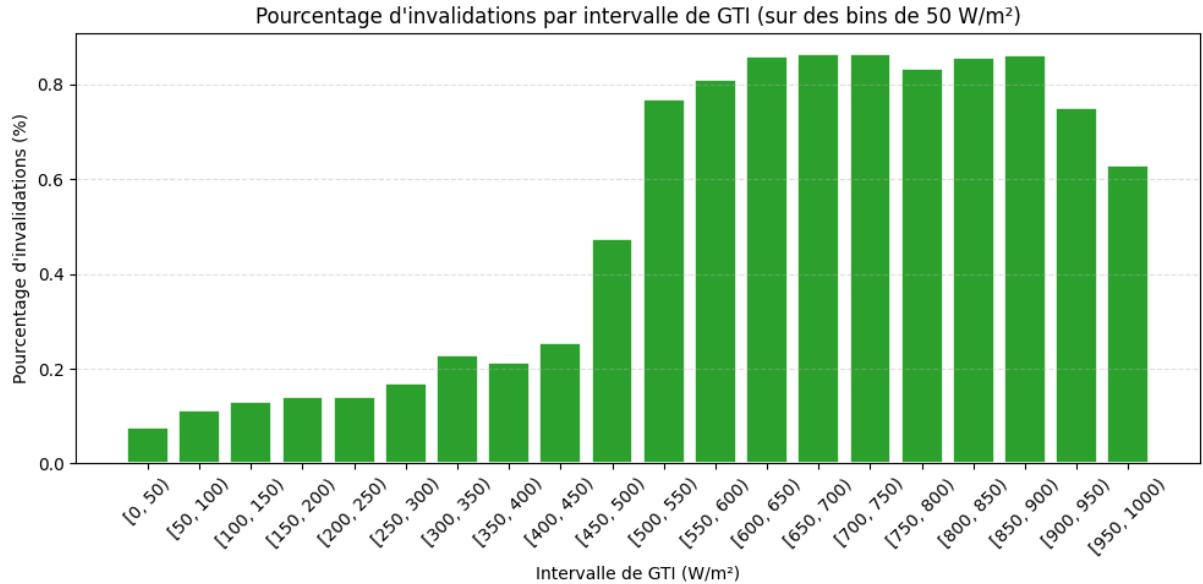


FIGURE 8 – Distribution des taux d'invalidations en fonction des mesures **GTI**

Les figures 7 et 8 montrent la distribution des invalidations en fonction des mesures de **GHI** et **GTI**. On observe que le taux d'invalidation est faible pour les faibles irradiances, puis augmente nettement avec l'irradiance pour atteindre un maximum entre 800 et 900 W/m² pour les mesures **GHI** et 600 et 900 W/m² pour les mesures **GTI**. Cela traduit le fait que les invalidations surviennent surtout lorsque le rayonnement est fort, période où les écarts entre capteurs sont les plus visibles.

À ce stade, après avoir exporté les bases de données depuis DataWin, nous disposons d'une série temporelle à pas de 10 minutes regroupant les mesures d'irradiance et de météo, accompagnées de la série temporelle de validation/invalidation de chaque mesure. Cependant, ces seules mesures ne suffisent pas pour permettre au modèle de déterminer avec un bon niveau de confiance si une mesure est valide ou non.

5 Création de features avancées pour améliorer l'apprentissage

Pour améliorer l'apprentissage et donner au modèle le plus d'informations utiles possible, nous avons enrichi la base de données initiale en ajoutant des variables explicatives supplémentaires. Ce choix s'appuie d'une part sur une bonne compréhension des différents scénarios de terrain, construite à travers de nombreux échanges avec l'équipe Mesure, qui s'occupe du processus de validation, et l'équipe Modélisation et Data Science, qui a l'habitude d'utiliser des modèles de machine learning sur des données solaires.

D'autre part, on a aussi choisi ces variables en se basant sur notre propre expérience : on a testé ce qui marche et ce qui ne marche pas, et on a affiné petit à petit notre compréhension de ce que le modèle doit apprendre à repérer.

Pour cela, on va présenter plus en détail les variables explicatives qu'on a choisies d'ajouter avant d'entraîner le modèle.

5.1 Azimut et élévation solaire

La position exacte du Soleil peut être identifiée à l'aide de deux coordonnées : l'azimut et l'élévation.

L'azimut est l'angle horizontal entre le nord et la position du Soleil.

L'élévation est l'angle vertical entre l'horizon et le Soleil : 0° signifie que le Soleil est sur l'horizon.

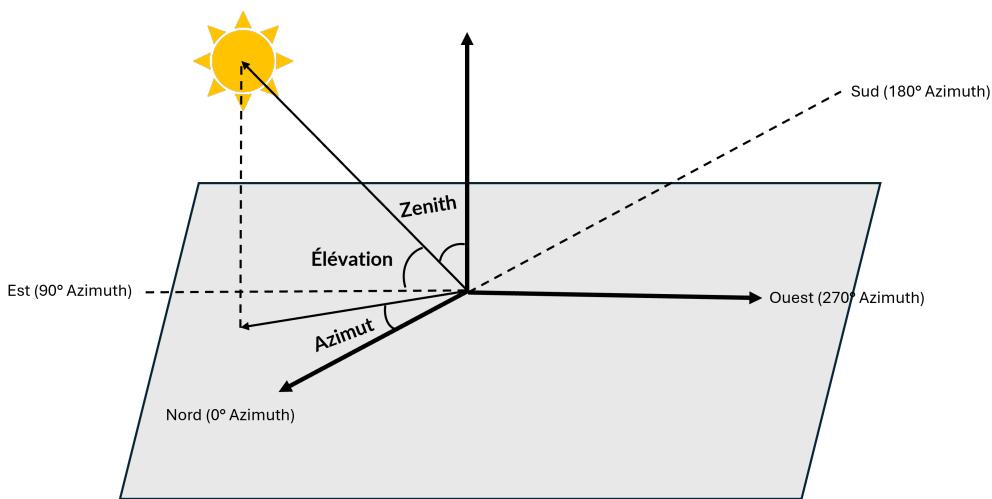


FIGURE 9 – Angles d'azimut et élévation solaire

Comme déjà mentionné, notre base de données contient une colonne **Date** indiquant le jour, l'heure et la minute de chaque mesure.

Pour éviter d'introduire un biais dans l'apprentissage, nous avons choisi de ne pas utiliser directement cette colonne, mais plutôt de la remplacer par deux colonnes : **l'azimut** et **l'élévation** du soleil.

Cette décision repose sur plusieurs points liés à l'apprentissage :

- La plupart des modèles de machine learning classiques ne sont pas capables de manipuler des dates et d'en tirer de l'information utile.
- Surtout, il n'y a aucune logique entre une date précise et une invalidation : une invalidation n'arrive pas systématiquement un 15 mars à 15h40. En revanche, la position du soleil (azimut et élévation) est directement liée au niveau d'irradiance mesuré.
- L'utilisation de la position solaire permet ainsi au modèle d'apprendre sur une campagne, puis de répéter ce comportement sur une autre, ce qui favorise sa capacité de généralisation. L'utilisation de la position solaire permet ainsi au modèle d'apprendre sur une campagne, puis de reproduire ce comportement sur une autre, ce

qui favorise sa capacité de généralisation. De plus, selon la position géographique, le Soleil n'a pas la même position au même instant.

L'azimut et l'élévation résolvent ce problème : ces variables offrent davantage de **variabilité** tout en permettant de représenter la position réelle du soleil à chaque instant. Elles facilitent aussi la distinction entre différentes campagnes, qui se trouvent à des **latitudes, longitudes et altitudes** distinctes.

Toujours dans l'optique de mieux comprendre l'invalidation des mesures en fonction des différentes variables explicatives utilisées, nous présentons ci-dessous une *heatmap* représentant le pourcentage d'invalidation en fonction de l'azimut et de l'élévation :

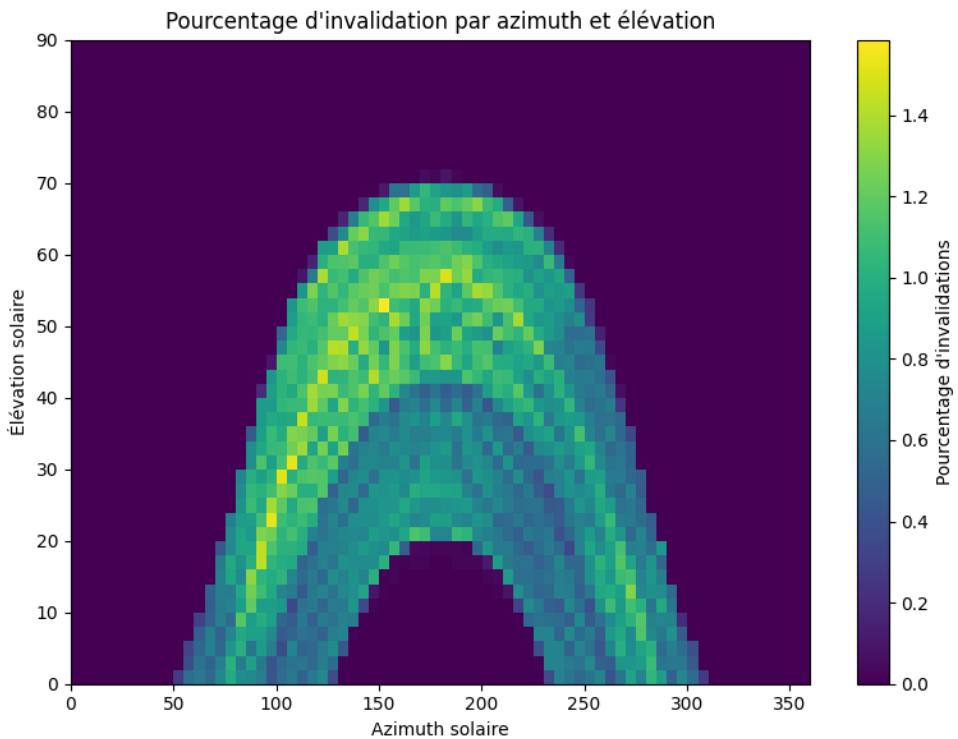


FIGURE 10 – Distribution des invalidations par azimut et élévation

En observant cette distribution, nous distinguons deux grandes saisons : l'hiver et l'été.

La partie supérieure du paraboloïde correspond à l'été, tandis que la partie inférieure correspond à l'hiver.

5.1.1 Présentation de la bibliothèque pvlib

`pvlib` [1] est une bibliothèque Python open-source développée par des chercheurs en énergie photovoltaïque à travers le monde. Elle fournit une large gamme d'outils et de fonctions pour simuler les performances des systèmes photovoltaïques.



FIGURE 11 – Logo **pvlib**

À titre d'exemple, **pvlib** permet de calculer très précisément la position solaire (azimut, élévation), ainsi que d'estimer différentes composantes d'irradiance (**GHI**, **GTI**, **DHI**, **DNI**). Il offre également des outils pour modéliser certains phénomènes physiques liés à l'énergie solaire, comme les pertes de productible.

5.1.2 Calculer l'azimut et l'élévation solaire sous **pvlib**

Pour calculer les azimuts et élévations solaires toutes les 10 minutes, nous utilisons la bibliothèque **pvlib**. Le calcul se déroule en deux étapes :

1. Récupérer la latitude, la longitude et l'altitude précises de chaque campagne solaire à partir de la base de données **DataWin**, afin de créer l'objet **pvlib.location.Location**.

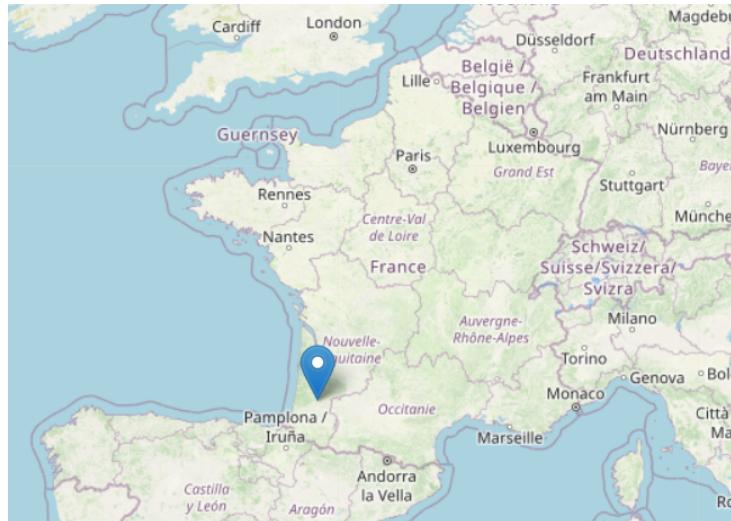


FIGURE 12 – Visualisation géographique des sites

2. Appliquer la fonction **get_solarposition** sur ce dernier, en lui donnant en entrée les dates, et optionnellement les températures et pressions.

En sortie, on récupère un tableau contenant les azimuts et élévations.

En été, pour une des campagnes solaires, ci-dessous un exemple d'élévations et d'azimuts solaires :

TABLE 3 – Exemples d’élévations et d’azimuts solaires

Date	Élévation	Azimut
2024-07-09 07 :20	11.3°	69.8°
2024-07-09 14 :10	68.2°	197.5°
2024-07-09 16 :10	52.9°	247.3°
2024-07-09 21 :10	0.8°	300.8°
2024-07-10 04 :20	-15.1°	37.5°

5.2 Utilisation de pvlib pour estimer le rayonnement solaire théorique

Le **GHI clearsky** représente la quantité d’irradiance horizontale que l’on recevrait au sol s’il n’y avait aucun nuage, autrement dit dans des conditions atmosphériques idéales. C’est une valeur théorique très utile que nous exploitons, car elle permet de comparer les mesures **GHI** réelles obtenues avec ce qu’on devrait normalement observer. Grâce à cette comparaison entre les **GHI** et les **GHI clearsky**, on peut repérer d’éventuelles mesures invalides.

Théoriquement, une mesure de **GHI** ne peut pas dépasser son **GHI clearsky** même sur une journée ensoleillée. Comme nous l’observons dans la figure 4, pendant la période où le **GHI** dépasse le **GHI clearsky**, la mesure a été automatiquement flaguée comme invalide.

Une fois les localisations (latitude, longitude et altitude) de chacune des campagnes récupérées depuis la base de données, nous utilisons la fonction `get_clearsky()` de `pvlib` qui calcule le **GHI clearsky**.

Ci-dessous, un exemple de tracé du **GHI clearsky** comparé à une mesure **GHI** :

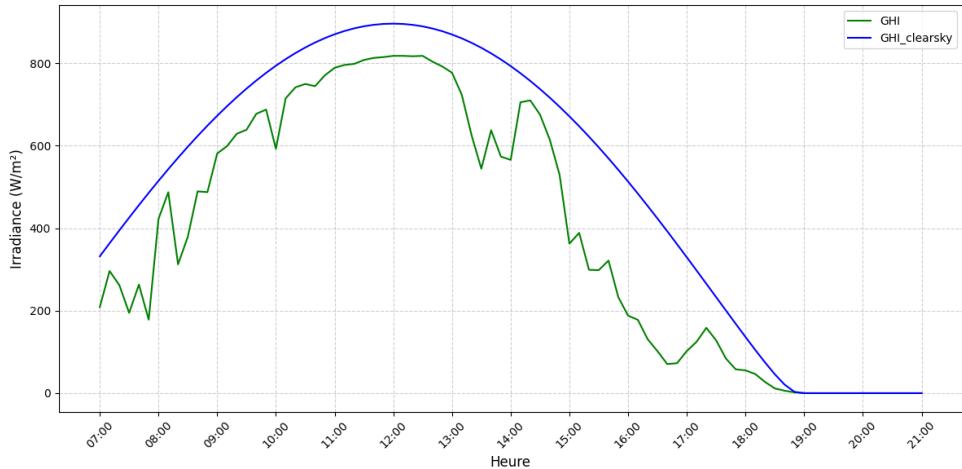


FIGURE 13 – Tracé de **GHI clearsky**

5.2.1 Indices beau-temps

Nous ajoutons également à l'ensemble des variables explicatives ce qu'on appelle des indicateurs beau temps, en anglais appelés *clearness index*. Ces indicateurs mesurent le rapport entre la production réelle **GHI** et la production théorique qu'on aurait dû observer dans des conditions idéales **GHI clearsky**. Pour chaque mesure **GHI_j**, nous ajoutons une colonne dans laquelle l'indice beau-temps $k_j^{clearsky}$ est calculé comme suit :

$$k_j^{clearsky} = \frac{\text{GHI}_j}{\text{GHI_clearsky}}$$

5.3 Écarts absolus entre les pyranomètres

Dans les campagnes centralisées, où tous les capteurs **GHI** et **GTI** sont installés très proches les uns des autres, il est pertinent de comparer les mesures du même type deux à deux afin de détecter d'éventuels écarts anormaux. Dans notre cas, on calcule tous les écarts absolus, deux à deux, entre les mesures **GHI** prises toutes les 10 minutes, ainsi que pour les mesures **GTI**.

Cette idée est très utile, car si les capteurs de même nature sont censés mesurer la même chose, un gros écart sur une certaine période peut indiquer un problème.

Même dans des situations un peu compliquées, comme le passage d'un nuage, cette méthode reste fiable dans le cas des campagnes centralisées : à l'échelle de 10 minutes, tous les capteurs observent globalement les mêmes conditions.

En pratique, sur toutes les campagnes centralisées, nous disposons au maximum de 5 mesures **GHI** et de 7 mesures **GTI**. Nous créons alors $\binom{5}{2} = 10$ colonnes d'écarts pour les mesures **GHI** et $\binom{7}{2} = 21$ colonnes d'écarts pour les mesures **GTI**, sur chaque campagne. Ces écarts sont calculés toutes les 10 minutes.

En fonction des capteurs **GHI** et **GTI** réellement disponibles sur chaque campagne, nous remplissons ces colonnes avec les écarts correspondants. Pour les combinaisons absentes (par exemple, lorsqu'un capteur est manquant), les colonnes concernées sont remplies avec des valeurs NaN.

Ci-dessous, un exemple d'écarts calculés entre trois mesures **GTI**

$$\Delta_1 = | GHI_1 - GHI_2 | \quad \Delta_2 = | GHI_2 - GHI_3 | \quad \Delta_3 = | GHI_1 - GHI_3 |$$

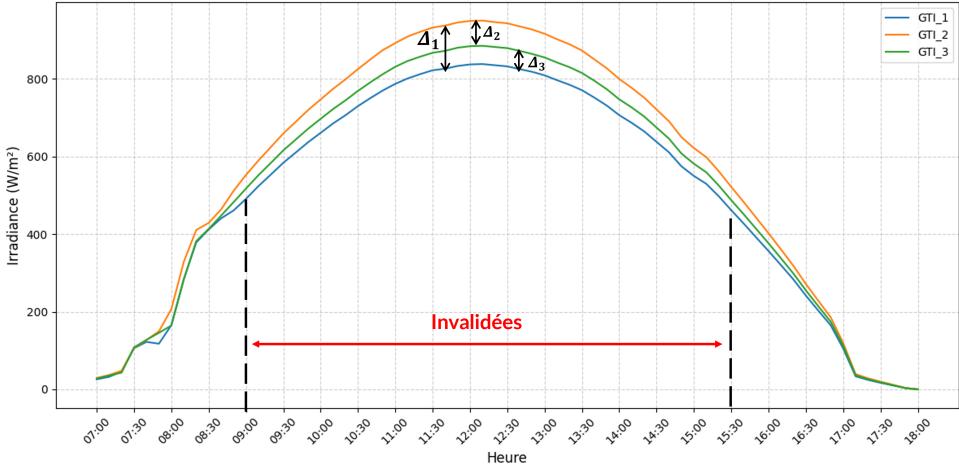


FIGURE 14 – Écarts calculés entre trois mesures **GTI**

Pour cette journée, toutes les mesures **GTI** ont été invalidées de **9h00** jusqu'à **15h30**, les écarts absolu s étant supérieurs à la limite acceptable.

5.4 Données satellitaires (SolEye)

SolEye est une application web développée par ENGIE Green qui permet de générer des données solaires à partir d'images satellitaires provenant du satellite Meteosat-10.

Les données satellites récupérées via l'application SolEye se sont révélées très utiles, puisqu'elles servent de données d'irradiance de référence supplémentaire que nous ajoutons aux données existantes. De plus, elles ont l'avantage d'être toujours disponibles et également d'être une source fiable, comme l'a montré une étude de Natural Power en 2022.

Ces données satellitaires d'irradiance **GHI**, **DHI** et **DNI** m'ont été fournies pour les campagnes centralisées à partir de 2008, par l'équipe Data Science & Modélisation.

5.4.1 Fraction diffuse

La fraction diffuse, appelée en anglais *diffuse fraction*, est un indicateur défini comme le rapport entre le rayonnement diffus horizontal **DHI** et le rayonnement global horizontal **GHI**, selon la formule suivante :

$$k^d = \frac{\text{DHI}}{\text{GHI}}$$

Cet indicateur est particulièrement utile pour caractériser les conditions atmosphériques. Par exemple, dans des journées très nuageuses, cet indicateur k^d est élevé car le rayonnement diffus domine. À l'inverse, dans une journée où il fait beau-temps, la composante directe **DNI** est forte et la fraction diffuse est faible.

Les mesures **DHI** sont rarement disponibles sur l'ensemble des campagnes solaires. C'est

pourquoi, pour calculer la fraction diffuse, nous utilisons les données **DHI** issues de l'application SolEye, qui ont l'avantage d'être toujours disponibles pour toutes les campagnes solaires que nous traiterons.

5.5 Cumuls d'irradiance

Les dernières variables explicatives que nous ajoutons sont les cumuls pour chaque instrument de mesure **GHI**. L'idée principale est de pouvoir les comparer aux cumuls obtenus à partir du **GHI** fourni par l'application SolEye, qui est généralement plus stable et consistant, surtout sur de longues périodes.

Par ailleurs, pour les mesures **GHI** et **GTI**, l'objectif est aussi de donner au modèle, à chaque pas de temps, une information sur ce qui s'est passé avant.

5.5.1 Cumul journalier

Par définition, le cumul journalier d'irradiance est calculé à chaque pas de temps, et représente l'irradiance totale reçue depuis le début de la journée jusqu'à l'instant considéré. Ces cumuls journaliers sont calculées pour chacune des mesure **GHI_i** et **GTI_j** à l'instant t_n comme suit :

$$C_i(t_n) = \sum_{k=0}^n \mathbf{GHI}_i(t_k) \quad C'_i(t_n) = \sum_{k=0}^n \mathbf{GTI}_j(t_k)$$

5.5.2 Cumul décalé

En analysant certaines journées comportant des mesures invalidées, on remarque que, dans certains cas, les invalidations s'étendent sur plusieurs jours.

Pour en tenir compte, nous proposons d'observer, pour chaque journée, ce qui s'est passé les jours précédents (la veille, deux jours avant, trois jours avant). L'idée de calculer ces *cumuls décalés* sur 1 jour, 2 jours, 3 jours et 7 jours repose sur le fait que cumuler les mesures permet de réduire la variabilité temporelle et ainsi de mieux détecter les écarts de comportement.

Concrètement, cela revient à calculer, pour chaque journée, le cumul d'irradiance mesurée la veille, sur les deux jours précédents, et sur les trois jours précédents.

Pour disposer d'une métrique de référence avec laquelle comparer les cumuls issus des mesures **GHI**, nous utilisons les données SolEye. Ces données sont en effet adaptées à ce type de comparaison, car elles sont suffisamment consistantes sur plusieurs jours.

La figure ci-dessous illustre de manière heuristique pourquoi il est pertinent de comparer les mesures **GHI** aux données SolEye :

En effet, la figure 24 montre que l'irradiance **GHI** diminue chaque jour : l'aire sous sa courbe devient de plus en plus petite. À l'inverse, l'irradiance **SolEye** reste stable pendant quatre jours.

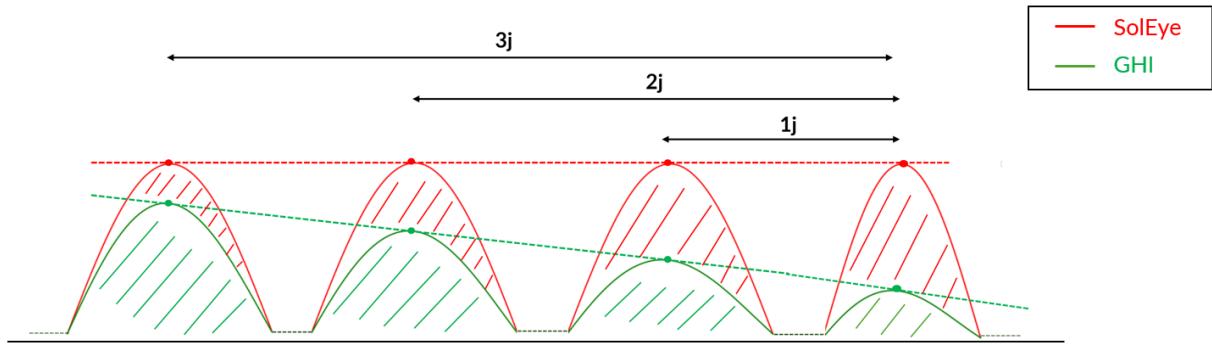


FIGURE 15 – Cumul décalé 1 jour, 2 jours, 3 jours.

En cumulant les valeurs, on lisse et on cache les variations journalières brusques, ce qui permet de mieux repérer la tendance générale d'un capteur à dysfonctionner.

5.6 Uniformisation des mesures GTI

5.6.1 Algorithme de transposition inverse

Les inclinaisons et orientations des mesures **GTI** peuvent varier d'un capteur à l'autre au sein d'une même campagne, et même d'une campagne à l'autre. De plus, en raison des disponibilités différentes des mesures (**GHI**, **GTI**) selon les campagnes, et de l'incapacité du modèle à faire la distinction entre toutes ces configurations différentes, cela peut poser problème lors de l'apprentissage.

Pour contourner ceci, nous proposons une approche d'uniformisation des mesures **GTI** qui consiste à transformer toutes les mesures **GTI** sur toutes les campagnes, en des mesures **GHI**.

Pour cela, nous nous appuyons sur l'algorithme de transposition inverse [2], déjà implémenté dans la bibliothèque `pvlib`.

Le modèle de transposition inverse permet de convertir les mesures **GTI** en mesures **GHI**. En d'autres termes, il permet d'estimer ce que pourrait donner une mesure **GHI**, à partir de la mesure **GTI**.

En entrée, le modèle de transposition inverse prend les mesures **GTI** d'un capteur, ainsi que son inclinaison, son orientation, et la position du soleil à chaque pas de temps. En sortie, il fournit une estimation des mesures **GHI**.

Dans le graphique ci-dessous, nous visualisons deux courbes : la première représente les mesures d'un capteur **GHI**, et la seconde correspond aux mesures **GTI** d'un autre capteur, transformées en **GHI** à l'aide du modèle de transposition inverse. Nous désignerons cette seconde courbe par **GHI_from_GTI**.

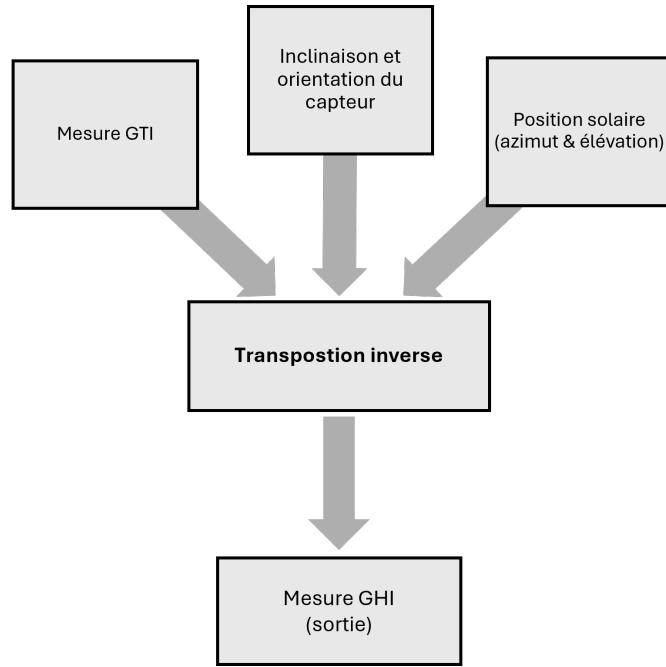


FIGURE 16 – Diagramme illustrant l'algorithme de transposition inverse

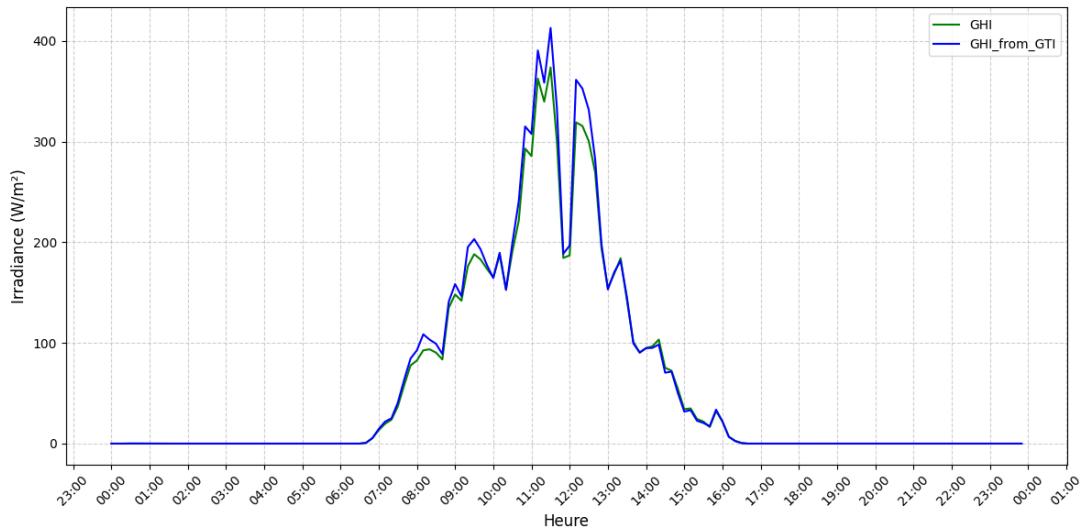


FIGURE 17 – Estimation des mesures **GHI** à partir des mesures **GTI** en utilisant la transposition inverse

Ce graphique illustre notre objectif : estimer les mesures **GHI** à partir des mesures **GTI**, afin d'uniformiser les données issues de capteurs aux caractéristiques différentes, en utilisant l'algorithme de transposition inverse.

6 Méthodes et métriques

Le jeu de données utilisé regroupe des mesures d'irradiance et météo de 110 campagnes centralisées, soit un dataset d'environ 18 millions de lignes et 25 colonnes (avant l'ajout de nos variables explicatives supplémentaires).

Au cours du stage, il a été décidé de débuter les expérimentations avec l'apprentissage sur un sous-ensemble de 10 campagnes centralisées. Ce choix a été effectué en concertation avec l'équipe Mesures, en s'appuyant sur les critères suivants :

- un taux d'invalidation significatif sur la campagne, récupéré à partir des données dans DataWin : nous nous intéressons aux campagnes présentant un écart relativement élevé entre la disponibilité brute et la disponibilité valide, c'est-à-dire une disponibilité invalide relativement élevée

Département	Zone	Energie	Périmètre	Type	Moyen Mesure	Etat	Hauteur mesures	Dispo brute totale	Dispo valide
Manche	EOL Normandie	Éolien	SAMEOLE	Développement	Aucun	En Opération	81	100%	81,56% -> 100%
Sarthe	EOL Nord Ouest	Éolien	ENGIE GREEN	Développement	DL Campbell	En Opération	101	100%	95,64% -> 99,97%
Dordogne	PV-Nouvelle Aquitaine	SOLAIRE CAS	ENGIE GREEN	Développement	DL Campbell	En Opération	2	100%	95,83% -> 100%
Pyrénées-Orientales	EOL Sud Est	Éolien	ENGIE GREEN	Exploitation	DL Campbell	En Opération	5	100%	99,86% -> 100%
Orne	EOL Normandie	Éolien	ENGIE GREEN	Développement	DL Campbell	En Opération	93	100%	99,29% -> 99,86%
Puy-de-Dôme	EOL Sud Est	Éolien	ENGIE GREEN	Développement	DL Campbell	En Opération	82	100%	100%
Haute-Garonne	PV-Occitanie	SOLAIRE CAS	ENGIE GREEN	Exploitation	DL Campbell	En Opération	2,5	100%	99,78%
Oise	France - Ombrière	SOLAIRE Ombrière	ENGIE GREEN	Exploitation	DL Campbell	En Opération	8	100%	99,98% -> 100%
Oise	EOL Nord Est	Éolien	ENGIE GREEN	Développement	DL Campbell	En Opération	99	100%	99,71% -> 100%
Aveyron	PV-Occitanie	SOLAIRE CAS	ENGIE GREEN	Exploitation	DL Campbell	En Opération	3,5	100%	99,99%
Landes	PV-Nouvelle Aquitaine	SOLAIRE CAS	ENGIE GREEN	Exploitation	DL Campbell	En Opération	2,5	100%	100%
Landes	PV-Nouvelle Aquitaine	SOLAIRE CAS	ENGIE GREEN	Exploitation	DL Campbell	En Opération	2,5	100%	100%
Indre-et-Loire	EOL Nord Ouest	Éolien	ENGIE GREEN	Développement	DL Campbell	En Opération	99	100%	97,75% -> 100%
Somme	EOL Nord Est	Éolien	LCV	Exploitation	DL Campbell	En Opération	40	99,99%	86,55% -> 98,21%
Val d'Oise	PV-Nord	SOLAIRE CAS	ENGIE GREEN	Développement	DL Campbell	En Opération	2,5	99,99%	99,34% -> 99,99%
Landes	PV-Nouvelle Aquitaine	SOLAIRE CAS	ENGIE GREEN	Exploitation	DL Campbell	En Opération	2,5	99,99%	90,81% -> 99,99%
Manche	EOL Normandie	Éolien	SAMEOLE	Développement	DL Campbell	En Opération	99	99,99%	99,67% -> 99,99%
Yonne	EOL Sud Est	Éolien	ENGIE GREEN	Exploitation	DL Campbell	En Opération	63	99,99%	98,86% -> 99,57%
Aude	PV-Occitanie	SOLAIRE CAS	ENGIE GREEN	Exploitation	DL Campbell	En Opération	3,5	99,99%	99,94% -> 99,99%
Vaucluse	PV-PACA	SOLAIRE CAS	ENGIE GREEN	Exploitation	DL Campbell	En Opération	4,7	99,99%	98,74% -> 98,74%
Hérault	PV-Occitanie	SOLAIRE CAS	ENGIE GREEN	Exploitation	DL Campbell	En Opération	3	99,99%	99,92% -> 99,92%
Landes	PV-Nouvelle Aquitaine	SOLAIRE CAS	ENGIE GREEN	Exploitation	DL Campbell	En Opération	4,7	99,98%	96,69% -> 99,01%
Loire-Atlantique	EOL Nord Ouest	Éolien	LCV	Exploitation	DL Campbell	En Opération	40	99,98%	96,28% -> 99,86%
Loir-et-Cher	PV-Nord	SOLAIRE CAS	ENGIE GREEN	Développement	DL Campbell	En Operation	2	99,97%	99,47% -> 99,97%

FIGURE 18 – Disponibilité brute et disponibilité valide par campagne

- une instrumentation relativement simple, avec des caractéristiques **GTI** peu variées.
- une bonne disponibilité des mesures sur la campagne (peu de données manquantes).

Après avoir traité manuellement les données de 10 campagnes centralisées, un code unificateur a été développé afin de généraliser le traitement à l'ensemble des campagnes. Ce code permet notamment de gérer automatiquement les différences d'instrumentation entre les campagnes (nombre et type de capteurs disponibles), ainsi que d'appliquer l'algorithme de transposition inverse sur les mesures **GTI**.

Approche pour améliorer l'explicabilité des données (finalement non retenue).

Afin d'optimiser la capacité du modèle à apprendre malgré les différences d'instrumentation **GHI** et **GTI** entre les campagnes, nous avons envisagé de modifier la structure du jeu de données. Plutôt que de conserver la base sous la forme classique — colonnes **GHI**, colonnes **GTI**, colonnes de mesures météo — nous avons tenté une transformation

du dataset de la manière suivante :

- Transformer les colonnes comportant les mesures **GTI** en **GHI** par transposition inverse
- Empiler toutes les colonnes de mesures **GTI** transformées (en **GHI**) ainsi que les colonnes de mesures **GHI** déjà existantes, pour constituer une seule grande colonne **GHI**.
- Pour chaque type de variable météorologique (température, pression, humidité, pluviosité), nous avons empilé toutes les colonnes sous une seule colonne unique.
- En ce qui concerne le dataset des états de validité, nous empilons également ses 12 colonnes d'états de **GHI** et **GTI**, sous une unique colonne des états de **GHI** et **GTI_to_GHI**.

Cependant, même si cette approche est en théorie la plus adaptée pour résoudre le problème de différences de structure entre les campagnes, car elle permet d'éviter la confusion entre les mesures **GHI** et **GTI**, ainsi qu'entre les différents **GTI** entre eux, elle génère en pratique un jeu de données très volumineux, ce qui rend l'entraînement du modèle particulièrement coûteux en temps d'exécution.

6.1 Algorithme Random Forest

A priori, il n'existe pas de choix universel pour le modèle à utiliser. Cependant, étant donné que notre jeu de données est bien structuré sous forme de tableau (des variables explicatives bien définies), il est bien connu que les forêts aléatoires donnent de très bons résultats sur ce type de données tabulaires.

6.1.1 Arbre de décision

L'arbre de décision est en principe un modèle qui prend des décisions en découplant progressivement les données. À chaque étape, il choisit la variable qui sépare le mieux les exemples, puis il divise les données en sous-groupes. Le processus se répète sur chaque sous-groupe jusqu'à ce que l'arbre atteigne un certain niveau de précision ou qu'il n'y ait plus assez d'exemples. À la fin, chaque feuille donne une prédiction. Pour nous, c'est la de la mesure.

Algorithm 1 Construction d'un arbre de décision

1: **Initialisation :**

- Placer tous les exemples à la racine de l'arbre.

2: **À chaque nœud de l'arbre :**

1. Calculer le **gain d'information** pour chaque variable explicative possible.
2. Sélectionner la variable qui maximise ce gain d'information.
3. Diviser les données du nœud en sous-groupes
4. Créer une branche pour chaque sous-groupe.

3: **Appliquer récursivement** les étapes précédentes sur chaque sous-groupe (c'est-à-dire sous-nœud).4: **Critères d'arrêt :**

- Tous les exemples du nœud appartiennent à la même classe (pureté maximale).
- Le gain d'information obtenu est inférieur à un seuil minimal.
- La profondeur maximale de l'arbre est atteinte.
- Le nombre d'exemples dans le noeud est inférieur à un seuil minimal.

5: **Sortie :** Un arbre de décision. Chaque feuille prédit une classe (dans notre cas : mesure valide ou invalide).

6.1.2 Random Forest

L'idée de Random Forest est de construire plusieurs arbres de décision, chacun entraîné sur un échantillon aléatoire du jeu de données, qu'on obtient par une méthode appelée sampling with replacement. Cela signifie que l'on tire aléatoirement des lignes du jeu de données d'origine, en autorisant les répétitions. Chaque arbre est donc entraîné sur une version légèrement différente des données. Lors de la prédiction, les arbres votent, et la classe majoritaire est choisie.

Algorithm 2 Algorithme Random Forest

1. **Initialisation :**
 - Fixer $B = \text{nombre d'arbres à construire}$.
 - Fixer $m = \text{nombre de variables explicatives à sélectionner aléatoirement à chaque nœud}$ ($m < p$, où p est le nombre total de variables).
 2. **Pour** $b = 1$ à B :
 - (a) Générer un échantillon aléatoire avec remise de taille N à partir du jeu de données d'entraînement.
 - (b) Construire un arbre de décision :
 - i. À chaque noeud :
 - Sélectionner aléatoirement m variables explicatives parmi les p disponibles.
 - Calculer le **gain d'information** pour chacune de ces m variables.
 - Choisir la variable qui maximise le gain d'information et séparer les données.
 - ii. Répéter récursivement jusqu'à atteindre un critère d'arrêt (pureté maximale, profondeur maximale, ou nombre minimal d'exemples).
 3. **Prédiction :**
 - Pour une nouvelle observation, chaque arbre prédit une classe.
 - La classe finale est déterminée par **vote majoritaire**.
-

6.1.3 XGBoost

XGBoost est une méthode qui construit plusieurs arbres, les uns après les autres. Chaque nouvel arbre essaie de corriger les erreurs des arbres précédents, en se concentrant sur les exemples mal prédis. Cela permet d'avoir un modèle plus précis. En plus, XGBoost est rapide et empêche que le modèle apprenne trop bien (ce qui peut être mauvais).

Algorithm 3 Algorithme XGBoost

1. **Initialisation :**
 - Placer des poids égaux sur tous les exemples du jeu de données.
 2. **Pour chaque itération** $b = 1$ à B :
 - (a) Entraîner un arbre de décision sur les données pondérées.
 - (b) Identifier les exemples mal classés par les arbres précédents.
 - (c) Augmenter leur poids pour que l'arbre suivant se concentre davantage sur eux.
 3. **Prédiction finale :**
 - Combiner les prédictions de tous les arbres en leur donnant un poids selon leur performance.
-

6.2 Entraînement par validation croisée

La validation croisée est une méthode utilisée pour vérifier que le modèle se généralise bien à des données non vues (méthode plus complexe qu'une validation classique)

Elle consiste à diviser le jeu de données en k sous-ensembles (appelés folds). Le modèle est entraîné sur $k - 1$ de ces sous-ensembles, puis testé sur le fold restant. On répète cette opération k fois, en changeant de fold de validation à chaque itération.

Mathématiquement, pour chaque itération i :

- Le modèle est entraîné sur $k - 1$ folds
- Il est validé sur le i -ème fold

On calcule une métrique de performance (par exemple, l'erreur logarithmique) pour chaque fold, puis on fait la moyenne (ou médiane ou percentiles) des résultats :

$$\begin{aligned} \text{Erreur}_{\text{cv}} &= \frac{1}{k} \sum_{i=1}^k \text{Erreur}_{\text{test}_i} \\ &= \frac{1}{k} \sum_{i=1}^k \left(-\frac{1}{n_{\text{test}}} \sum_{j=1}^{n_{\text{test}}} [y_{\text{test}_j} \log(\hat{y}_{\text{test}_j}) + (1 - y_{\text{test}_j}) \log(1 - \hat{y}_{\text{test}_j})] \right) \end{aligned}$$

où :

- $y_{\text{test}_j} \in \{0, 1\}$ est la vraie classe de l'exemple j ;
- $\hat{y}_{\text{test}_j} \in [0, 1]$ est la probabilité que $y_{\text{test}_j} = 1$;
- n_{test} est le nombre d'exemples dans le jeu de validation.

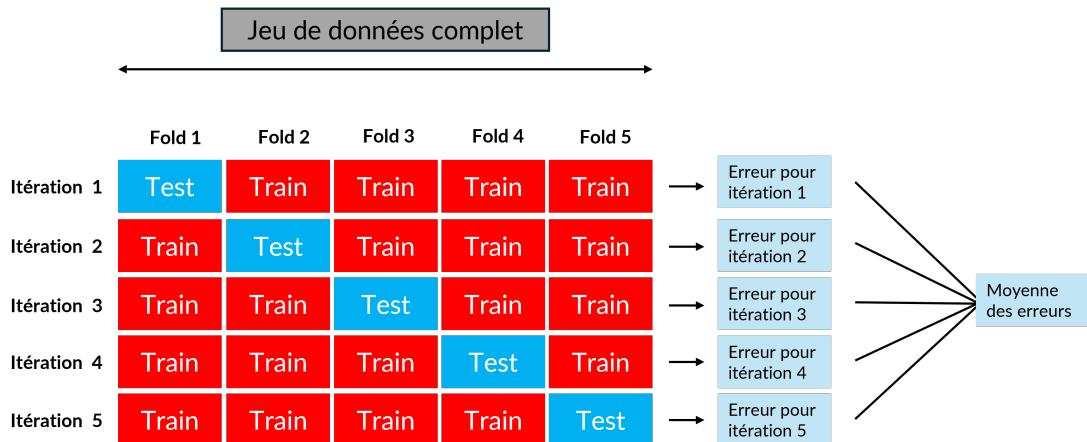


FIGURE 19 – Exemple de validation croisée sur 5 folds

6.3 Scores considérés pour évaluer le modèle

Le modèle utilisé est un modèle de classification binaire, car à partir d'une base de données contenant des mesures d'irradiance et de météo, l'objectif est de prédire si chaque mesure est valide (label 1) ou invalide (label 0).

6.3.1 Précision, rappel et F-score

Pour évaluer les performances d'un modèle de classification binaire, nous utilisons des métriques classiques : la précision, le rappel, et le F-score. Nous expliquons ci-dessous comment ces scores sont calculés.

Matrice de confusion Pour évaluer les performances du modèle de classification, nous utilisons la matrice de confusion.

Celle-ci permet de visualiser sous forme d'une matrice les erreurs de prédiction en comparant les classes prédictes avec les classes réelles. Elle est structurée comme suit :

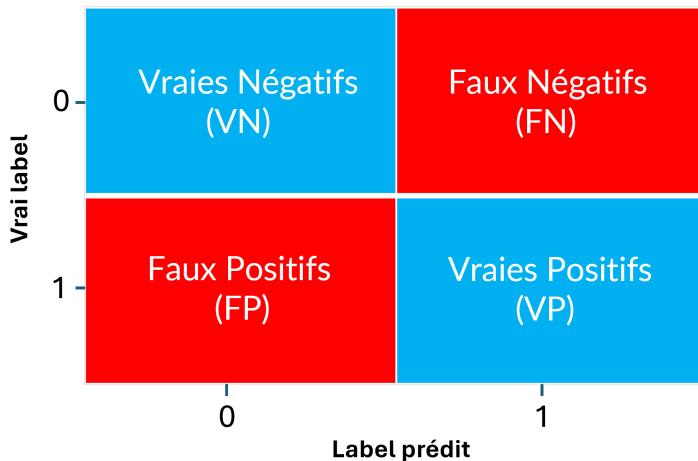


FIGURE 20 – Matrice de confusion

- **Vrais Positifs (VP)** : le modèle prédit la classe positive et la vraie classe est bien positive ;
- **Faux Positifs (FP)** : le modèle prédit la classe positive alors que la vraie classe est négative ;
- **Vraies Négatifs (VN)** : le modèle prédit la classe négative et la vraie classe est bien négative ;
- **Faux Négatifs (FN)** : le modèle prédit la classe négative alors que la vraie classe est positive.

Précision, rappel et F-score Ces trois scores sont dérivés directement de la matrice de confusion. Ils peuvent être calculés pour chaque classe (par exemple : classe « mesure valide » ou classe « mesure invalide »), en considérant successivement l'une comme la classe positive (celle qu'on cherche à détecter) et l'autre comme la classe négative. Cela permet d'évaluer les performances du modèle sur chacune des deux classes.

- **Précision** : parmi toutes les prédictions faites pour une classe, combien sont correctes ?

$$\text{Précision} = \frac{\text{nb_invalidations_de_la_cellule}}{\text{nb_mesures_de_la_cellule}}$$

Autrement dit, à quel point nos prédictions sur cette classe sont-elles bonnes ?

- **Rappel** : parmi tous les exemples réels de cette classe, combien ont été correctement trouvés ?

$$\text{Rappel} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

Autrement dit, à quel point le modèle parvient-il à ne pas rater les exemples de cette classe ?

- **F-score** : il combine la précision et le rappel dans un seul score en prenant leur moyenne harmonique.

$$F\text{-score} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

Dans tout ce qui suit, nous désignerons par prec_1 et rappel_1 la précision et le rappel calculés pour la classe « mesure valide », et par prec_0 et rappel_0 ceux calculés pour la classe « mesure invalide ».

Pour interpréter plus concrètement les valeurs de précision et de rappel, prenons un exemple :

Exemple Pour donner un ordre d'idée, sur une année de mesures avec un pas de 10 minutes (environ 57 000 valeurs), un rappel de 0.90 signifie que si une semaine complète de mesures est invalide (7 jours \times 24 heures \times 6 mesures par heure), le modèle en détecte correctement environ 90%, soit 6.3 jours. Cela montre aussi que la différence entre 0.90 et 0.91, même si elle paraît faible, correspond déjà à plus d'une demi-journée supplémentaire de mesures détectées sur l'année.

7 Bilan des résultats

Dans cette section, nous présentons le bilan des résultats obtenus avec l'apprentissage sur les mesures d'irradiance. Au cours du stage, plusieurs approches ont été testées, et les performances ont été progressivement améliorées grâce à l'ajout de variables explicatives pertinentes.

7.1 Première approche non retenue

7.1.1 Empilement des colonnes d'irradiance et de météo

Dans un premier temps, nous avons testé une approche consistant à empiler toutes les mesures d'irradiance dans une seule colonne **GHI**, ceci après avoir transformé mesures **GTI**, et à regrouper les mesures météo par type dans des colonnes distinctes. Cette méthode s'est révélée peu avantageuse, car coûteuse en temps de calcul CPU (environ 720 minutes) et en mémoire.

The diagram illustrates the process of stacking (empiling) columns from two separate datasets into a single dataset. An arrow labeled "Empiler" points from the left table to the right table.

Left Table (Initial Data):

Dates	GHI	GTI_to_GHI	Température 1	Température 2	Humidité
10h	300	301.2	10	20	100
10h10	325	325.3	12	22	100
10h20	350	352.1	14	24	200
10h30	375	375.9	16	26	200

Right Table (Stacked Data):

Dates	GHI	Temps	Hums
10h00	300	10	100
10h10	325	12	100
10h20	350	14	200
10h30	375	16	200
10h00	301.2	20	100
10h10	325.3	22	100
10h20	352.1	24	200
10h30	375.9	26	200

FIGURE 21 – Exemple d’empilement des colonnes d’irradiance et de météo

Avec cette approche, nous avons pu réaliser une validation croisée avec le modèle XGBoost sur 36 campagnes centralisées, ce qui a généré un dataset d’environ 90 millions de lignes. Les résultats obtenus étaient :

$$\text{Précision}_1 \approx 0,998, \quad \text{Rappel}_1 \approx 0,999, \quad \text{Précision}_0 \approx 0,719, \quad \text{Rappel}_0 \approx 0,561$$

Cette précision et ce rappel pour la classe 0 sont peu satisfaisants, puisque notre objectif est d’obtenir des scores aussi proches que possible de 1.

7.2 Approche retenue

Avec l’approche qui consiste à conserver plusieurs colonnes de **GHI** et plusieurs colonnes de **GTI**, nous pouvons traiter directement les bases de données des 110 campagnes centralisées. En effet, la base de données générée est significativement moins lourde que dans le cas où les mesures seraient empilées, et le modèle parvient ainsi à apprendre plus facilement. Nous avons réalisé une validation croisée sur l’ensemble des 110 campagnes centralisées avec XGBoost, en ajoutant progressivement des variables explicatives pour améliorer les scores.

Dans les tests ci-dessous, nous avons conservé uniquement les mesures enregistrées en journée. Celles prises pendant la nuit n’ont pas été retenues, car il n’y a aucun intérêt à prédire leur validité.

Test 1 : Mesures d’irradiances et météo uniquement (aucune feature supplémentaire) :

$$\text{Précision}_0 = 0.884, \quad \text{Rappel}_0 = 0.698, \quad F_1 = 0.78$$

Test 2 : Test 1 + écarts **GHI** et **GTI** :

$$\text{Précision}_0 = 0.907, \quad \text{Rappel}_0 = 0.779, \quad F_1 = 0.839$$

Test 3 : Test 2 + colonne **GHI_clearsky** :

$$\text{Précision}_0 = 0.922, \quad \text{Rappel}_0 = 0.794, \quad F_1 = 0.853$$

Test 4 : Test 3 + indices de beau temps et fraction diffuse :

$$\text{Précision}_0 = 0.924, \quad \text{Rappel}_0 = 0.797, \quad F_1 = 0.855$$

Test 5 : Test 4 + cumuls d'irradiances et cumuls des mesures SolEye :

$$\text{Précision}_0 = 0.988, \quad \text{Rappel}_0 = 0.976, \quad F_1 = 0.982$$

Test 6 : Test 5 + écarts entre **GHI** et **GHI_clearsky** + mesures **GTI** transformées en **GHI** :

$$\text{Précision}_0 = 0.987, \quad \text{Rappel}_0 = 0.974, \quad F_1 = 0.98$$

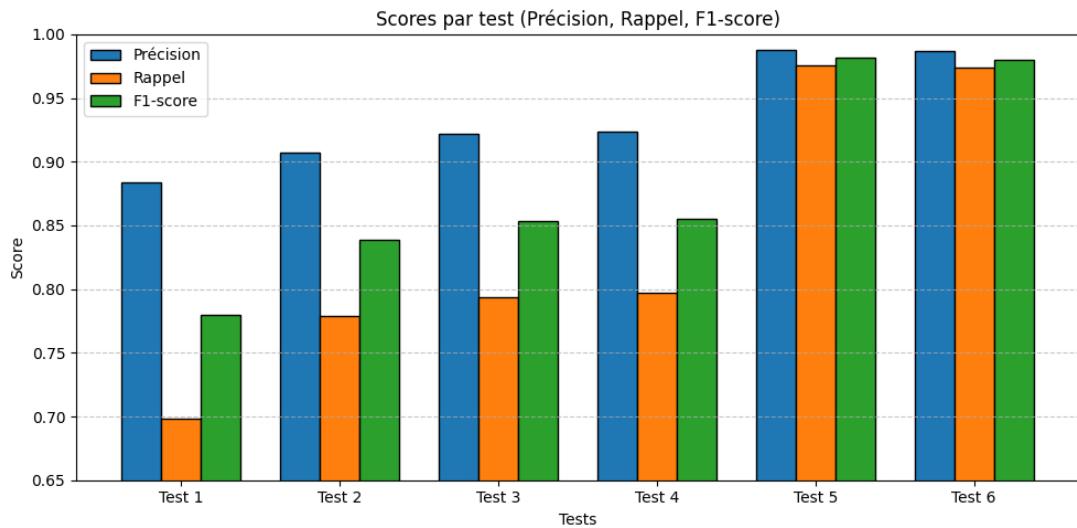


FIGURE 22 – Scores de chaque test

Ces tests montrent qu'en ajoutant à chaque étape de nouvelles variables explicatives, le modèle parvient à détecter de plus en plus de motifs et, par conséquent, à mieux identifier les mesures invalides. Cela se reflète directement dans l'amélioration des scores obtenus à chaque test.

7.3 Deuxième approche non retenue

La deuxième approche non retenue consiste à remplacer dans notre tableau de mesures d'irradiance de **GHI** et **GTI**, les colonnes **GTI** par **GTI_to_GHI**.

Le calcul des **GTI_to_GHI** sur toutes les mesures **GTI** de l'ensemble des campagnes, puis le remplacement des colonnes **GTI** par ces nouvelles valeurs, n'a pas donné les résultats attendus :

$$\text{Précision}_0 = 0.87, \quad \text{Rappel}_0 = 0.695, \quad \text{F-score} \approx 0.77$$

En théorie, cette transformation aurait dû améliorer les performances. Cependant, les résultats montrent l'inverse, probablement pour la raison suivante : en observant la distribution du nombre de campagnes en fonction de (**GHI**, **GTI**) (voir figure 2), on remarque qu'un nombre important de campagnes présentent un nombre d'instruments **GTI** supérieur au **GHI**. Ainsi, transformer systématiquement les valeurs **GTI** en **GHI** pourrait entraîner une perte d'information importante.

7.4 Cas d'application réel

Dans cette section, nous appliquons l'approche retenue, qui consiste à conserver les colonnes de **GHI** et de **GTI**. Pour tester le modèle dans un cas réaliste, nous l'entraînons sur l'ensemble de la base des 110 campagnes centralisées (hors janvier 2025), puis nous l'évaluons sur ce mois récent, janvier 2025, pour toutes les campagnes.

Le modèle XGBoost (`n_trees = 400, max_depth = 22, learning_rate = 0.1`) obtient, pour la classe 0, une précision de 0,33 et un rappel de 0,28, tandis que pour la classe 1, la précision et le rappel atteignent 0,99.

Ces résultats sont peu satisfaisants pour la classe 0. Pour analyser cela de plus près, nous examinons les ratios $\frac{\text{TN}}{\text{total zéros}}$ et $\frac{\text{FN}}{\text{total zéros}}$ par campagne, en ne considérant que les campagnes contenant des exemples de classe 0.

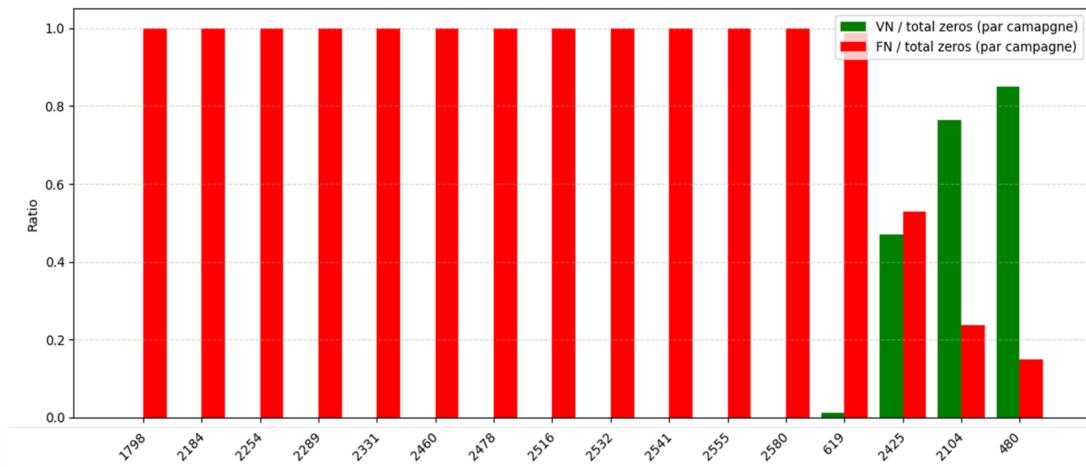


FIGURE 23 – Ratios VN et FN sur le nombre de zéros par campagne

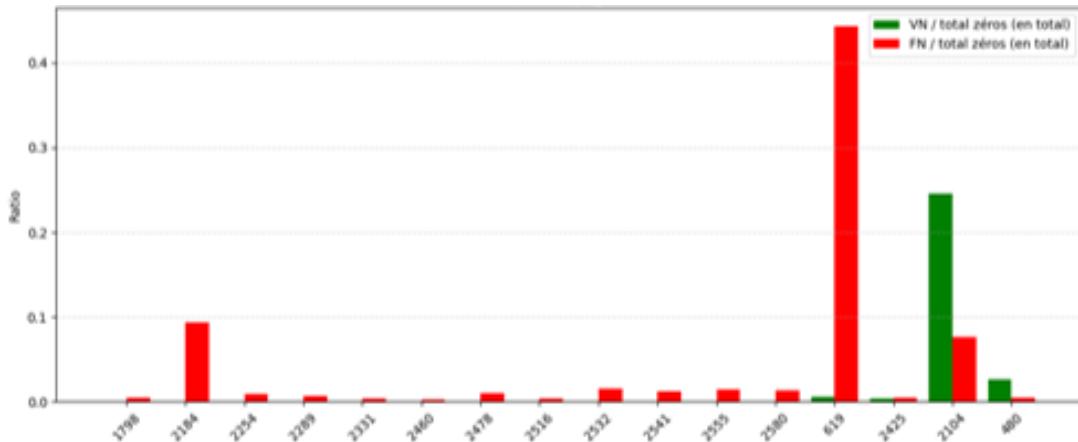


FIGURE 24 – Ratios VN et FN sur le nombre de zéros sur toutes les campagnes

Ces graphiques montrent que la campagne 619 est la principale source d'erreurs sur la classe 0 : près de 45% des classifications totales sur la classe 0 (mesures invalides) sont mauvaises (prédictions en classe 1) et proviennent de cette campagne (graphe 2), avec très peu de 0 bien prédits (graphe 1). À l'inverse, environ 25% des classifications totales sur la classe 0 sont bonnes (0 prédit correctement) sur la campagne 2104 (graphe 2) avec 77% de 0 bien prédits sur cette campagne (graphe 1).

Pour la campagne 2184, qui présente un ratio $\frac{FN}{\text{total zéros}}$ relativement élevé, j'ai appris qu'une intervention a eu lieu en mars 2025 pour changer l'emplacement des pyromètres lors de leur remise en place, ce qui pourrait expliquer ce résultat.

7.4.1 Analyse par journée - Campagne 619

En examinant de plus près la campagne **619**, des invalidations apparaissent sur **cinq jours**. Sur **quatre jours**, elles couvrent toute la journée de **8h00 à 16h00**, et sur **un jour**, elles s'étendent de **11h40 à 13h10** et concernent **six mesures GTI**. Le modèle parvient à détecter correctement cette courte période :

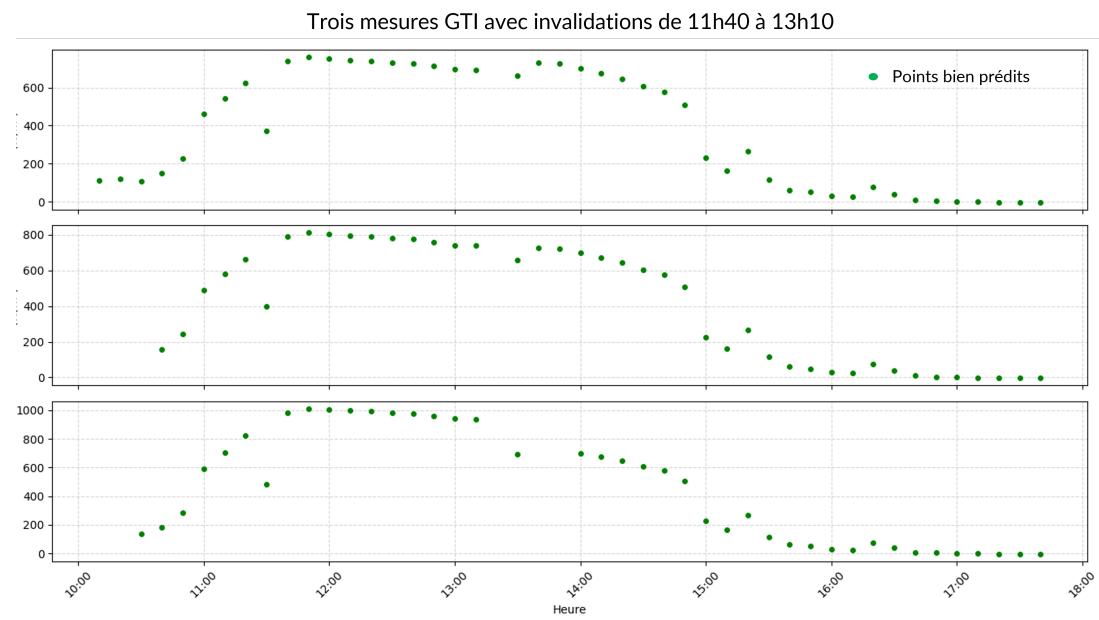


FIGURE 25 – Invalidations de 11h40 à 13h10 bien prédites par le modèle.

En revanche, le modèle échoue sur les longues périodes d'invalidation :

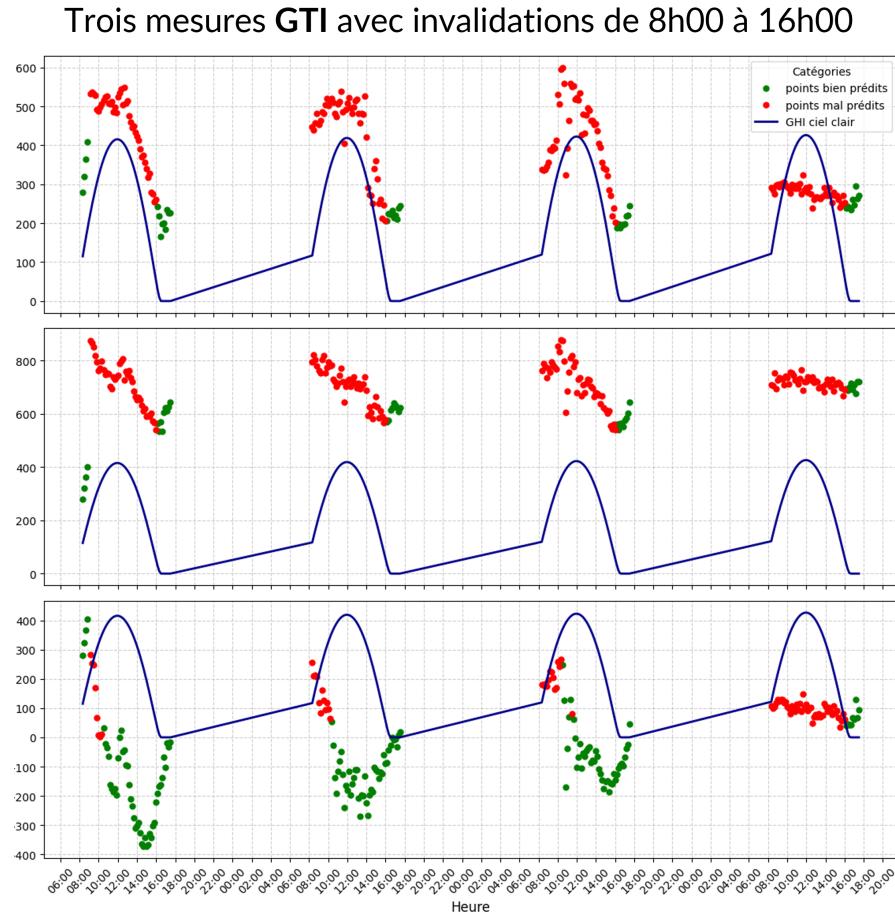


FIGURE 26 – Invalidations de 8h00 à 16h00 mal prédites par le modèle.

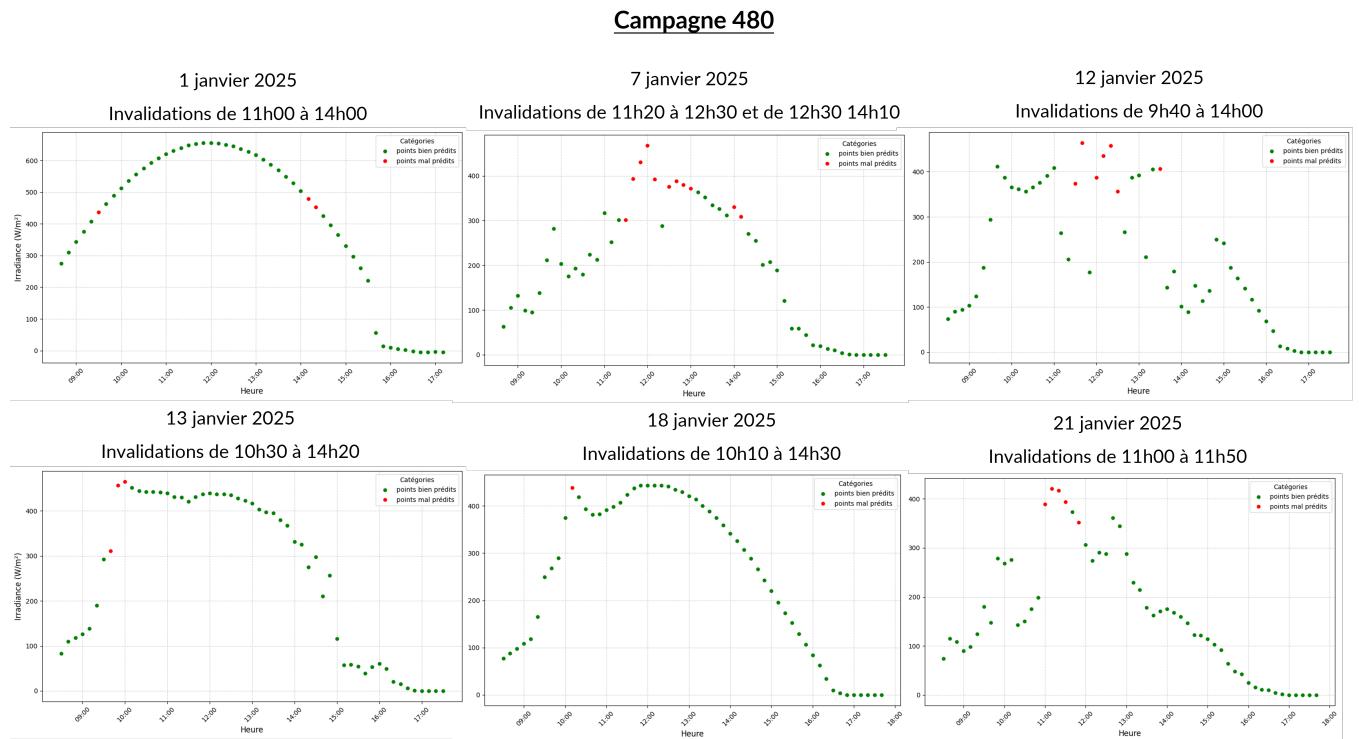
Nous observons que, même lorsque le modèle aurait de bonnes raisons de marquer ces mesures comme invalides (écart **GHI** très importants, irradiance proche de zéro en pleine journée), il ne le fait pas toujours. Cela nous ramène à un problème déjà identifié : la base de données de validation n'est pas toujours fiable. À certaines dates, de gros écarts ne sont pas marqués, ou des irradiances aberrantes ne sont pas signalées, ce qui induit le modèle en erreur. Toutefois, le modèle arrive quand même à détecter les valeurs négatives et quelques autres positives.

Ces mesures sont aberrantes et pointent clairement sur des problèmes dans les capteurs. Comme ces situations sont très atypiques, le modèle n'a pas été capable de les détecter.

Pour y remédier, j'ai tenté d'ajouter les colonnes **GTI_to_GHI**, ainsi que les écarts entre **GHI_clearsky** et **GHI**. Ces ajouts n'ont toutefois pas apporté d'amélioration significative.

7.4.2 Analyse par journée - Campagne 480

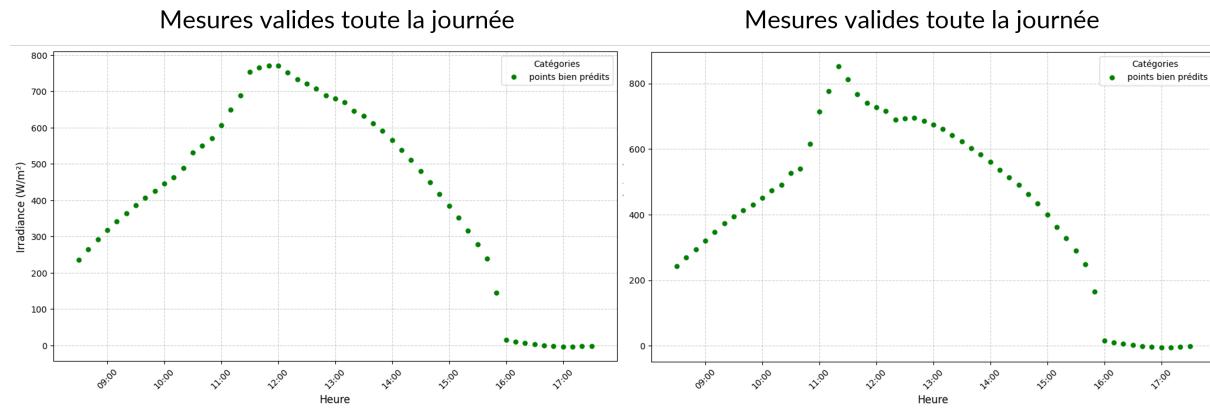
Sur la figure ci-dessous, nous présentons le comportement du modèle sur chaque journée présentant des périodes d'invalidation en janvier 2025. Les points en vert correspondent aux prédictions correctes (par exemple : classe 1 prédite 1, ou classe 0 prédite 0), tandis que les points en rouge indiquent les erreurs du modèle (par exemple : classe 1 prédite 0, ou classe 0 prédite 1).



Le 1er, 12, 13 et 18 janvier, le modèle détecte bien les longues périodes d'invalidation, mais en manque quelques-unes. C'est attendu, car les techniciens Mesures pourraient

prendre de la marge lors des invalidations. Le 7 et le 21 janvier, les invalidations courtes ne sont pas entièrement détectées. Néanmoins, le modèle repère qu'une anomalie existe.

Ces deux graphiques illustrent des journées où le modèle a correctement identifié que toutes les mesures étaient valides.



En résumé, sur la campagne 480, le modèle parvient à détecter les petites que les longues périodes d'invalidation. Même s'il ne les identifie pas toutes, il en repère au moins certaines, ce qui permet déjà de mettre en évidence la présence d'une anomalie.

Toutefois, il est important de rappeler que nous ne pouvons pas avoir une vérité absolue : le technicien prend volontairement de la marge car mieux vaut trop invalider que laisser passer des mesures fausses. Par conséquent, il ne faut pas s'attendre à un accord parfait entre notre modèle et les décisions du technicien.

7.4.3 Analyse des prédictions du modèle par colonne GHI/GTI

Après avoir calculé les scores de précision et de rappel sur l'échantillon de test du mois de janvier, nous avons souhaité approfondir cette évaluation en analysant plus finement la performance du modèle sur chaque colonne GHI/GTI, afin de comprendre comment il effectue ses prédictions.

Pour cela, nous présentons ci-dessous les matrices de confusion par colonne, à la fois en valeurs brutes et en pourcentages. Ces résultats peuvent sembler, à première vue, décevants, mais nous expliquerons par la suite quels éléments il convient réellement d'interpréter et de prioriser dans cette analyse.

TABLE 4 – Matrice de confusion par colonne

Colonne	FN	VN	FP	VP
GHI_1	244	0	105	178994
GHI_2	517	8	29	178789
GHI_3	249	0	0	179094
GHI_4	0	0	0	179343
GHI_5	0	0	0	179343
GTI_1	18	181	332	178812
GTI_2	27	189	264	178863
GTI_3	13	164	255	178911
GTI_4	0	0	198	179145
GTI_5	0	0	193	179150
GTI_6	0	0	0	179343
GTI_7	0	0	5	179338

TABLE 5 – Précisions par colonne pour les classes 0 et 1

Colonne	Précision ₀	Précision ₁
GHI_1	0.0000	0.9994
GHI_2	0.0152	0.9998
GHI_3	0.0000	1.0000
GHI_4	–	1.0000
GHI_5	–	1.0000
GTI_1	0.904	0.9982
GTI_2	0.875	0.9985
GTI_3	0.926	0.9986
GTI_4	–	0.9989
GTI_5	–	0.9989
GTI_6	–	1.0000
GTI_7	–	1.0000

Egalement, pour visualiser à quel point nous ratons chaque classe, visualisons les rappels par classe par colonne :

TABLE 6 – Rappels par colonne pour les classes 0 et 1

Colonne	Rappel ₀	Rappel ₁
GHI_1	0.0000	0.9994
GHI_2	0.2162	0.9971
GHI_3	0	0.9986
GHI_4	0	1
GHI_5	0	1
GTI_1	0.3521	0.9999
GTI_2	0.4175	0.9998
GTI_3	0.3914	0.9999
GTI_4	0	1
GTI_5	0	1
GTI_6	0	1
GTI_7	0	1

Interprétation des faibles précisions et rappels pour GHI Pour les GHI, la précision est faible : en effet, le modèle n'a pas bien prédit la classe 0. Cela peut s'expliquer par le fait que, récemment, les techniciens portent moins d'attention aux mesures GHI. Ainsi, certaines invalidations qui auraient dû être effectuées conformément à la base d'entraînement ne l'ont pas été.

Par ailleurs, le rappel est également faible : le modèle n'a trouvé presque aucun zéro. Cela est probablement dû au fait que les causes des invalidations réalisées en janvier sur les GHI ne sont pas liées aux variables explicatives fournies au modèle, mais plutôt à des problèmes techniques ou logistiques associés aux capteurs.

En effet, nous remarquons que le modèle performe nettement mieux sur les mesures GTI que sur les mesures GHI. Pour cela, examinons le nombre total d'invalidations par colonne GHI/GTI durant le mois de janvier

TABLE 7 – Nombre d'invalidations et de validations par colonne

Colonne	Nombre de invalidations	Nombre de validations
GHI_1	105	179238
GHI_2	37	179306
GHI_3	0	179343
GHI_4	0	179343
GHI_5	0	179343
GTI_1	513	178830
GTI_2	453	178890
GTI_3	419	178924
GTI_4	198	179145
GTI_5	193	179150
GTI_6	0	179343
GTI_7	5	179338

Ceci montre que sur le mois de janvier, il y a nettement plus d'invalidations sur les GTI que sur les GHI. Cela correspond à l'information reçue : récemment, les techniciens en charge n'utilisent plus les mesures GHI et portent donc moins d'attention à les invalider. Pour le voir plus clairement, on peut visualiser le nombre d'invalidation par type de mesure **GHI** et **GTI** :

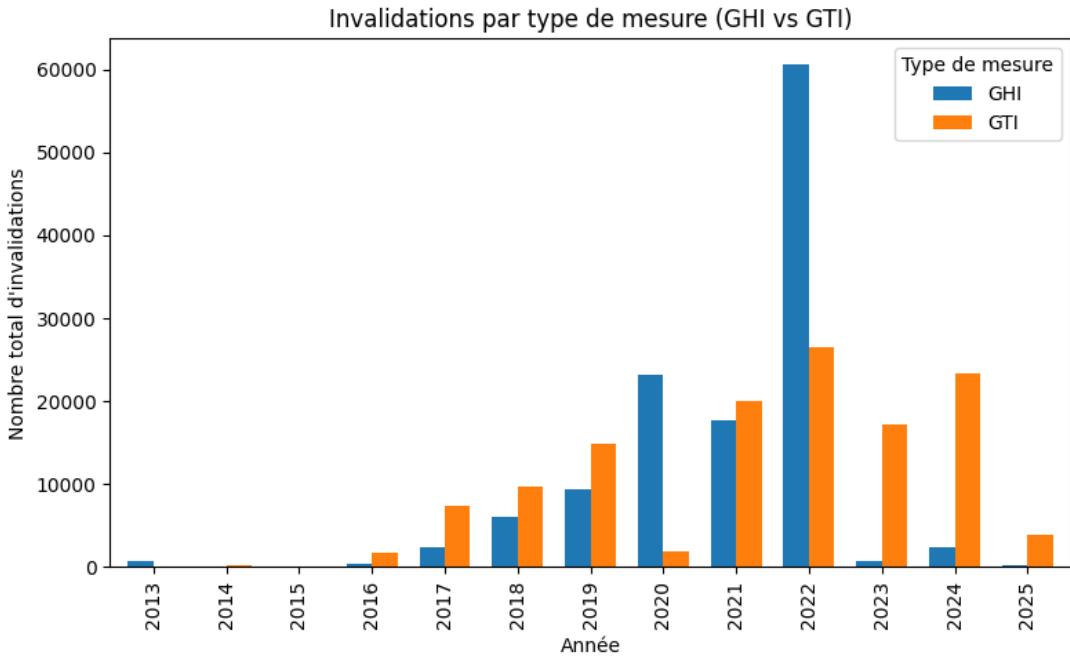


FIGURE 27 – Nombre d’invalidations par type de mesure

On voit bien qu’entre 2023 et 2025, les invalidations sur GHI sont beaucoup moins nombreuses que sur GTI. Comme le modèle a été entraîné sur toute la base couvrant plusieurs années, et que les personnes et méthodes d’invalidation changent, cela impacte les prédictions récentes sur les GHI.

7.5 Test du modèle sur des mois plus anciens

Nous nous intéressons maintenant à l’entraînement sur l’ensemble de la base de données et au test du modèle sur des données plus anciennes. L’objectif est d’observer comment le modèle se comporte face à la manière, relativement ancienne, dont les invalidations étaient réalisées.

7.5.1 Mois de novembre 2021

Ce mois a été choisi car il contient un nombre important d’invalidations, aussi bien pour les mesures GHI que pour les mesures GTI (voir figure 27). Ci-dessous, nous présentons la précision et le rappel par colonne pour ce mois.

TABLE 8 – Précisions par colonne pour les classes 0 et 1

Colonne	Précision ₀	Précision ₁
GHI_1	0.985	0.99
GHI_2	0	0.99
GHI_3	0.953	0.98
GHI_4	–	1
GHI_5	–	1
GTI_1	1.0	1
GTI_2	–	1
GTI_3	–	1
GTI_4	–	1
GTI_5	–	1
GTI_6	–	1
GTI_7	–	1

TABLE 9 – Rappels par colonne pour les classes 0 et 1

Colonne	Rappel ₀	Rappel ₁
GHI_1	0.187	0.99
GHI_2	0	1
GHI_3	0.096	0.99
GHI_4	–	1
GHI_5	–	1
GTI_1	0.171	1
GTI_2	–	1
GTI_3	–	1
GTI_4	–	1
GTI_5	–	1
GTI_6	–	1
GTI_7	–	1

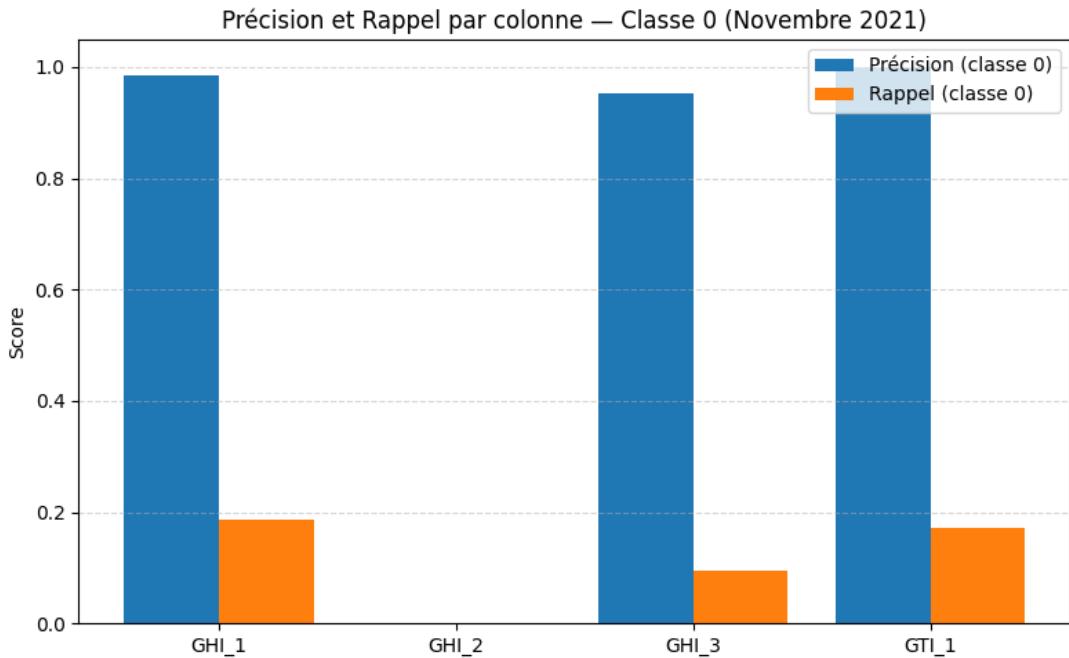


FIGURE 28 – Précision et rappel par colonne (classe 0)

Une précision élevée combinée avec un rappel faible sur la classe 0 pour les mois anciens indique que le modèle est très sûr de ses prédictions lorsqu'il annonce un 0 (il ne le prédit que lorsqu'il a une forte confiance que c'est bien un 0), mais qu'il manque la majorité des vrais 0.

Ce faible rappel est probablement lié à l'ancienne manière d'invalider (avant 2022 inclus). À cette époque, les invalidations étaient plus fréquentes, avec une marge plus large, et pouvaient se propager sur plusieurs jours consécutifs. Or, le modèle ne prend pas en compte les invalidations sur plusieurs jours.

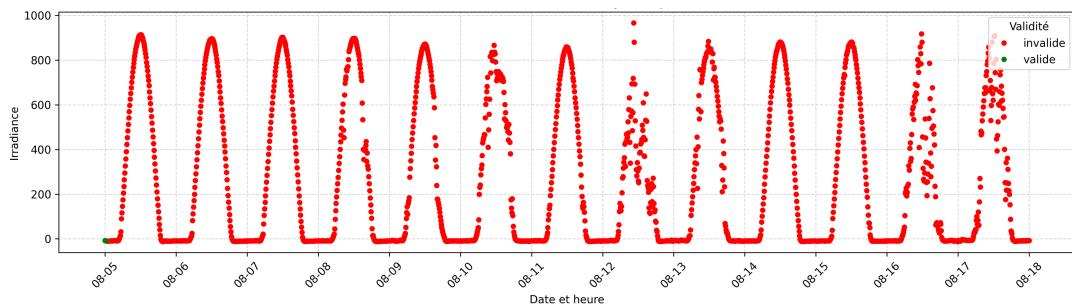


FIGURE 29 – Invalidations effectuées sur plusieurs jours consécutifs.

Concrètement, ces invalidations prolongées étaient faites par les techniciens mesure lorsqu'ils observaient que les capteurs enregistraient une valeur négative de nuit inférieure à -7. Dès lors, toute la journée était invalidée.

7.5.2 Mois de février 2022

Ce mois a été choisi car il comporte lui aussi un grand nombre d'invalidations pour les mesures GHI et GTI, et il fait partie des mois présentant le plus d'invalidations dans l'année (voir figure 27). Les tableaux suivants présentent la précision et le rappel par colonne pour ce mois.

TABLE 10 – Précisions par colonne pour les classes 0 et 1

Colonne	Précision ₀	Précision ₁
GHI_1	0,997	0,97
GHI_2	1,0	0,97
GHI_3	0,999	0,97
GHI_4	–	1
GHI_5	–	1
GTI_1	0,994	0,99
GTI_2	–	1
GTI_3	1,0	0,99
GTI_4	–	1
GTI_5	–	1
GTI_6	–	1
GTI_7	–	1

TABLE 11 – Rappels par colonne pour les classes 0 et 1

Colonne	Rappel ₀	Rappel ₁
GHI_1	0,542	0,99
GHI_2	0,537	1
GHI_3	0,515	0,99
GHI_4	–	0,99
GHI_5	–	1
GTI_1	0,937	1
GTI_2	–	1
GTI_3	0,963	1
GTI_4	–	1
GTI_5	–	1
GTI_6	–	1
GTI_7	–	1

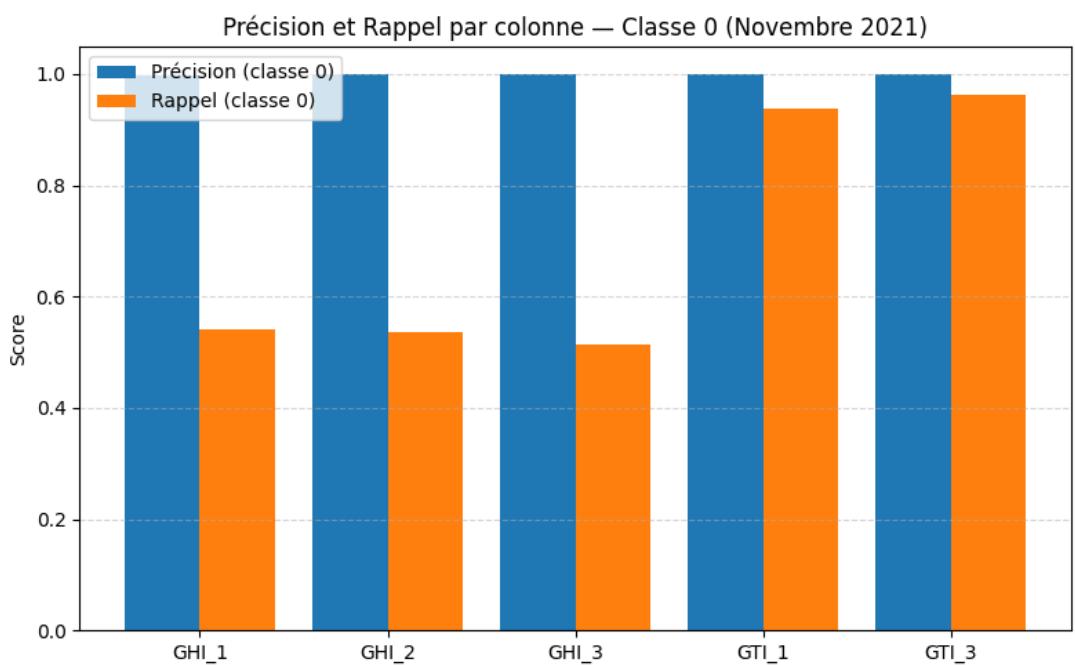


FIGURE 30 – Précision et rappel par colonne (classe 0)

8 Conclusion

Bilan sur le fonctionnement de l'approche de Machine Learning

En conclusion, les résultats obtenus en validation croisée ont montré que les variables explicatives sélectionnées (écart entre mesures des pyranomètres, mesures satellites, cumuls d'irradiance, GHI théorique, indices, etc.) interagissent de manière pertinente avec le modèle XGBoost. Leur intégration s'est révélée très bénéfique pour la qualité des prédictions.

En revanche, même avant le stage, il n'était pas certain à 100% que l'approche de Machine Learning fonctionnerait, notamment en raison du fort déséquilibre entre les mesures invalides et celles qui sont valides. Plus important encore, la base de données de validations/invalidations n'est pas toujours parfaitement correcte, ce qui est inévitable et connu. Pour espérer obtenir de meilleurs résultats, il est essentiel que le processus de validation et la manière d'invalider soient plus clairs, plus uniformes, mieux documentés et appliqués de façon plus régulière.

Par ailleurs, l'approche qui consiste à empiler les colonnes n'a pas été retenue : compte tenu de ses performances et du coût important en calcul et en mémoire, nous avons choisi de l'abandonner et de conserver le format de tableau avec des colonnes séparées pour **GHI** et **GTI**.

Le test du modèle sur un mois récent a donné des résultats peu satisfaisants pour la majorité des campagnes, mais a bien fonctionné pour un petit nombre d'entre elles, où le modèle était en accord avec la manière d'invalider de l'utilisateur. De même, sur des mois plus anciens, le modèle a produit de très bonnes prédictions, mais a manqué dans plusieurs cas un grand nombre de mesures invalides, en partie à cause des invalidations portant sur plusieurs journées consécutives.

Retour sur l'expérience de stage

En tant qu'ancien diplômé d'un master de mathématiques fondamentales et, après ce stage, futur diplômé d'un master de mathématiques appliquées, il s'agissait de ma première expérience en entreprise. Ce fut une expérience très enrichissante, notamment grâce à l'opportunité de travailler au sein d'une équipe de Data Science & Modélisation particulièrement compétente, dont fait partie mon encadrant. J'ai pu évoluer dans cet environnement stimulant et collaboratif, et j'espère mettre à profit les connaissances acquises lors de ce stage dans de futures opportunités, idéalement dans le domaine des énergies renouvelables, un domaine qui est désormais devenu une véritable passion.

Annexe : Entropie et gain d'information

Définition de l'entropie

L'entropie mesure l'impureté d'un ensemble d'exemples de la base de données. Soit p_1 la fraction d'exemples de la *classe positive* (par exp. label 1), et $p_0 = 1 - p_1$ celle de la *classe négative*.

L'entropie H est définie par :

$$H(p_1) = -p_1 \log_2(p_1) - p_0 \log_2(p_0)$$

avec la convention $0 \log_2(0) = 0$.

Elle prend sa valeur maximale $H = 1$ lorsque $p_1 = 0.5$ (ensemble parfaitement équilibré) et vaut 0 lorsque $p_1 = 0$ ou $p_1 = 1$ (ensemble pur).

Définition du gain d'information

Soit un noeud d'arbre de décision contenant un ensemble S d'exemples. On considère une partition de S en deux sous-ensembles S^{gauche} et S^{droite} en fonction d'une caractéristique (feature).

On note :

$$\begin{aligned} p_1^{\text{root}} &: \text{proportion de positifs dans } S, \\ p_1^{\text{gauche}} &: \text{proportion de positifs dans } S^{\text{gauche}}, \\ p_1^{\text{droite}} &: \text{proportion de positifs dans } S^{\text{droite}}, \\ w^{\text{gauche}} &= \frac{|S^{\text{gauche}}|}{|S|}, \quad w^{\text{droite}} = \frac{|S^{\text{droite}}|}{|S|}. \end{aligned}$$

Le **gain d'information** associé à cette division est :

$$\text{IG} = H(p_1^{\text{root}}) - \left[w^{\text{gauche}} H(p_1^{\text{gauche}}) + w^{\text{droite}} H(p_1^{\text{droite}}) \right]$$

Il mesure la réduction de l'impureté (diminution de l'entropie) obtenue en réalisant le split. Lors de l'apprentissage d'un arbre de décision, on choisit le split donnant le IG maximal.

Références

- [1] Kevin S. ANDERSON et al. “pvlib python : 2023 project update”. In : *Journal of Open Source Software* 8.92 (2023), p. 5994. DOI : [10.21105/joss.05994](https://doi.org/10.21105/joss.05994). URL : <https://doi.org/10.21105/joss.05994>.
- [2] Anton DRIESSE, Adam R. JENSEN et Richard PEREZ. “A continuous form of the Perez diffuse sky model for forward and reverse transposition”. In : *Solar Energy* 267 (2024), p. 112093. ISSN : 0038-092X. DOI : <https://doi.org/10.1016/j.solener.2023.112093>. URL : <https://www.sciencedirect.com/science/article/pii/S0038092X23007272>.