

# RAPPORT DE STAGE INGENIEUR



Ecole nationale d'ingénieur de  
Gabes

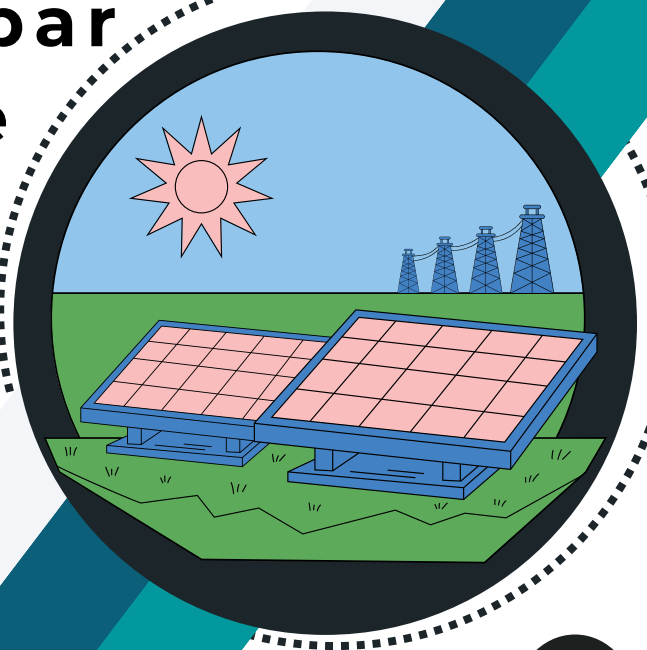


Elaboré par : Ghada Saad

Spécialité: génie électrique  
automatique GEA 3

Projet: Prédiction de  
l'énergie solaire par  
l'apprentissage  
automatique

Encadré par :  
Mohamed Hedi  
Riahi



# Remerciement

A travers ces lignes, je tiens à remercier tous les employés, cadres et dirigeants de l'Esprit pour leur assistance et encadrement durant ce stage.

En préambule à ce projet, je souhaite exprimer ma profonde gratitude envers les encadrants de ESPRIT : Madame Amal HDHILI et Monsieur Mohamed Hedi RIAHI . Grace à leurs précieux conseils et leur expertise, notre stage a été plus enrichissant et motivant. Ainsi que tout les responsables qui m'ont supporté durant cette période .

# SOMMAIRE

<b>Introduction générale</b>	<b>5</b>
<b>Chapitre 1: Etude préliminaires</b>	<b>6</b>
1. Introduction	6
2. Présentation de l'organisation d'accueil Esprit	6
3. Contexte du projet:	8
<b>Chapitre 2: compréhension des données</b>	<b>12</b>
1. Introduction	12
2. Les énergies renouvelables	12
3. Modélisation des données	16
4. Aperçu des données	20
5. Conclusion	25
<b>Chapitre 3: préparation des données</b>	<b>26</b>
1. Introduction	26
2. Nettoyage des données	26
<b>Chapitre 4: Modélisation avec série temporelle</b>	<b>30</b>
1. Introduction	30
2. Modélisation	30
3. Conclusion	38
<b>Conclusion générale</b>	<b>39</b>
<b>Bibliographie</b>	<b>40</b>

# Liste des figures

<b>Figure 1: Logo Esprit</b>	<b>6</b>
<b>Figure 2: Esprit</b>	<b>6</b>
<b>Figure 3: Les applications et les branches de l'intelligence artificielle.</b>	<b>11</b>
<b>Figure 4: logo python</b>	<b>11</b>
<b>Figure 5 :logo Microsoft excel</b>	<b>11</b>
<b>Figure 6:Types des énergies</b>	<b>12</b>
<b>Figure 7 : consommation énergétique et production électrique en 2020</b>	<b>13</b>
<b>Figure 8: Energie renouvelable et énergie non renouvelable</b>	<b>14</b>
<b>Figure 9: panneau photovoltaïque</b>	<b>14</b>
<b>Figure 10 : partition de la capacité d'électricité 2010-2027</b>	<b>15</b>
<b>Figure 11: schéma explicatif du CRISP DM</b>	<b>16</b>
<b>Figure 12: capacité d' électricité renouvelable installé par secteur</b>	<b>19</b>
<b>Figure 13: Production de l'énergie solaire a Sfax S</b>	<b>20</b>
<b>Figure 14: Capture des colonnes</b>	<b>20</b>
<b>Figure 15: Description du colonne "Prod"</b>	<b>22</b>
<b>FIGURE 16: BOITE À MOUSTACHES</b>	
<b>23FIGURE 17: CONTRÔLE DE LA NORMALITÉ</b>	<b>24</b>
<b>FIGURE 18: INFORMATIONS SUR LES VALEURS MANQUANTES</b>	<b>25</b>
<b>FIGURE 19: LES VALEURS ABERRANTES</b>	<b>26</b>
<b>FIGURE 20: TRANSFORMATION DES VALEURS ABERRANTES EN NA</b>	<b>27</b>
<b>FIGURE 21: L'ENSEMBLE DES VALEURS MANQUANTES TOTALE</b>	<b>27</b>
<b>FIGURE 22: LES DONNÉES APRÈS IMPUTATION</b>	<b>28</b>
<b>FIGURE 23: PAS DE VALEURS MANQUANTES APRÈS IMPUTATION</b>	<b>29</b>
<b>FIGURE 24: LES VALEURS DE PROD JOURNALIÈREMENT</b>	<b>31</b>
<b>FIGURE 25: CRÉATION DU SÉRIE TEMPORELLE</b>	<b>31</b>
<b>FIGURE 26:GRAPHE REPRÉSENTANT LA SÉRIE TEMPORELLE</b>	<b>32</b>
<b>FIGURE 27:TEST DE STATIONNARITÉ DU SÉRIE TEMPORELLE</b>	<b>33</b>
<b>FIGURE 28:CHOIX DU MEILLEUR AIC POUR UN MEILLEUR MODÈLE</b>	<b>34</b>
<b>FIGURE 29:SYNTHÈSE DU MODÈLE</b>	<b>34</b>
<b>FIGURE 30:REPRÉSENTATION GRAPHIQUE DU MODÈLE ARMA</b>	<b>35</b>
<b>FIGURE 31:RÉSIDUS</b>	<b>36</b>
<b>FIGURE 32: VÉRIFICATION DE STATIONNARITÉ</b>	<b>36</b>
<b>FIGURE 33: MOYENNE DES RÉSIDUS</b>	<b>36</b>
<b>FIGURE 34: AUTOCORRELATION DES RÉSIDUS</b>	<b>37</b>

# Introduction générale

L'énergie solaire est devenue l'une des sources d'énergie renouvelable les plus prometteuses pour répondre aux besoins croissants en électricité dans le monde. Cependant, la variabilité de l'énergie solaire en fonction des conditions météorologiques et des cycles diurnes peut rendre sa prévision un défi de taille. La disponibilité d'une prédiction précise de la production d'énergie solaire est essentielle pour planifier efficacement l'intégration de l'énergie solaire dans le réseau électrique, optimiser la gestion des ressources, et garantir un approvisionnement stable en électricité.

Dans cette optique, les techniques de machine learning se sont avérées être des outils puissants pour la prédiction de l'énergie solaire. L'une de ces méthodes, l'utilisation de modèles ARMA (AutoRegressive Moving Average) basés sur des séries temporelles, offre une approche particulièrement robuste pour modéliser et prédire les variations de la production d'énergie solaire.

Ce projet vise à démontrer l'efficacité d'un modèle ARMA pour la prédiction de l'énergie solaire en utilisant des données de séries temporelles historiques. Nous explorerons les différentes étapes du processus, de la collecte des données à la création et à l'évaluation du modèle ARMA. L'objectif ultime est de fournir une méthode précise et fiable pour anticiper la production d'énergie solaire, permettant ainsi une meilleure gestion des ressources énergétiques et une intégration plus efficace de l'énergie solaire dans le réseau électrique.

Ce rapport présentera en détail la méthodologie utilisée et les résultats obtenus pour la prédiction (prévision) de l'énergie solaire. Il contribuera à l'essor continu des énergies renouvelables en offrant une approche innovante pour surmonter les défis liés à la variabilité de l'énergie solaire.

# Chapitre 1:

## Etude préliminaires

### 1.Introduction :

D'abord, avant de détailler la réalisation de notre projet, nous faisons nécessairement une étude préalable de son contexte. Ainsi, nous commençons par présenter l'organisme d'accueil. Par suite, nous continuons en expliquant le cadre de projet.

Finalement, nous présentons la méthodologie adoptée pour l'élaboration de ce projet.

### 2.Presentation de l'organisation d'accueil Esprit :

L'École supérieure privée d'ingénierie et de technologie ou ESPRIT est une école d'ingénieurs privée de Tunisie basée à l'Ariana et agréée par le ministère de l'Enseignement supérieur et de la Recherche scientifique (agrément no2003-03). ESPRIT a également une branche école de commerce, la ESPRIT Business School. Depuis 2020, elle appartient au groupe Honoris United Universities<sup>2</sup>. En 2021, Entreprises Magazine classe ESPRIT, la première école d'ingénieurs de Tunisie.



Figure 1: Logo Esprit [1]



Figure 2: Esprit [1]

Avec plus de 10 000 étudiants et 240 enseignants permanents, Esprit est le plus grand établissement privé d'enseignement supérieur en Tunisie.

Fondé en 2003 il s'est rapidement forgé une réputation d'excellence, grâce à sa proximité avec l'univers entrepreneurial, ses partenariats avec les universités étrangères, et sa méthode d'apprentissage par problèmes et projets. ESPRIT forme des ingénieurs opérationnels dans 4 spécialités : la télécommunication, la technologie de l'information et de la communication, le génie civil, l'électromécanique.

En septembre 2017, l'ensemble des formations d'Esprit ont obtenu l'accréditation d'EUR-ACE fournie par l'organisation française CTI (Comité des titres d'ingénieurs). L'école est également membre du consortium de CDIO (Conceive, Design, Implement, Operate) fondé par Massachusetts Institute of Technology.

Pour finir Esprit est membre de la conférence des Grandes Ecoles (CGE) qui labellise les écoles d'ingénieur et de commerce les plus prestigieuses du monde. Depuis sa création en 2003, Esprit a évolué et plusieurs entités ont vu le jour : Esprit Entreprise (formation en direction des entreprises) et ESB (formation en management) et Top Skills (cabinet de formation continue).

ESPRIT est membre du réseau Honoris United Universities, le premier réseau panafricain d'enseignement supérieur privé engagé à former la nouvelle génération de leaders et de professionnels africains capables d'avoir un impact sur leurs sociétés et leurs économies dans un monde globalisé. Intelligence collaborative, agilité culturelle et mobilité sont au cœur de la vision d'Honoris en matière d'enseignement supérieur.

Honoris United Universities fusionne les savoirs et l'expertise de ses institutions membres pour développer un capital humain africain de classe mondiale, compétitif sur des marchés de plus en plus numérisés du travail et des start-ups.

## Contact de l'Esprit:

## Services de l'école

### Les directions opérationnelles:



#### Adresse

1, 2 rue André Ampère - 2083 - Pôle  
Technologique - El Ghazala.



#### Email

[contact@esprit.tn](mailto:contact@esprit.tn)



#### Téléphone

T (216) 70 250 000

- Direction de la formation.
- Direction de la RDI(ESPRIT-TECH)
- Direction cours de soir TIC
- Direction administrative et financière
- Direction des relations extérieurs.
- Direction des systèmes d'information.
- Direction des infrastructures, Hygiène ,  
Sécurité et Environnement.
- Direction d'Esprit Language Center.
- Direction démarche qualité et amélioration  
continue.
- Service communication.
- Esprit entreprise(Formation continue)

## 3.Contexte du projet:

### 3.1.Introduction :

Récemment, la production d'énergie a fait l'objet d'études approfondies en raison du risque d'approvisionnement les crises et les changements climatiques mondiaux. La production d'énergies renouvelables joue un rôle essentiel dans la croissance économique d'un pays. Le photovoltaïque solaire est considéré comme une ressource nécessaire pour production d'électricité.

### 3.2.But de projet de stage:

Pour assurer l'intégration sécurisée des systèmes photovoltaïques (PV) dans le réseau intelligent, les prévisions énergétiques sont un élément essentiel des systèmes de gestion de l'énergie. En réponse à l'opinion publique demande des clients, la recherche en prévision de l'énergie solaire a reçu un degré important au cours de la dernière décennie. En conséquence, le but de ce stage est de mettre en place une machine algorithmes d'apprentissage pour prédire la production d'énergie sur les paramètres météorologiques et temporels en se basant sur un modèle ARMA.



### **3.3.Problématique :**

L'évolution des énergies renouvelables décentralisées, comme le solaire et l'éolien, pose des défis pour STEG en termes d'intégration efficace. Les prévisions précises sur de courtes, moyennes et longues durées sont essentielles pour gérer la variabilité de ces sources, influencées par les conditions climatiques. Les centrales conventionnelles ne suffisent pas pour assurer cette flexibilité. L'objectif est d'obtenir des prévisions fiables pour garantir une intégration efficace des énergies renouvelables.

### **3.4.Solution proposée :**

La solution consiste à mettre en place un système de prévision de l'énergie solaire avec son interface. Ce sera le cas. mis en œuvre pour améliorer la précision des prévisions de puissance stable et réduire la les coûts. En outre, il sera très bénéfique pour les utilisateurs en termes de temps et ils pourront évaluer de nombreuses propriétés en peu de temps

### **3.5.Outils et technologies utilisées :**

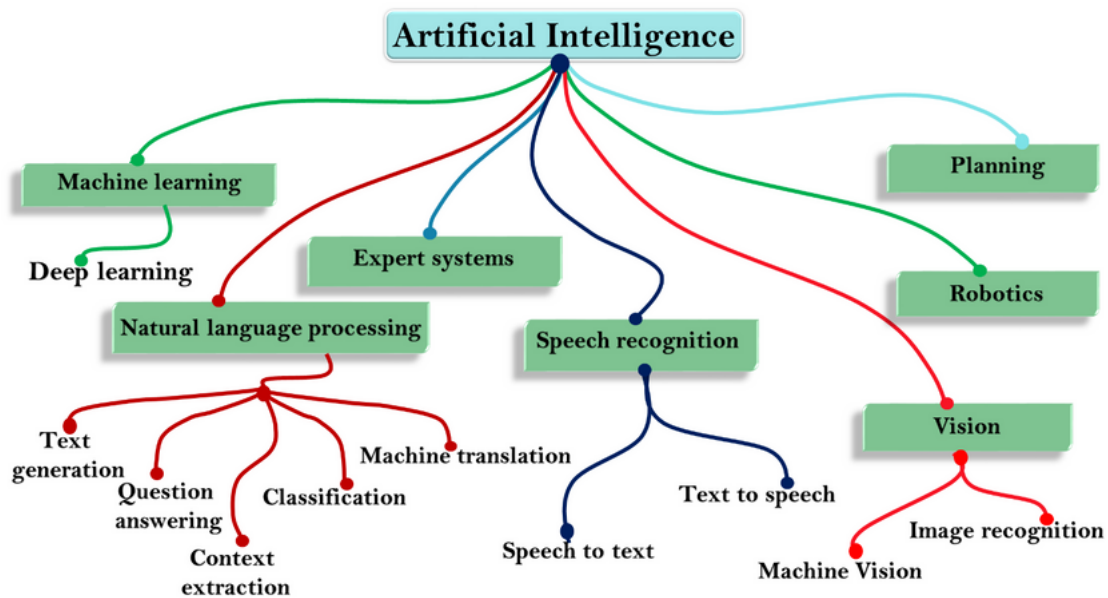
- **Intelligence artificielle:**

L'intelligence artificielle (faite par l'homme) (pouvoir de la pensée) (IA) [12] est une méthode informatique qui tente d'imiter les capacités cognitives humaines d'une manière très simple pour résoudre l'ingénierie problèmes qui ne peuvent être résolus par des techniques informatiques conventionnelles. L'essence de l'IA technologie pour résoudre n'importe quel problème d'ingénierie est d'apprendre à travers l'entrée et la sortie des exemples des données fournies, de sorte que même si la relation de base est inconnue ou la signification physique est difficile à expliquer, la relation fonctionnelle subtile entre les données peut être capturé.

En AI, la machine n'est pas programmée pour résoudre un problème, mais il peut apprendre et résoudre des problèmes plus complexes.

- **Les applications de l'Intelligence artificielle:**

Nous pouvons plonger plus profondément dans les différentes applications et domaines de l'IA à travers la figure ci-dessous:

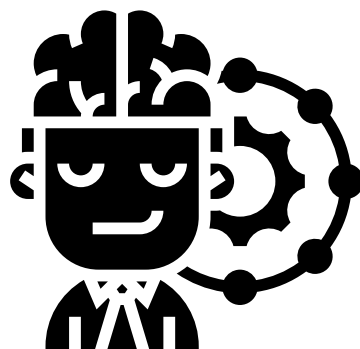


**Figure 3: Les applications et les branches de l'intelligence artificielle.**

- **Machine learning ( Apprentissage automatique):**

L'apprentissage automatique peut être considéré comme l'un des plus influents et des plus puissants technologies dans le monde d'aujourd'hui. Plus important encore, nous sommes loin de réaliser tout son potentiel.

Il ne fait aucun doute qu'il continuera de prévaloir dans un avenir prévisible. L'apprentissage automatique est un outil qui transforme l'information en connaissance. Au cours des 50 dernières années, les données ont explosé. À moins que vous analysiez et trouvez des schémas cachés, ce tas de données est inutile. Technologie d'apprentissage automatique est utilisé pour trouver automatiquement des modèles de base valables dans des données complexes, sinon il sera difficile pour nous de les trouver. Des modèles de problèmes cachés et des connaissances peuvent être utilisés pour prédire événements futurs et exécuter diverses décisions complexes.[15]



- **Les séries temporelles:(Données de série temporelles):**

Les données de séries chronologiques [18], également appelées données horodatées, sont une séquence de points de données. indexés dans l'ordre chronologique. Ces points de données sont généralement constitués de mesures successives effectuées provenant de la même source sur un intervalle de temps et sont utilisés pour suivre les changements au fil du temps.

Les données en type de séries temporelles(chronologiques) sont également l'un des types de données les plus courants disponibles aujourd'hui. Les données peuvent être comme la production d'une centrale solaire évolue au fil des ans, comme dans notre projet. Tout ce qui change avec le temps forme des séries temporelles.

- **Python:**

Python [23] est largement considéré comme le langage préféré pour l'apprentissage automatique parce qu'il est open source, a une excellente capacité de traitement des données, ainsi que la possibilité d'interagir avec presque toutes les langues et plates-formes tierces.



**Figure 4: logo python.**

- **Microsoft Excel :**

Microsoft Excel [27] appartient à la famille Microsoft Office et comprendre les données préliminaires



**Figure 5 :logo Microsoft excel.**

- **Pandas :**

Pandas est une bibliothèque open-source en Python utilisée pour la manipulation et l'analyse de données.

- **Matplotlib.pyplot:**

est une bibliothèque Python pour créer des visualisations graphiques.

- **NumPy:**

est une bibliothèque Python pour le calcul numérique et la manipulation de tableaux multidimensionnels.

- **statsmodels :**

Statsmodels est une bibliothèque Python pour l'estimation, la modélisation et l'analyse statistique.

# Chapitre 2:

## compréhension des données

### 1.Introduction:

Dans cette section, nous plongerons dans l'univers des données relatives à l'énergie solaire, mettant l'accent sur leur compréhension et leur explication. À travers des exemples concrets, nous démontrerons comment l'exploitation intelligente de ces données peut améliorer notre utilisation de l'énergie solaire.

### 2.Les energies renouvelables:

Les énergies renouvelables sont des énergies provenant de sources naturelles qui se renouvellent à un rythme supérieur à celui de leur consommation. La lumière du soleil et le vent, par exemple, constituent de telles sources qui se renouvellent constamment. Les sources d'énergie renouvelables sont abondantes et sont présentes partout autour de nous.

En revanche, les combustibles fossiles (charbon, pétrole et gaz) sont des ressources non renouvelables qui mettent des centaines de millions d'années à se constituer. Les combustibles fossiles, lorsqu'ils sont brûlés pour produire de l'énergie, provoquent des émissions de gaz à effet de serre nocifs, tels que le dioxyde de carbone.

La production d'énergie renouvelable génère bien moins d'émissions que la combustion de combustibles fossiles. Afin de faire face à la crise climatique, il est primordial de passer des combustibles fossiles, qui sont actuellement à l'origine de la majeure partie des émissions, aux sources d'énergie renouvelables.

Les énergies renouvelables sont désormais moins chères dans la plupart des pays et permettent de créer trois fois plus d'emplois que les combustibles fossiles.

Le secteur mondial de l'électricité connaît une transition progressive de l'énergie thermique conventionnelle. Les objectifs fixés par les pays à travers le monde indiquent une part croissante des énergies renouvelables dans le bouquet énergétique mondial pour les 20 prochaines années. La part des énergies renouvelables a été de 8,6% dans le mix énergétique en 2010 et devrait atteindre 22,5 % en 2020 selon une récente étude thématique rapport de recherche RE [8] par «GlobalData» [9].

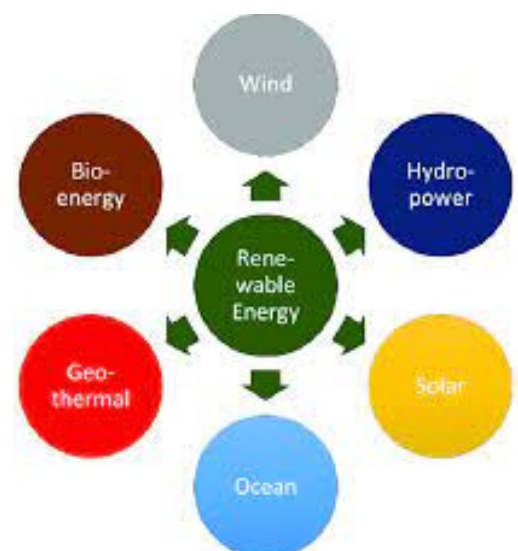
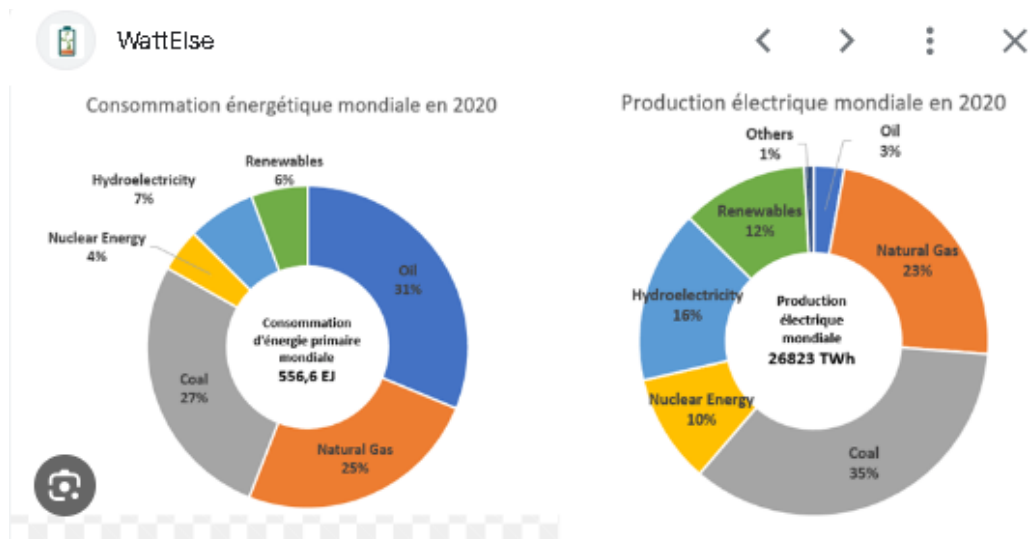


Figure 6:Types des énergies.



Les chiffres clés de l'énergie en 2020 - WattElse

[Consulter](#)

**Figure 7 : consommation énergétique et production électrique en 2020[1]**

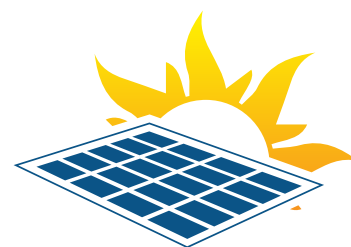
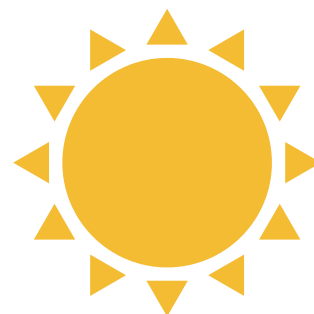
## 2.1.Énergie solaire:

L'énergie solaire est la plus abondante de toutes les ressources énergétiques et peut même être exploitée par temps nuageux. La vitesse à laquelle l'énergie solaire est interceptée par la Terre est environ 10 000 fois supérieure à la vitesse à laquelle l'humanité consomme de l'énergie.

Les technologies de l'énergie solaire permettent de produire de la chaleur, du froid, de l'éclairage naturel, de l'électricité et des carburants pour une multitude d'applications. Elles consistent à convertir la lumière du soleil en énergie électrique, soit au moyen de panneaux photovoltaïques, soit au moyen de miroirs qui concentrent le rayonnement solaire.

Même si tous les pays ne disposent pas de la même quantité d'énergie solaire, l'énergie solaire directe peut contribuer de manière importante au bouquet énergétique de chaque pays.

Les coûts de fabrication des panneaux solaires ont chuté de façon spectaculaire au cours de ces dix dernières années : non seulement sont-ils devenus abordables, mais il s'agit souvent de la forme d'électricité la moins chère. Les panneaux solaires ont une durée de vie d'environ 30 ans et se déclinent en différentes teintes selon le type de matériau utilisé pour leur fabrication.



Les sources d'énergie fossiles finiront par s'épuiser, c'est inévitable compte tenu de l'offre limitée. Mais là-bas sont des solutions, à savoir l'énergie renouvelable, un terme qui s'explique tout seul. C'est de l'énergie qui peut être renouvelé encore et encore. Il est facilement disponible à la lumière du soleil, au vent, à l'eau courante et en d'autres sources naturelles. L'énergie renouvelable est là pour être récoltée et peut être collectée avec un investissement énergétique relativement faible, utilisant de l'énergie courante comme par exemple des panneaux photovoltaïques

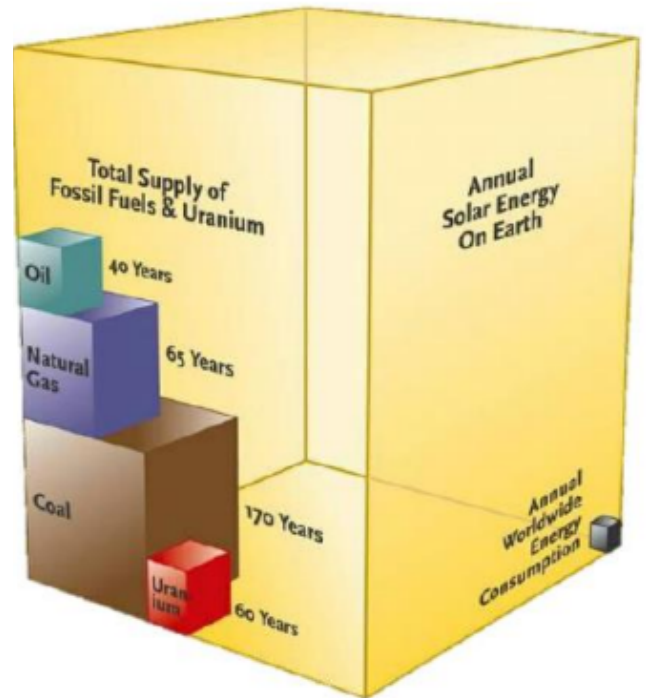


Figure 2.2: Renewable Vs. Non-renewable

Figure 8: Energie renouvelable et énergie non renouvelable[1]

## 2.2.Energie photovoltaïque:

L'énergie photovoltaïque est une énergie obtenue grâce aux rayonnements du soleil. Elle est ensuite récupérée par des **panneaux solaires** qui la transforment en électricité.

C'est le rayonnement émis par les modules photovoltaïques composés de cellules solaires. Ces derniers fonctionnent sur le principe de l'égalité des chances de l'effet photoélectrique. Plusieurs cellules sont reliées entre elles pour former un solaire photovoltaïque. module et plusieurs modules sont regroupés pour constituer une installation solaire. L'électricité produite peut être consommée sur place ou alimentée par un réseau de transport ou de distribution. Il se compose généralement d'un générateur photovoltaïque, d'un système de stockage (en option), d'un dispositif auxiliaire source (groupe diesel, éolienne, réseau, etc. ), systèmes d'interface (convertisseurs, réseau, un système de contrôle et de commande (système de surveillance, armoires électriques, cartes électroniques, etc. ) et en usage courant à des fins spécifiques. Cette utilisation (éclairage, réfrigération, pompage, communication. . . ) est utilisé dans différents secteurs (santé, éducation, agriculture, énergie. . .).

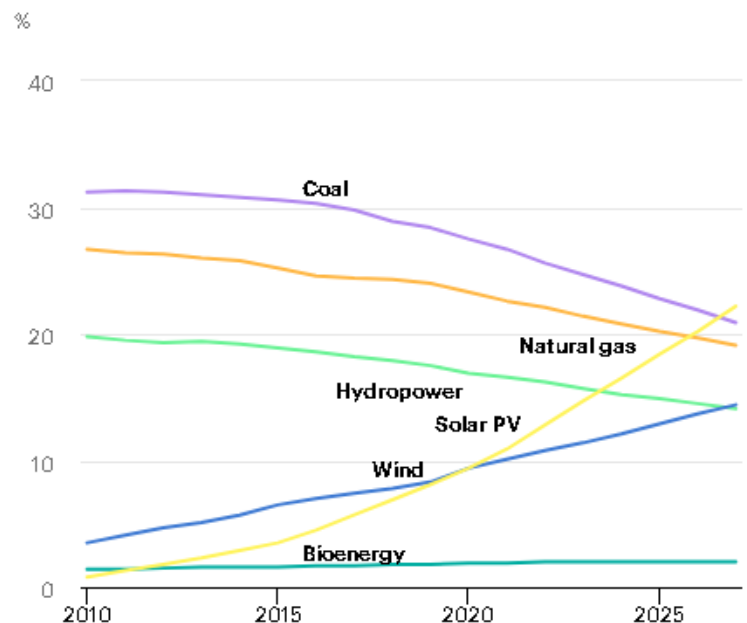


Figure 9: panneau photovoltaïque[1]

La partie principale de l'installation est le générateur photovoltaïque, qui se compose de plusieurs modules formés par une association soit en série, soit en parallèle de cellules élémentaires convertissant l'énergie solaire (sous forme de rayonnement) en énergie électrique. Une cellule peut produire 1.5w pour un soleil de 100w / m2 avec une tension de 0.6v

L'énergie solaire photovoltaïque domine toujours les augmentations de capacité d'énergie renouvelable.

La puissance installée du solaire photovoltaïque devrait dépasser celle du charbon d'ici 2027, devenant ainsi la plus importante au monde.



**Figure 10 : partition de la capacité d'électricité 2010-2027**  
[1]

### • les Facteurs affectant les panneaux solaires photovoltaïques:

Plusieurs facteurs influencent les performances des panneaux solaires photovoltaïques :

- **Ensoleillement** : La quantité de lumière solaire reçue affecte directement la production d'électricité. Des régions avec plus d'ensoleillement ont une production plus élevée.
- **Inclinaison et orientation** : L'angle d'inclinaison des panneaux et leur orientation par rapport au soleil sont cruciaux. Un ajustement optimal permet de capturer davantage de lumière solaire.
- **Nettoyage et entretien** : La propreté des panneaux est essentielle. La saleté, la poussière et d'autres contaminants peuvent réduire l'efficacité, donc un entretien régulier est nécessaire pour maximiser la production d'électricité.
- **Humidité** : L'augmentation du niveau d'humidité affecte fortement le courant, la tension et la puissance des panneaux solaires photovoltaïques en les diminuant.
- **Température** : La température a un impact négatif sur les panneaux solaires photovoltaïques, réduisant leur efficacité, provoquant une perte de tension et pouvant raccourcir leur durée de vie. Les systèmes de refroidissement peuvent être utilisés pour atténuer cet effet.



### 3.Modélisation des données :

#### 3.1.Introduction:

Le processus CRISP-DM nécessite l'acquisition de données(ou l'accès aux données). Cette collecte initiale comprend le chargement des données nécessaires pour comprendre les données. Une bonne compréhension des données nous permettra d'avoir une bonne phase de prétraitement. Le client nous a fourni des données brutes sur lesquelles nous travaillons. Nous nous concentrerons donc sur le processus de collecte des données nécessaires pour mieux l'explorer et comprendre quelques intuitions primaires

#### 3.2.Presentation CRISP-DM:

En tant que méthodologie, CRISP-DM comprend des descriptions des phases typiques d'un projet et des tâches comprises dans chaque phase, et une explication des relations entre ces tâches. En tant que modèle de processus, CRISP-DM offre un aperçu du cycle de vie de l'exploration de données. Notre tâche donc est la compréhension du métier



Figure 11: schéma explicatif du CRISP DM]



### 3.3. Les équations réagissant et influant sur l'énergie solaire :

$$\frac{d\mathbf{V}_3}{dt} = -2\boldsymbol{\Omega} \times \mathbf{V}_3 - \frac{1}{\rho} \nabla_3 p - \nabla_3 \Phi + \mathbf{F} \quad , \text{momentum equation,} \quad (1)$$

$$\frac{dT}{dt} = \frac{R}{C_p} \frac{T}{p} \frac{dp}{dt} + \frac{Q}{C_p} \quad , \text{thermodynamic equation,} \quad (2)$$

$$\frac{d\rho}{dt} = -\rho \nabla_3 \cdot \mathbf{V}_3 \quad , \text{continuity equation,} \quad (3)$$

$$\frac{dq}{dt} = M \quad , \text{water vapor equation,} \quad (4)$$

$$p = \rho R T \quad , \text{equation of state,} \quad (5)$$

#### Explication:

- $\mathbf{V}_3$  est le vent tridimensionnelle du vent.
- $\boldsymbol{\Omega}$  est l'angle vecteur de vitesse de la terre
- $\rho$  est la densité de l'air .
- $P$  pression
- $T$  est température.
- $q$  est la quantité spécifique de vapeur d'eau.
- $M$  est l'humidité.
- $F$  est la force.
- $\partial/\partial t$  est la dérivée partielle par rapport au temps.
- $\nabla$  est l'opérateur de gradient, qui représente la variation spatiale.
- $C_p$  est la capacité thermique des panneaux solaires

## **Équation de l'irradiance solaire en fonction de l'angle d'incidence solaire :**

$$I = I_0 * \cos(\theta)$$

- $I$  : L'irradiance solaire sur la surface (en watts par mètre carré).
- $I_0$  : L'irradiance solaire en dehors de l'atmosphère (constante solaire, en watts par mètre carré).
- $\theta$  : L'angle d'incidence solaire.

## **Équation de la puissance électrique produite par un panneau solaire photovoltaïque :**

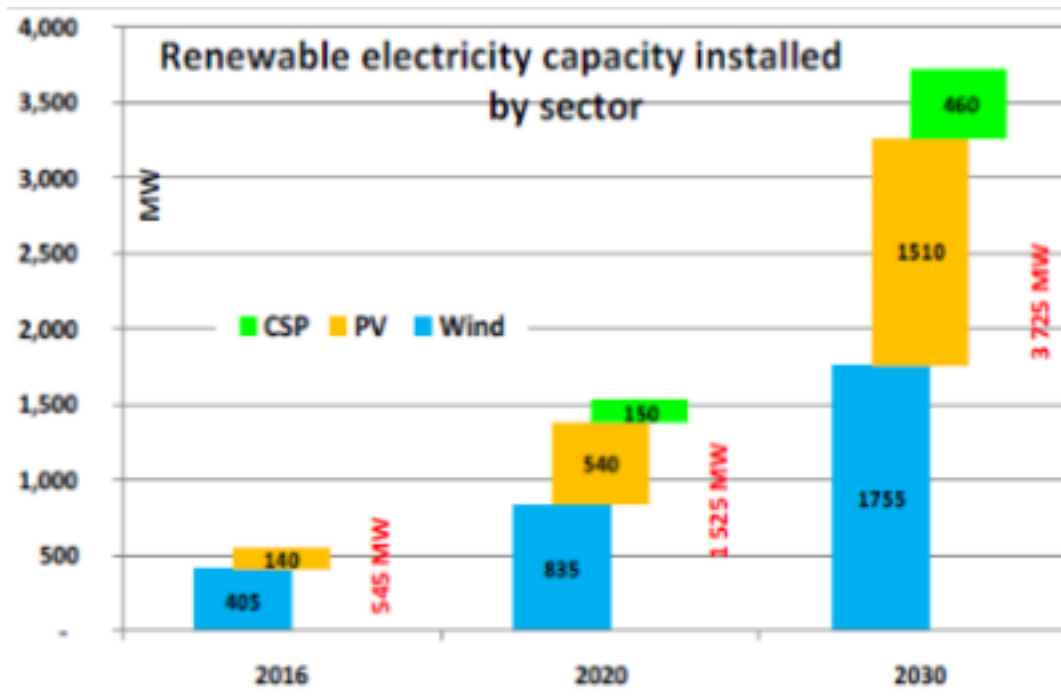
$$P = \eta * A * I$$

- $P$  : La puissance électrique produite (en watts).
- $\eta$  : L'efficacité du panneau solaire photovoltaïque.
- $A$  : L'aire de la surface du panneau exposée au soleil (en mètres carrés).
- $I$  : L'irradiance solaire sur la surface (en watts par mètre carré)

## **Équation de bilan thermique pour la température des panneaux solaires :**

$$T = T_0 + (I * \alpha - U * (T - T_a)) / (m * C_p)$$

- $T$  : La température des panneaux solaires (en degrés Celsius).
- $T_0$  : La température initiale des panneaux solaires (en degrés Celsius).
- $I$  : L'irradiance solaire sur la surface (en watts par mètre carré).
- $\alpha$  : Le coefficient d'absorption thermique des panneaux solaires.
- $U$  : Le coefficient de transfert de chaleur entre les panneaux et l'environnement.
- $T_a$  : La température ambiante (en degrés Celsius).
- $m$  : La masse des panneaux solaires (en kilogrammes).
- $C_p$  : La capacité thermique des panneaux solaires .



**Figure 12: capacité d'électricité renouvelable installé par secteur]**

==> Ces équations et modèles permettent de comprendre et de prédire l'évolution des facteurs climatiques qui affectent la production d'énergie solaire. Cependant, il est important de noter que la modélisation précise de ces facteurs peut varier en fonction de la région géographique, des caractéristiques spécifiques du système solaire et des données disponibles.

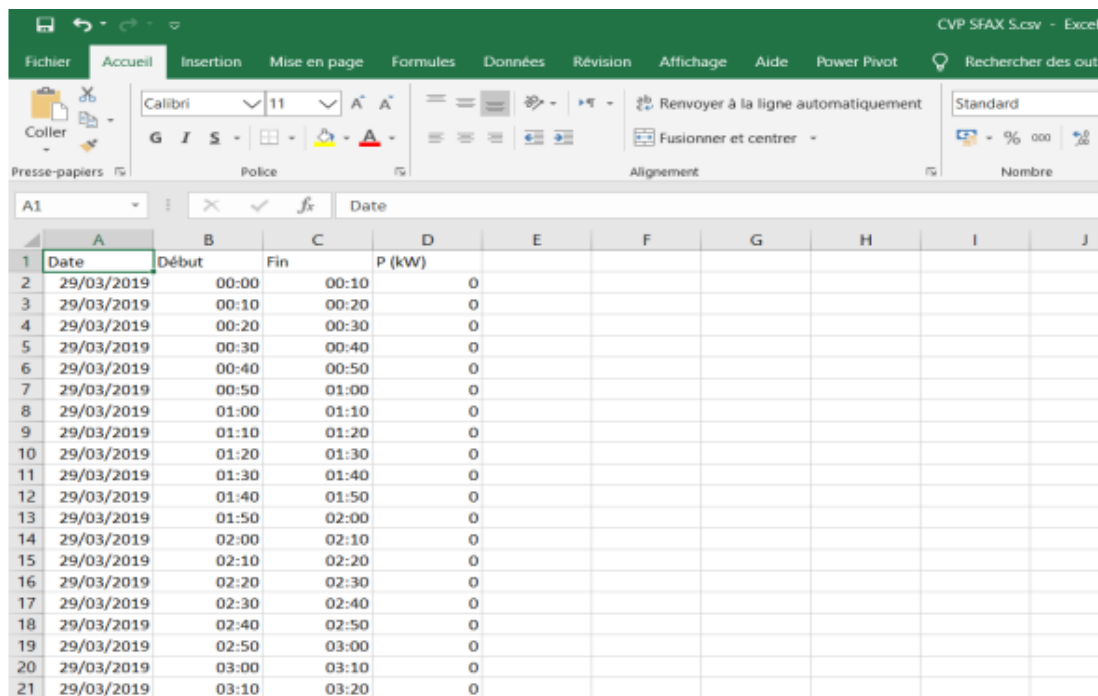
==> La capacité croissante d'électricité renouvelable installée est un signe prometteur de la transition énergétique mondiale vers des sources d'énergie plus durables. Au fil des années, les secteurs tels que l'énergie éolienne, solaire, hydraulique et biomasse ont montré une croissance constante, témoignant de l'engagement des gouvernements, des entreprises et des citoyens envers un avenir plus propre. Cette tendance est encourageante, car elle contribue à réduire les émissions de gaz à effet de serre, à créer des emplois verts et à renforcer la sécurité énergétique. Cependant, pour garantir un avenir énergétique véritablement durable, il est essentiel de poursuivre ces investissements, d'innover dans la technologie et de mettre en place des politiques incitatives pour accélérer la transition vers les énergies renouvelables.

En conclusion, la croissance de la capacité d'électricité renouvelable dans différents secteurs est une étape cruciale vers un avenir énergétique plus propre et plus durable.

## 4.Aperçu des données:

### 4.1.Description des sorties de l'énergie solaire:

Les données de production d'énergie solaire fournies par la centrale solaire de Sfax correspondent à la production réelle de l'usine entre le 29-03-2019 et le 05-05-2020 et a obtenu une étape d'enregistrement de 10min. Le jeu de données contient 58095 lignes dans le fichier Excel.



Date	Début	Fin	P (kW)
29/03/2019	00:00	00:10	0
29/03/2019	00:10	00:20	0
29/03/2019	00:20	00:30	0
29/03/2019	00:30	00:40	0
29/03/2019	00:40	00:50	0
29/03/2019	00:50	01:00	0
29/03/2019	01:00	01:10	0
29/03/2019	01:10	01:20	0
29/03/2019	01:20	01:30	0
29/03/2019	01:30	01:40	0
29/03/2019	01:40	01:50	0
29/03/2019	01:50	02:00	0
29/03/2019	02:00	02:10	0
29/03/2019	02:10	02:20	0
29/03/2019	02:20	02:30	0
29/03/2019	02:30	02:40	0
29/03/2019	02:40	02:50	0
29/03/2019	02:50	03:00	0
29/03/2019	03:00	03:10	0

Figure 4.3: Solar energy production in Sfax S

Figure 13: Production de l'énergie solaire a Sfax S [5]

Les valeurs de cette figure sont 0 car elles sont enregistrées de 00:00 à 03:10 ce qui est logique puisqu'il n'y a pas de rayonnement solaire la nuit, en d'autres termes la production d'énergie solaire sera égale à 0

### 4.2.Description des variables:

```
df.shape
```

```
(2716, 22)
```

```
df.columns
```

```
Index(['SITE', 'HUMIDITE_RELATIVE(%)', 'TEMPERATURE(°K)', 'TEMPERATURE(°C)',  
      'VENT_MERIDIEN', 'VENT_ZONAL', 'VENT(m/s) a 10m', 'VENT(m/s) a 2m',  
      'SUNSHINE_DURATION', 'SUNSHI_DURATION', 'SURFFLU_RAY_SOLA(w/m2)',  
      'RAY SOLAIRE DIRECT + DIFFUS (KW/m2)', 'RAY SOLAIRE ALBEDO (KW/m2)',  
      'RAY SOLAIRE GLOBAL (KW/m2)', 'SURFFLU_RAY_THER(w/m2)',  
      'RAY_THER (KW/m²)', 'SURFNEBUL_TOTALE(%)', 'SURFPRESSION(Pa)', 'DATE.1',  
      'DEBUT', 'FIN', 'Prod'],  
      dtype='object')
```

Figure 14: Capture des colonnes [5]

Notre ensemble de données brutes contient 22 colonnes comme indiqué dans la figure ci-dessus.

- **Site :**

Il représente le nom de l'endroit où se trouve la centrale solaire à Sfax.

- **Humidité relative (%):**

L'humidité relative est égale à 100 fois le rapport de la pression partielle de vapeur d'eau divisé par la pression partielle de vapeur saturée. Cette vapeur saturée est toujours calculée en ce qui concerne l'eau liquide, même à des valeurs de températures négatives.

$$\text{Relative Humidity} = \frac{\text{actual vapor density}}{\text{saturation vapor density}} \times 100\%$$

- **Température (°K) et (°C):**

Il s'agit de la température locale au niveau considéré. L'unité est le degré Kelvin (degré Celsius + 273,15 par définition).

- **Vent méridien à 10m (V m. s-1):**

C'est la composante méridienne du vent horizontal, avec la convention: positif pour un vent venant du sud et négatif pour un vent venant du nord.

- **Vent de zone à 10m (U m. s-1):**

C'est la composante zonale du vent horizontal, avec la convention: positif pour un vent venant de l'ouest et négatif pour un vent venant de l'est.

- **Durée d'ensoleillement (heure):**

Il s'agit de la durée pendant laquelle la surface du sol est irradiée par le rayonnement solaire direct (c. -à-d. la lumière du soleil atteignant la surface de la terre directement du soleil).

- **Rayonnement solaire (direct + diffus) (w/m2) :**

Cumul (depuis le début de la simulation) du flux solaire descendant à la surface.

- **Albédo :**

C'est la mesure de la réflexion diffuse du rayonnement solaire sur le rayonnement solaire total et mesurée sur une échelle de 0, correspondant à un corps noir qui absorbe tout le rayonnement incident, à 1, correspondant à un corps qui réfléchit tout le rayonnement incident.

- **Rayonnement solaire global (KW/m²):**

C'est la somme du rayonnement solaire direct + diffus et de l'albédo.

- **Rayonnement thermique de surface (w/m2):**

Cumul (depuis le début de la simulation) du flux thermique descendant à la surface.

- **Nébulosité totale (%):**

C'est la nébulosité "totale" qui est diagnostiquée en tenant compte du nuage combiné fractions (plus convective stratiforme) pour tous les niveaux verticaux des modèles.

- **Pression superficielle (Pa):**

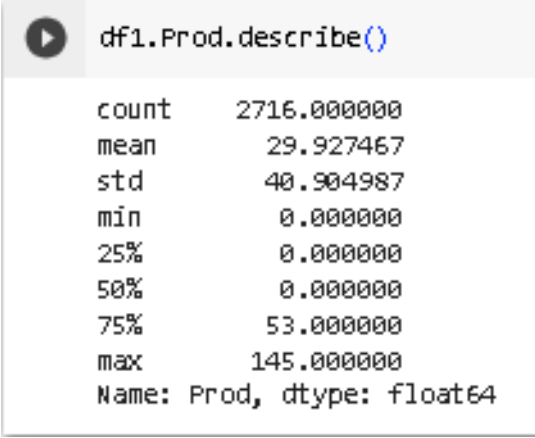
C'est la pression superficielle qui est utilisée comme variable pronostique. Donc, ce n'est pas la pression de soulagement, puisqu'il s'agit de la pression de relief interne utilisée dans le modèle ALADIN. En revanche, les allègements proposés ici sont calculés avec des valeurs zéro en mer. L'unité est Pascal (il faut donc diviser par 100 pour trouver hPa, anciennement millibars).

- **Prod :**

Elle correspond à la valeur de la production d'énergie solaire.

### 4.3.Description des données:

Nous commençons par notre variable cible qui est la valeur de la production d'énergie solaire, qui nous permet d'avoir les informations nécessaires à son sujet, telles que la moyenne, le minimum et valeur maximale. L'objectif principal de l'application étant de prédire la production solaire, nous vise à faire une description détaillée de la colonne "Prod".



```
df1.Prod.describe()
```

count	2716.000000
mean	29.927467
std	40.904987
min	0.000000
25%	0.000000
50%	0.000000
75%	53.000000
max	145.000000
Name: Prod, dtype: float64	

**Figure 15: Description du colonne "Prod" [5]**

Notre ensemble de données brutes contient 22 colonnes comme indiqué dans la figure ci-dessus. De ce chiffre, nous pouvons voir que nous avons un nombre exact de 2716 de valeur de production d'énergie solaire , nous remarquons une petite différence entre la valeur minimale qui est 0. Ce qui explique qu'il n'y a pas de production d'électricité spécialement la nuit que nous allons traiter dans les données

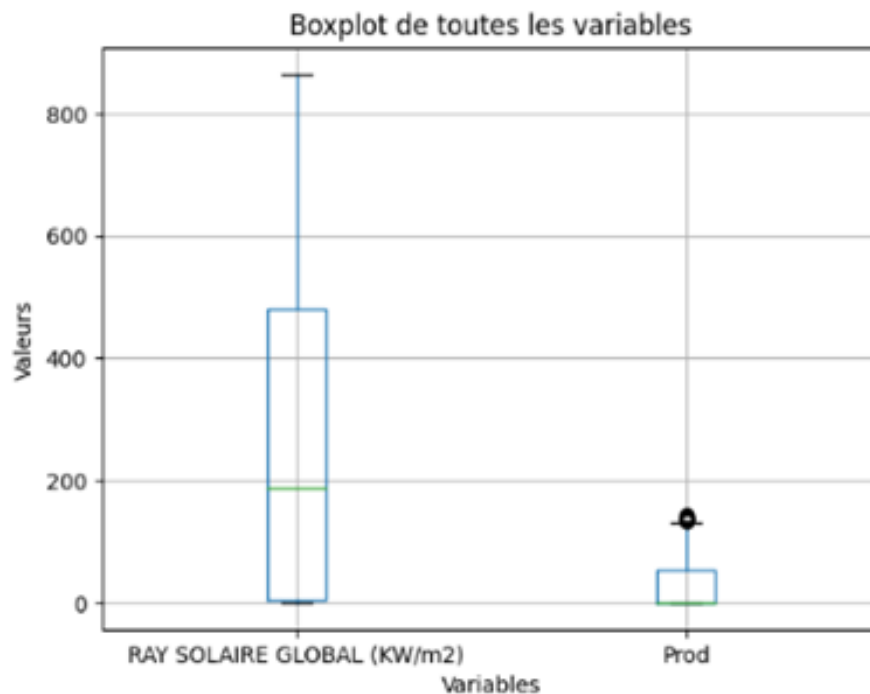
l'étape de préparation et la valeur maximale qui est 145. Nous constatons en particulier que la la médiane (0) est beaucoup plus faible que la moyenne (29,92). En d'autres termes, plus de la moitié des valeurs de la production est inférieure à la moyenne et moins de la moitié est supérieure à la moyenne.

## 4.5.Extraction de données:

### 4.5.1.Boîtes à moustaches:

La "boîte à moustaches" (box plot en anglais) est un type de graphique statistique qui permet de représenter graphiquement la distribution de données numériques.Box Plot est la représentation visuelle des groupes représentatifs de données numériques à travers leurs quartiles. Boxplot est également utilisé pour détecter les valeurs aberrantes dans l'ensemble de données. Il présente le résumé des données efficacement avec une boîte simple et des intervalles. Cela nous permet de comparer facilement les groupes. Boxplot résume un échantillon de données utilisant les 25e, 50e et 75e percentiles. Ces percentiles sont aussi appelés quartile inférieur, quartile médian et quartile supérieur. Les valeurs aberrantes peuvent influence notre modèle, le but principal d'un boxplot est d'établir un seuil qui définit un observation en tant que valeur aberrante.

```
[ ] import matplotlib.pyplot as plt
df1.boxplot()
plt.title("Boxplot de toutes les variables")
plt.xlabel("Variables")
plt.ylabel("Valeurs")
plt.show()
```



**Figure 16: Boite à moustaches [5]**

A partir de la courbe de la variable 'Prod', nous remarquons qu'il y a des valeurs au-dessus de la moustache supérieure du boxplot qui sont des valeurs aberrantes, nous fournissant des informations sur des comportements spécifiques.

### 4.5.2. Test de normalité:

Un élément essentiel de la compréhension des données est le «test de normalité» qui sert à déterminer si un ensemble de données est bien modélisé par une distribution normale et pour calculer la probabilité qu'une variable aléatoire sous-jacente à l'ensemble de données est normalement distribuée.

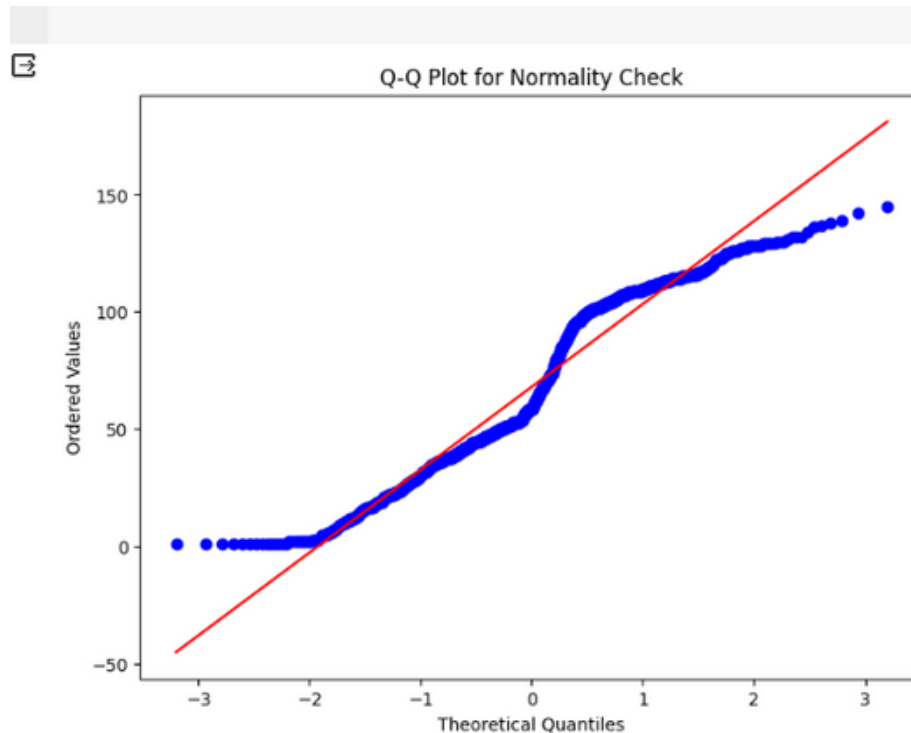


Figure 17: Contrôle de la normalité [5]

```
The value of skewness is:  
0.03313387433174709  
The value of kurtosis is:  
-1.237296430568
```

Pour un ensemble de données normal, l'asymétrie et la kurtosis devraient être nulles. Il y a une très bonne asymétrie positive qui est de 0,033 ainsi qu'un indice de kurtose de -1,23 qui sont clairement un nombre négatif proche de zéro.

### 4.5.3. Analyse des valeurs manquantes:

Nous avons analysé les valeurs manquantes ainsi que leurs distributions. La stratégie s'appliquait aux données manquantes seront déduites des résultats de cette analyse. Ce qui suit: figure 18 montre le script et le résultat des valeurs manquantes de chaque variable.



missing

	feature	percentage nan
14	RAY SOLAIRE GLOBAL (KW/m2)	0.388807
13	RAY SOLAIRE ALBEDO (KW/m2)	0.388439
12	RAY SOLAIRE DIRECT + DIFFUS (KW/m2)	0.388439
11	SURFFLU_RAY_SOLA(w/m2)	0.000368
15	SURFFLU_RAY_THER(w/m2)	0.000368
1	DATE	0.000000
20	FIN	0.000000
19	DEBUT	0.000000
18	DATE.1	0.000000
17	SURFPRESSION(Pa)	0.000000
16	SURFNEBUL_TOTALE(%)	0.000000
0	SITE	0.000000
10	SUNSHINE_DURATION	0.000000
9	VENT(m/s) a 2m	0.000000
8	VENT(m/s) a 10m	0.000000
7	VENT_ZONAL	0.000000
6	VENT_MERIDIEN	0.000000
5	TEMPERATURE(°C)	0.000000
4	TEMPERATURE(°K)	0.000000
3	HUMIDITE_RELATIVE(%)	0.000000
2	HEURE	0.000000
21	Prod	0.000000

[9]

```
df1.isna().sum()
print(df1.isnull().sum())
print(df1.isnull().values.any())
print(df1.isnull().sum().sum())
```

```
SITE 0
DATE 0
HEURE 0
HUMIDITE_RELATIVE(%) 0
TEMPERATURE(°K) 0
TEMPERATURE(°C) 0
VENT_MERIDIEN 0
VENT_ZONAL 0
VENT(m/s) a 10m 0
VENT(m/s) a 2m 0
SUNSHINE_DURATION 0
SURFFLU_RAY_SOLA(w/m2) 1
RAY SOLAIRE DIRECT + DIFFUS (KH/m2) 1055
RAY SOLAIRE ALBEDO (KH/m2) 1055
RAY SOLAIRE GLOBAL (KH/m2) 1056
SURFFLU_RAY_THER(w/m2) 1
SURFNEBUL_TOTALE(%) 0
SURFPRESSION(Pa) 0
DATE.1 0
DEBUT 0
FIN 0
Prod 0
dtype: int64
True
3168
```

**Figure 18: Informations sur les valeurs manquantes [5]**

D'après le graphique ci-dessus, le pourcentage maximal des valeurs manquantes d'une certaine fonctionnalité n'est pas si élevé. Pour la caractéristique «RAY\_SOLAIRE GLOBAL (KW/m<sup>2</sup>)», le pourcentage de valeur est de 38%, ce qui est un pourcentage significatif qui peut être traité en remplissant une valeur estimée dans les champs vides tels que la moyenne des colonnes par exemple ou simplement la baisse d'entreeux. L'approche adoptée dans notre projet sera expliquée dans le chapitre suivant.

## 5.Conclusion:

Dans ce chapitre, nous avons établi une explication du processus que nous avons suivi pour nous assurer que les données étape de compréhension, a traité des données manquantes et des valeurs aberrantes ainsi que des données illogiques, et a tenu un aperçu des données. Ces données seront transformées et corrigées au cours de la préparation des données , qui fera l'objet du prochain chapitre.

# Chapitre 3:

## préparation des données

### 1.Introduction:

L'une des phases les plus importantes et les plus longues d'un projet d'exploration de données est la préparation. La préparation des données pourrait nécessiter 70 % du temps et des efforts d'un projet. Donner l'énergie suffisante pour les premières étapes de la compréhension des affaires et de la compréhension des données joue évidemment un rôle important dans la réduction de ce surcoût.

### 2.Nettoyage des données:

Dans tout projet de science des données, il est pratiquement impossible d'avoir un cas parfait et un base de données aussi propre, homogène que complète. Par conséquent, il est absolument nécessaire de corriger plusieurs ambiguïtés. Ces problèmes doivent être corrigés manuellement, soit dans le fichier csv / excel ou avec un script python pendant la phase de prétraitement et nous allons essayer d'améliorer la qualité des données en insérant des valeurs par défaut appropriées. Nous devons choisir les transformations nécessaires en fonction de l'objectif commercial qui contribue à accroître la précision du modèle.

Commençant par l'extraction des valeurs aberrantes:

	RAY SOLAIRE GLOBAL (KH/m2)	SURFFLU_RAY_THER(w/m2)	SURFNEBUL_TOTALE(%)
2252	650.588889	-3618800	0.30001
2276	639.866667	-3443100	0.31789
2396	774.533333	-4068300	0.064988
2404	743.444444	-2985900	0.015549
2468	682.577778	-3145800	0.30001
2525	367.533333	-2733000	0.76298
2605	680.111111	-4138200	0
2620	NaN	NaN	NaN

	SURFPRESSION(Pa)	DATE.1	DEBUT	FIN	Prod
2252	101980	2020-03-08	12:00:00	12:10:00	145
2276	102290	2020-03-11	12:00:00	12:10:00	138
2396	101030	2020-03-26	12:00:00	12:10:00	142
2404	100930	2020-03-27	12:00:00	12:10:00	136
2468	101640	2020-04-04	12:00:00	12:10:00	139
2525	101970	2020-04-11	15:00:00	15:10:00	137
2605	100660	2020-04-21	15:00:00	15:10:00	134
2620	NaN	2020-04-23	12:00:00	12:10:00	134

Figure 19: Les valeurs aberrantes [5]

On transforme les points aberrantes en points NA.

```
transformation des pts aberrantes en NA

import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

# Mise à NA pour certaines valeurs d'âge
df1.loc[134, 'Prod'] = np.nan
df1.loc[134, 'Prod'] = np.nan
df1.loc[137, 'Prod'] = np.nan
df1.loc[139, 'Prod'] = np.nan
df1.loc[136, 'Prod'] = np.nan
df1.loc[142, 'Prod'] = np.nan
df1.loc[138, 'Prod'] = np.nan
df1.loc[145, 'Prod'] = np.nan
```

Figure 20: Transformation des valeurs aberrantes en NA [5]

Après la transformation des points aberrantes en Na en trouve seulement des NA dans notre base de données.

Les valeurs manquantes dans une base de données, souvent notées comme "NaN" (Not a Number) ou "NA" (Not Available), sont des valeurs qui sont absentes ou inconnues pour certaines entrées de données. Ces valeurs manquantes peuvent survenir pour diverses raisons, comme des erreurs lors de la collecte ou de la saisie des données, des données non disponibles pour certaines observations, etc.

```
df1.isna().sum()
print(df1.isnull().sum())
print(df1.isnull().values.any())
print(df1.isnull().sum().sum())
```

SITE	0
DATE	0
HEURE	0
HUMIDITE_RELATIVE(%)	0
TEMPERATURE(°K)	0
TEMPERATURE(°C)	0
VENT_MERIDIEN	0
VENT_ZONAL	0
VENT(m/s) a 10m	0
VENT(m/s) a 2m	0
SUNSHINE_DURATION	0
SURFFLU_RAY_SOLA(W/m2)	1
RAY SOLAIRE DIRECT + DIFFUS (KW/m2)	1055
RAY SOLAIRE ALBEDO (KW/m2)	1055
RAY SOLAIRE GLOBAL (KW/m2)	1056
SURFFLU_RAY_THER(W/m2)	1
SURFNEBUL_TOTALE(%)	0
SURFPRESSION(Pa)	0
DATE.1	0
DEBUT	0
FIN	0
Prod	59
dtype: int64	
True	
3227	

Figure 21: L'ensemble des valeurs manquantes totale [5]

Maintenant on peut faire **l'imputation des données**.

Pour traiter les valeurs manquantes, nous allons supprimer toutes les valeurs NaN de la base de données et la suppression de toutes les valeurs zéro dans la colonne «Prod» parce que nous n'aurons pas besoin de valeurs de production solaires lorsque la valeur est nulle. Elles seront inutiles et désorientantes pour la formation du modèle.

La suppression des valeurs manquantes n'affecte pas la base de données ou le modèle car les valeurs NAN ne sont pas en grand nombre et négligeable en comparant avec toutes le données.

```
[33] df1=df1[(df1[['Prod']]!=0).all(axis=1)]
      df1=df1.dropna()
      df1
```

df1

		SITE	DATE	HEURE	HUMIDITE_RELATIVE(%)	TEMPERATURE(*K)	TEMPERATURE(*C)	RAY SOLAIRE DIRECT + DIFFUS (KH/m2)	RAY SOLAIRE ALBEDO (KH/m2)	RAY SOLAIRE GLOBAL (KH/m2)
3	Sfax_S	2019.06.01	09:00:00		0.51241	295.64	22.49	426.732407	85.346481	512.078889
4	Sfax_S	2019.06.01	12:00:00		0.5384	297.05	23.9	643.62963	128.725926	772.355556
5	Sfax_S	2019.06.01	15:00:00		0.61581	295.06	21.91	583.888889	116.777778	700.666667
6	Sfax_S	2019.06.01	18:00:00		0.67318	292.4	19.25	200.277778	40.055556	240.333333
11	Sfax_S	2019.06.02	09:00:00		0.50887	295.48	22.33	429.319444	85.863889	515.183333
...	...	...	...		...	...	...	...	...	...
2707	Sfax_S	2020.05.04	09:00:00		0.41503	299.49	26.34	398.265741	79.653148	477.918889
2708	Sfax_S	2020.05.04	12:00:00		0.49105	300.39	27.24	698.333333	139.666667	838.000000
2709	Sfax_S	2020.05.04	15:00:00		0.58245	297.73	24.58	579.722222	115.944444	695.666667
2710	Sfax_S	2020.05.04	18:00:00		0.72702	294.18	21.03	160.37037	32.074074	192.444444
2715	Sfax_S	2020.05.05	09:00:00		0.78145	294.97	21.82	382.058333	76.411667	458.470000

**Figure 22: Les données après imputation [5]**

Cela est fait pour obtenir un ensemble de données propre et adapté à l'analyse ultérieure.

Pour traiter les valeurs manquantes, nous avons supprimé toutes les valeurs NaN du dataset et la suppression de toutes les valeurs zéro dans la colonne 'Prod' parce que nous n'aurons pas besoin de valeur d'énergie solaire lorsque la valeur est égale à zéro. Elles seront inutiles et désorientantes pour la formation du modèle.

Vérification de la propreté de l'ensemble de données :

```
df1.isna().sum()
print(df1.isnull().sum())
print(df1.isnull().values.any())
print(df1.isnull().sum().sum())
```

SITE	0
DATE	0
HEURE	0
HUMIDITE_RELATIVE(%)	0
TEMPERATURE(°K)	0
TEMPERATURE(°C)	0
VENT_MERIDIEN	0
VENT_ZONAL	0
VENT(m/s) a 10m	0
VENT(m/s) a 2m	0
SUNSHINE_DURATION	0
SURFFLU_RAY_SOLA(w/m2)	0
RAY SOLAIRE DIRECT + DIFFUS (KW/m2)	0
RAY SOLAIRE ALBEDO (KW/m2)	0
RAY SOLAIRE GLOBAL (KW/m2)	0
SURFFLU_RAY_THER(w/m2)	0
SURFNEBUL_TOTALE(%)	0
SURFPRESSION(Pa)	0
DATE.1	0
DEBUT	0
FIN	0
Prod	0
dtype: int64	
False	

**Figure 23: Pas de valeurs manquantes après imputation [5]**

==>La phase de préparation des données prend plus de 70 % du temps estimé du projet.

Ces efforts sont consacrés à la collecte, au nettoyage et à la préparation des données en vue d'une analyse ultérieure. Les données est maintenant utilisable et prêt à l'emploi dans la phase de modélisation.

# Chapitre 4:

# Modélisation avec série temporelle

## 1.Introduction:

Nous pouvons maintenant aborder la partie attendue de la modélisation. C'est probablement la partie du modèle de processus CRISP-DM que les spécialistes en DS attendent le plus. Les données sont propres, dans un formulaire prêt à être utilisé pour les modèles prédictifs. Le data mining offre de nombreuses techniques de modélisation, mais ils ne répondent pas tous à nos besoins. Dans l'exploration de données, il existe de nombreux algorithmes de prédiction, mais nous devons choisir celui qui répond à nos besoins et correspond à nos données. Dans ce chapitre, nous ferons la prévision à l'aide des séries temporelles.

## 2.Modélisation:

La modélisation des données joue un rôle important dans la croissance de toute entreprise qui comprend des décisions basées sur des algorithmes d'apprentissage automatique sont la clé de leur succès. Avoir les données dans le bon format et l'application d'algorithmes vous permet d'obtenir les réponses à votre entreprise questions et atteindre divers objectifs commerciaux plus facilement et plus rapidement pour permettre aux parties concernées à prendre des décisions basées sur les résultats de la modélisation. L'application de ces algorithmes et la modélisation des données diffère selon la nature de l'entreprise

### 2.1.Séries temporelles:

L'analyse des séries temporelles en apprentissage automatique est un processus essentiel pour comprendre et exploiter des données qui évoluent dans le temps. Pour ce faire, on commence par collecter et prétraiter les données temporelles, puis on explore leur structure pour en extraire des informations significatives. Ensuite, on divise ces données en ensembles d'apprentissage et de test pour former un modèle adapté, tel qu'un modèle ARIMA ou un réseau de neurones, que l'on optimise en fonction des performances. Une fois le modèle prêt, il peut être utilisé pour faire des prévisions sur de nouvelles données et, si nécessaire, être déployé en production pour automatiser des décisions basées sur les séries temporelles. La complexité de cette démarche dépend largement de la nature spécifique des données temporelles et des tendances qu'elles présentent.

Avant la création du série temporelle, on doit grouper les observations du variable prod par jour et agrégez-les en utilisant la moyenne:

```
import pandas as pd

# Groupez les observations par jour et agrégez-les en utilisant la moyenne,
df1['DATE'] = pd.to_datetime(df1['DATE'])
data= df1.groupby(df1['DATE'].dt.date).agg({'Prod': 'mean'})

# Affichez le DataFrame résultant
print(data)
```

DATE	Prod
2019-06-01	69.50
2019-06-02	80.25
2019-06-03	81.50
2019-06-04	77.50
2019-06-05	76.25
...	...
2020-04-30	93.50
2020-05-02	93.50
2020-05-03	93.00
2020-05-04	87.75
2020-05-05	58.00

[279 rows x 1 columns]

**Figure 24: Les valeurs de Prod journalièrement[8]**

## Création du série temporelle:

```
[ ] import pandas as pd

# Créer une série temporelle avec des dates et des valeurs
serie_temporelle = pd.Series(vecteur_prod,
                             index=pd.date_range(start='2019-06-01', periods=279, freq='D'))

# Afficher la série temporelle
print(serie_temporelle)
```

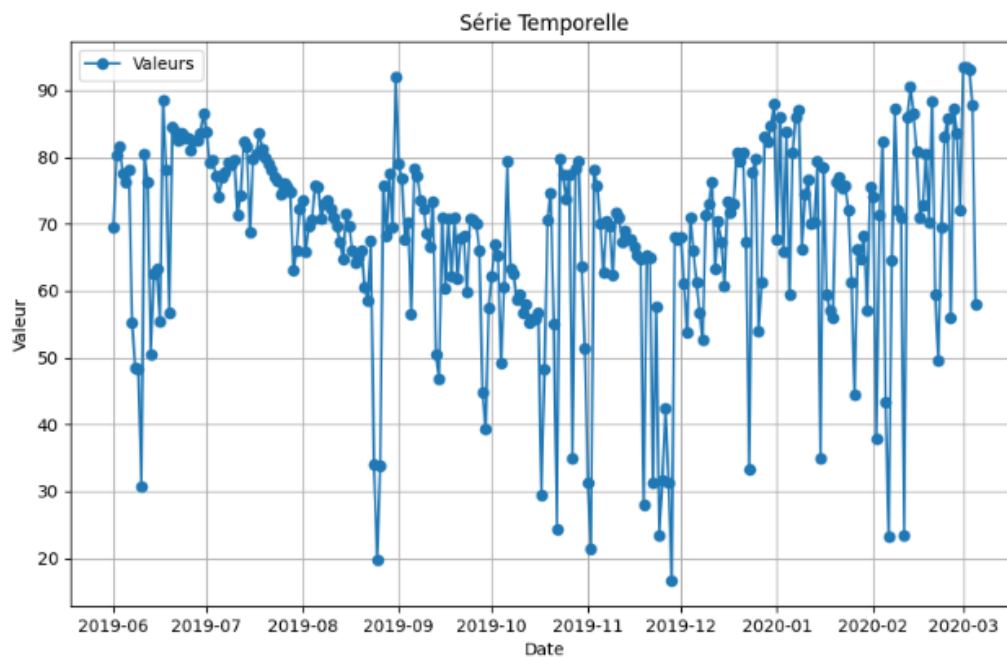
2019-06-01	69.50
2019-06-02	80.25
2019-06-03	81.50
2019-06-04	77.50
2019-06-05	76.25
...	...
2020-03-01	93.50
2020-03-02	93.50
2020-03-03	93.00
2020-03-04	87.75
2020-03-05	58.00

Freq: D, Length: 279, dtype: float64

**Figure 25: Création du série temporelle[8]**

La création d'une série temporelle pour la variable "prod" suit généralement un processus en plusieurs étapes. Il commence par l'importation des bibliothèques nécessaires, puis le chargement et le prétraitement des données historiques de production d'énergie solaire. Ensuite, une série temporelle est créée, avec les dates/heure en tant qu'index et les valeurs de production d'énergie associées. Des modèles de prédiction sont utilisés pour anticiper la production future, suivis de l'évaluation du modèle à l'aide de mesures de performance. Enfin, le modèle formé est utilisé pour effectuer des prédictions de la production d'énergie solaire à venir.

## Représentation graphique et interprétation:



**Figure 26:Graphe représentant la série temporelle [8]**

### Interprétation graphique :

- La valeur journalière de la production de l'énergie solaire varie entre 10 et 95 selon les dates et les différents autres variables influant dans cette production.
- Un zoom de la série (figure en bas) sur les mois de 9 à 12 de l'année 2019 montre une décroissance du niveau moyen de la série qui est due à la décroissance de l'énergie solaire dans l'hiver.
- Une série temporelle plus longue montrerait également des tendances saisonnières. Par exemple, en été, la production d'énergie solaire peut être plus élevée en raison de journées plus longues et de plus de luminosité solaire par rapport à l'hiver.
- Les baisses soudaines dans la courbe peuvent être causées par des nuages qui passent devant le soleil ou par des ombres de bâtiments ou d'arbres.
- On constate 3 pics : une au mois de Juin, une au mois d'Aout et l'autre au mois de Mars qui représente les maximums de la production en énergie solaire.

## 2.2.Analyse de stationnarité :

La stationnarité est un concept fondamental en séries temporelles en machine learning. Une série temporelle est dite stationnaire lorsqu'elle satisfait à certaines conditions qui rendent son comportement statistique relativement constant au fil du temps. La stationnarité est importante car de nombreux modèles de séries temporelles supposent que la série est stationnaire pour produire des prédictions précises.



```
[ ] import pandas as pd
    from statsmodels.tsa.stattools import adfuller

    # Effectuez le test ADF
    resultat_adf = adfuller(serie_temporelle)

    # Affichez les résultats
    print("Statistique du test ADF :", resultat_adf[0])
    print("P-valeur :", resultat_adf[1])

    # Interprétation des résultats
    if resultat_adf[1] < 0.05:
        print("La série temporelle est stationnaire (rejette l'hypothèse nulle)")
    else:
        print("La série temporelle n'est pas stationnaire (ne rejette pas l'hypothèse nulle)")
```

Résultat:

```
Statistique du test ADF : -4.281080866789191
P-valeur : 0.0004789428049439026
La série temporelle est stationnaire (rejette l'hypothèse nulle)
```

### Figure 27: Test de stationnarité du série temporelle [8]

Puisque le résultat du test ADF a une p-valeur (0.0004789...) inférieure au seuil 0,05, nous pouvons rejeter l'hypothèse nulle et conclure que la série est stationnaire.

La stationnarité est une propriété importante des séries temporelles car de nombreuses méthodes d'analyse et de modélisation des séries temporelles supposent que la série est stationnaire. Une série stationnaire est plus facile à modéliser et à prévoir car ses propriétés statistiques ne changent pas avec le temps.

- il y a une auto-corrélation entre les erreurs.
- Donc la série n'est pas assimilable à un bruit blanc.

## 2.3. Modélisation de la série temporelle : les modèles ARMA

Les modèles ARIMA (AutoRegressive Integrated Moving Average) sont des outils puissants en séries temporelles en machine learning. Ils combinent trois composantes essentielles : l'AutoRégressive (AR) pour les dépendances temporelles, l'Intégration (I) pour rendre les données stationnaires, et la Moyenne Mobile (MA) pour la dépendance par rapport aux erreurs. Les modèles ARIMA sont notés ARIMA(p, d, q), où p est l'ordre de l'AR, d est l'ordre d'intégration, et q est l'ordre de la MA. Ils sont particulièrement utiles pour modéliser des séries temporelles, même si elles ne sont pas stationnaires initialement. Les étapes typiques comprennent l'identification des ordres, l'ajustement du modèle, l'estimation des paramètres, l'évaluation de la performance, et la prévision. Les modèles ARIMA sont couramment utilisés pour la prévision et l'analyse de séries temporelles, offrant une méthodologie solide pour traiter une variété de données temporelles.

Pour un choix du meilleur modèle, on utilise les critères de sélection AIC :

L'AIC :l'Akaike Information Criterion est un critère couramment utilisé pour sélectionner le meilleur modèle parmi un ensemble de modèles concurrents dans le domaine de la statistique et de l'analyse de données. L'AIC est basé sur la théorie de l'information et est utilisé pour évaluer la qualité d'un modèle en tenant compte de sa capacité à ajuster les données tout en pénalisant la complexité du modèle. En général, l'objectif est de minimiser l'AIC pour sélectionner le modèle le plus approprié. , offrant une méthodologie solide pour traiter une variété de données temporelles.

```
print(f"Meilleurs paramètres (p, d, q) : {best_para}")
print(f"Meilleur AIC : {best_aic}")

/usr/local/lib/python3.10/dist-packages/statsmodels
warn('Non-stationary starting autoregressive para
/usr/local/lib/python3.10/dist-packages/statsmodels
warn('Non-invertible starting MA parameters found
/usr/local/lib/python3.10/dist-packages/statsmodels
warn('Non-stationary starting autoregressive para
/usr/local/lib/python3.10/dist-packages/statsmodels
warn('Non-invertible starting MA parameters found
/usr/local/lib/python3.10/dist-packages/statsmodels
warn('Non-stationary starting autoregressive para
/usr/local/lib/python3.10/dist-packages/statsmodels
warn('Non-invertible starting MA parameters found
Meilleurs paramètres (p, d, q) : (1, 0, 2)
Meilleur AIC : 2245.885039190296
```

==> Après le choix des paramètres p, d et q a travers AIC avec python on conclut que le meilleur choix est (p, d, q)= (1,0, 2) qui correspond au AIC minimal.

Figure 28:Choix du meilleur AIC pour un meilleur modèle[9]

Maintenant on synthétise le modèle:

==> Ici on constate que normalement les paramètres sont adéquates car tous les p-value inférieure à 0.05.

```
from statsmodels.tsa.arima.model import ARIMA

arma = ARIMA(serie_temporelle, order=(1,0,2)).fit()

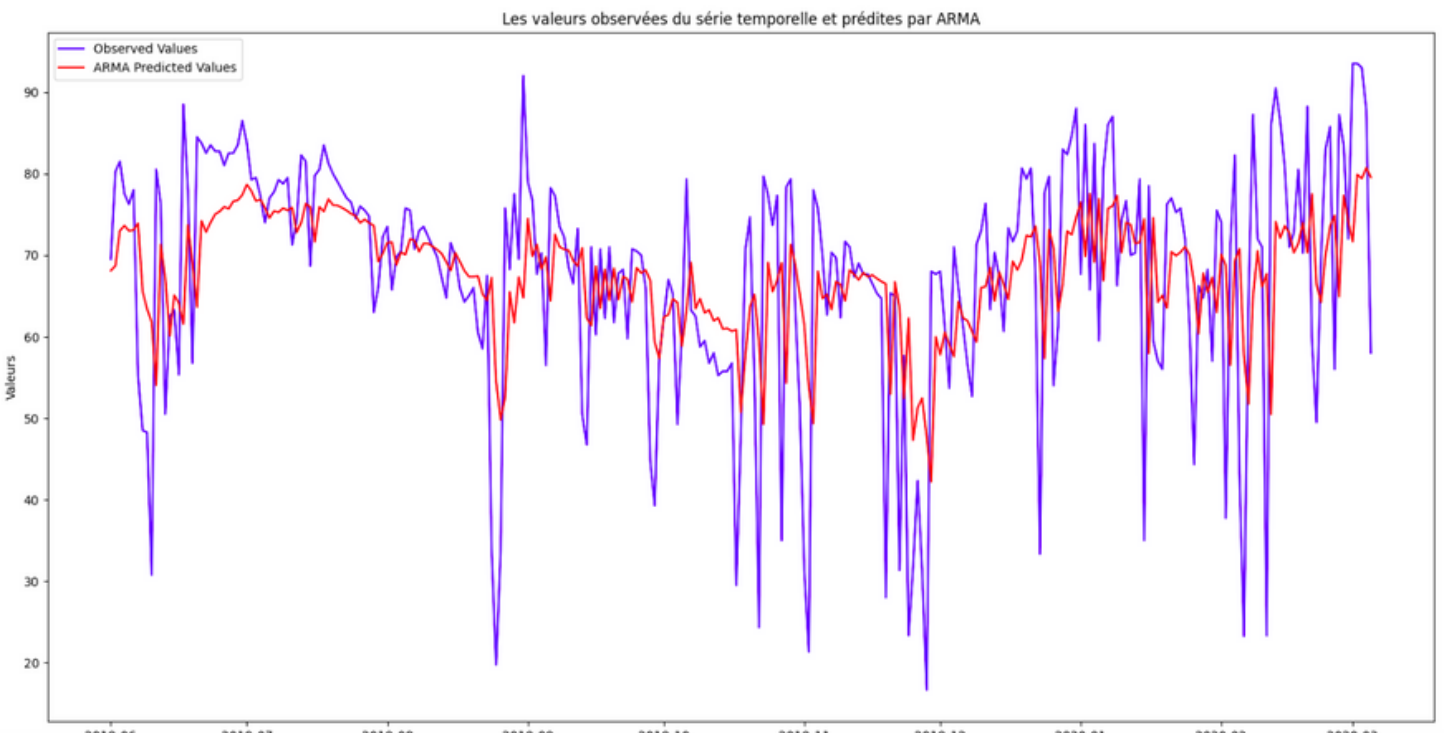
arma.summary()
```

```
/usr/local/lib/python3.10/dist-packages/statsmodels/tsa/st
warn('Non-stationary starting autoregressive parameters'
/usr/local/lib/python3.10/dist-packages/statsmodels/tsa/st
warn('Non-invertible starting MA parameters found.'
```

SARIMAX Results					
<b>Dep. Variable:</b>	y	<b>No. Observations:</b>	279		
<b>Model:</b>	ARIMA(1, 0, 2)	<b>Log Likelihood</b>	-1117.943		
<b>Date:</b>	Sat, 16 Sep 2023	<b>AIC</b>	2245.885		
<b>Time:</b>	15:14:14	<b>BIC</b>	2264.041		
<b>Sample:</b>	06-01-2019	<b>HQIC</b>	2253.168		
	- 03-05-2020				
<b>Covariance Type: opg</b>					
	<b>coef</b>	<b>std err</b>	<b>z</b>	<b>P&gt; z </b>	<b>[0.025 0.975]</b>
<b>const</b>	68.0949	3.441	19.789	0.000	61.350 74.839
<b>ar.L1</b>	0.9345	0.061	15.337	0.000	0.815 1.054
<b>ma.L1</b>	-0.5690	0.086	-6.635	0.000	-0.737 -0.401
<b>ma.L2</b>	-0.2142	0.063	-3.416	0.001	-0.337 -0.091
<b>sigma2</b>	176.7477	13.629	12.968	0.000	150.035 203.461
<b>Ljung-Box (L1) (Q):</b>	0.04	<b>Jarque-Bera (JB):</b>	49.97		
<b>Prob(Q):</b>	0.85	<b>Prob(JB):</b>	0.00		
<b>Heteroskedasticity (H):</b>	2.00	<b>Skew:</b>	-0.81		
<b>Prob(H) (two-sided):</b>	0.00	<b>Kurtosis:</b>	4.30		

Figure 29:Synthèse du modèle[9]

Visualisation graphique du résultat(modèle arma) et les valeurs réelles du série temporelle:



**Figure 30:Représentation graphique du modèle ARMA [9]**

==> On constate que les valeurs réelles (en bleu) et notre modèle (en rouge) sont presque confondus.

Ce modèle ARMA avec  $(p, d, q) = (1, 0, 2)$  présente le meilleur résultat qu'on peut avoir après le choix des paramètres et des modèles ARMA.

## 2.4.Les résidus:

Représenter les résidus d'un modèle ARMA sous forme graphique est une étape cruciale dans l'analyse des séries temporelles. Les résidus, également appelés erreurs de prédiction, représentent la différence entre les valeurs réelles de la série et les valeurs prédites par le modèle ARMA. L'analyse des résidus peut fournir des informations importantes sur la qualité du modèle et aider à identifier des modèles potentiels pour une amélioration.

Un modèle ARMA bien ajusté devrait générer des résidus qui ressemblent à un bruit blanc, c'est-à-dire des résidus aléatoires sans structure temporelle ou de corrélation significative. Si les résidus présentent des schémas ou des déviations par rapport à ces attentes, cela peut indiquer que le modèle ARMA actuel ne capture pas correctement la structure de la série temporelle, ce qui peut nécessiter des ajustements.

Alors,l'analyse des résidus sous forme graphique est un outil essentiel pour évaluer la qualité d'un modèle ARIMA et identifier d'éventuelles améliorations nécessaires. Cela permet de s'assurer que le modèle produit des prédictions fiables et cohérentes.

Extraction des résidus :

```
import pandas as pd
import numpy as np
import statsmodels.api as sm

#Extraction des résidus
residus = arma.resid

# Afficher les résidus
print("Résidus du modèle ARMA :")
print(residus)
```

Résidus du modèle ARMA :

2019-06-01	1.405119
2019-06-02	11.562722
2019-06-03	8.491036
2019-06-04	3.878730
2019-06-05	3.266952
...	
2020-03-01	21.832185
2020-03-02	13.616101
2020-03-03	13.586214
2020-03-04	7.026859
2020-03-05	-21.555409

Freq: D, Length: 279, dtype: float64

**Figure 31:Résidus [9]**

Un modèle ARMA est considéré comme adéquat lorsque ses résidus présentent quelques caractéristiques, parmi d'elles la stationnarité des résidus ce qui signifie qu'ils ne présentent pas de tendance globale ni de variation systématique dans le temps.

Vérification de la stationnarité des résidus :

```
[45] import statsmodels.api as sm

resultat_adf = sm.tsa.adfuller(residus)

# La statistique du test ADF
statistique_adf = resultat_adf[0]
# La p-valeur du test ADF
p_valeur_adf = resultat_adf[1]

if p_valeur_adf < 0.05:
    print("Les résidus sont stationnaires (p-valeur =", p_valeur_adf, ")")
else:
    print("Les résidus ne sont pas stationnaires (p-valeur =", p_valeur_adf, ")")
```

Les résidus sont stationnaires (p-valeur = 1.3929377152806235e-29 )

Donc, les résidus sont stationnaires ce qui favorise que notre modèle est le meilleur modèle.

**Figure 32: Vérification de stationnarité [9]**

```
moyenne_residus = np.mean(residus)
print("Moyenne des résidus :", moyenne_residus)
```

Moyenne des résidus : -0.03916958106261983

Donc, les résidus sont stationnaires ce qui favorise que notre modèle est le meilleur modèle.

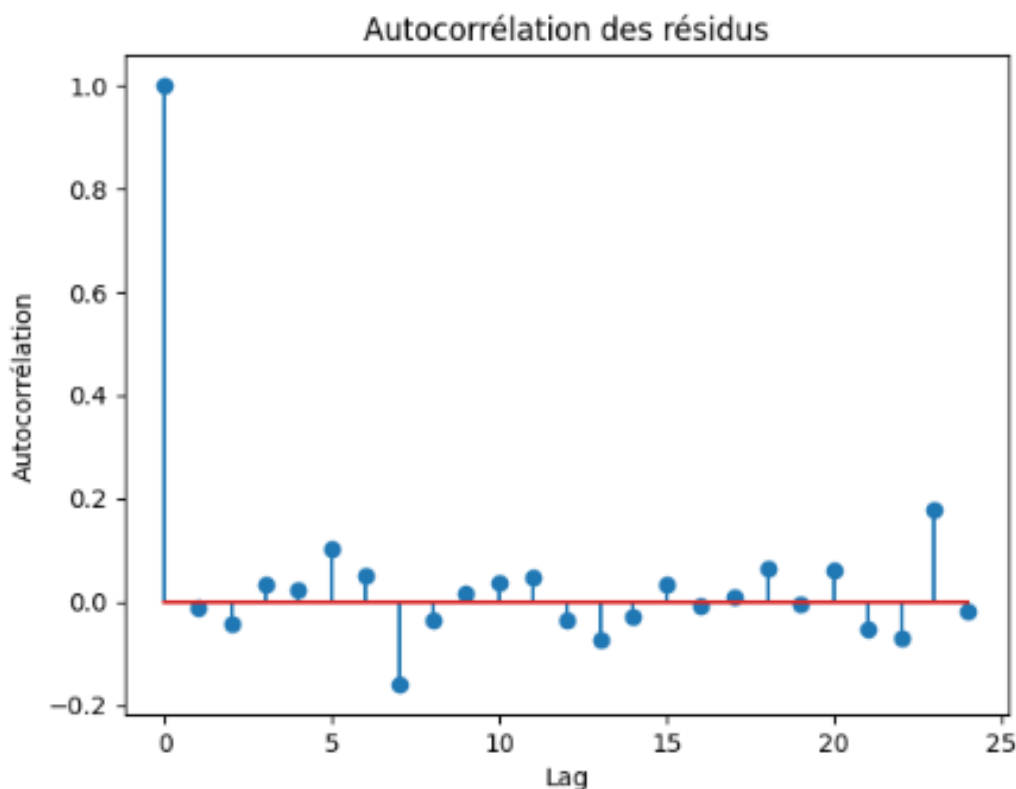
**Figure 33: Moyenne des résidus [9]**

## L'auto-corrélation des résidus :

L'autocorrélation est un concept essentiel à prendre en compte lors de la modélisation de séries temporelles avec des modèles ARIMA (AutoRegressive Integrated Moving Average). L'autocorrélation se réfère à la corrélation des valeurs d'une série avec elles-mêmes à différents décalages temporels.

```
[52] autocorr_residus = sm.tsa.acf(arma.resid)

# Tracé de l'autocorrélation
plt.stem(autocorr_residus)
plt.title("Autocorrélation des résidus")
plt.xlabel("Lag")
plt.ylabel("Autocorrélation")
plt.show()
```



**Figure 34: Autocorrelation des résidus [9]**

En voyant l'auto-corrélation, on constate que la majorité est très proche de 0. D'après de tous les observations précédentes, on conclut que notre modèle (ARMA avec  $(p, d, q) = (1, 0, 2)$ ) est le meilleur modèle.

l'autocorrélation a une influence significative sur le modèle ARIMA en aidant à identifier les ordres du modèle, en évaluant la qualité de la spécification du modèle, en optimisant la précision des prédictions et en évitant le surajustement.

L'analyse de l'autocorrélation est une étape essentielle dans le processus de modélisation des séries temporelles avec ARIMA.

### **3.Conclusion:**

En fin de compte, la modélisation avec le modèle ARMA est une compétence précieuse pour analyser et prévoir des données temporelles. Cependant, il est essentiel de comprendre les limites du modèle ARMA et de sélectionner le modèle approprié en fonction des caractéristiques de la série temporelle. Nous avons évalué et critiqué les algorithmes utilisés dans notre système. Basé sur les résultats de l'évaluation, nos objectifs commerciaux ont été atteints et l'entreprise est satisfaite.

# Conclusion générale

Aujourd'hui, la science des données est le domaine où l'on investit le plus dans l'informatique de la science et la technologie, notamment en termes de recherche et de développement. C'est pour cette raison pour laquelle un grand nombre continu d'entreprises se tournent vers cette science et travaillent à intégrer dans les différents départements de leurs domaines de travail, afin de se donner de meilleures chances d'être leaders sur leurs marchés. Avec la volatilité actuelle du marché, l'intégration de l'apprentissage automatique dans les entreprises est obligatoire pour comprendre et prévoir l'opinion/ le comportement du client afin de présenter des services adéquats.

Dans ce projet, nous avons commencé par étudier l'état de l'énergie solaire photovoltaïque et de l'apprentissage de machine. Ensuite, nous avons établi la méthodologie à suivre tout au long du projet et nous avons déterminé les outils techniques à utiliser. Par la suite, nous avons commencé la phase de mise en œuvre de notre système ainsi que la phase de compréhension et de préparation des données. Enfin, nous avons construit notre modèle prédictif, la phase de modélisation utilisant des analyses de séries temporelles et des composantes principales pour assurer la stabilité des résultats. Grâce à cette expérience, j'ai pu approfondir mes connaissances dans les secteurs concernés par le projet et surtout pour mieux comprendre le système et le travail à travers l'exploration des données. C'était aussi l'occasion pour moi d'affiner les capacités en termes de gestion de projet, conception de projet, énergie solaire photovoltaïque et en termes de modélisation des données de la science. Cela nous permet de mieux apprécier la polyvalence et l'intérêt de l'ingénieur.

En outre, le modèle utilisé nous a donné des résultats assez bons, mais de meilleurs résultats auraient été visible avec l'utilisation de Deep Learning avec une base de données plus grande qui couvre plus d'énergie solaire

# Bibliographie