République Tunisienne

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Carthage

Institut National des Sciences Appliquées et de Technologie

RAPPORT STAGE D'ETE

1ère année cycle d'ingénieur

Spécialité : Informatique Industrielle et Automatique

Développé par :

Rayen Ayadi Boukhchina

Prédiction de la production D'énergie solaire à l'aide de l'apprentissage automatique

Supervisé par : **Nour ben Lazrek**

Période de stage : **De 15/06/2024 à 15/07/2024**

Entreprise D'accueil :" Digimytch "

2023/2024

Table des matières

Table des figures

1. Introduction

1.1. Contexte et objectifs du projet

L'énergie solaire est devenue un pilier essentiel dans la transition énergétique mondiale grâce à son caractère durable et renouvelable. Cependant, maximiser l'efficacité des installations photovoltaïques nécessite une compréhension approfondie des facteurs qui influencent leur production d'énergie. La variabilité des conditions météorologiques, telles que l'intensité lumineuse, la température, et l'humidité, joue un rôle déterminant dans le rendement des panneaux solaires.

Dans ce contexte, le projet vise à développer un modèle prédictif basé sur le machine learning pour estimer la production en courant des panneaux solaires à partir de données environnementales collectées. En prédisant cette production avec précision, il devient possible d'optimiser l'exploitation des installations, de planifier les ressources et d'anticiper d'éventuels problèmes techniques.

L'objectif principal est de concevoir un système de prédiction performant et automatisé, permettant d'assister les gestionnaires dans la prise de décisions stratégiques liées à la gestion des parcs solaires. Ce modèle devra offrir des prédictions précises, tout en s'adaptant aux variations des données d'entrée, garantissant ainsi sa robustesse dans un contexte industriel.

1.2. Objectifs du stage et de l'étude

Ce stage a pour but de concevoir et de mettre en œuvre un pipeline complet de machine learning dédié à la prédiction de la production des panneaux solaires.

Les objectifs spécifiques sont :

- **Nettoyer et préparer les données** : Effectuer un prétraitement des données collectées pour corriger les incohérences, gérer les valeurs manquantes et standardiser les variables.
- **Créer des features pertinentes** : Utiliser le feature engineering pour enrichir les données en créant des variables telles que la production journalière ou des indicateurs météorologiques combinés.

- **Tester plusieurs algorithmes** : Comparer différentes techniques de machine learning (comme XGBoost, Random Forest, et Régression Linéaire) et évaluer leurs performances à l'aide de métriques comme le RMSE et le MSE.
- **Optimiser le modèle retenu** : Ajuster les hyperparamètres du modèle sélectionné afin d'obtenir les meilleurs résultats possibles.
- **Valider et interpréter les résultats** : S'assurer que les prédictions du modèle sont fiables et analyser les facteurs influençant le plus la production des panneaux solaires.

Ce projet s'inscrit dans une démarche visant à démontrer l'utilité des outils de machine learning dans le secteur des énergies renouvelables, tout en apportant des solutions concrètes pour améliorer la gestion et l'exploitation des installations solaires.

2.Présentation de l'entreprise d'accueil

Digimytch est une startup fondée en 2023, spécialisée dans le développement professionnel et le coaching. Elle se consacre à la formation et à l'accompagnement des individus cherchant à développer leurs compétences techniques et comportementales. Avec une équipe de 11 à 20 employés et des partenaires spécialisés dans divers domaines tels que le design graphique, la conception UI/UX, le marketing digital, la photographie, la création de contenu, et bien d'autres, Digimytch se positionne comme un acteur clé dans le secteur de la transformation numérique et de la formation. Récemment, Digimytch a ouvert une nouvelle branche dédiée aux énergies renouvelables, dans le but d'étendre ses activités vers des projets d'innovation durable. Cette initiative s'inscrit dans un contexte global de transition énergétique, où la demande pour des solutions écologiques et performantes dans le secteur des énergies renouvelables, notamment l'énergie solaire, est en pleine croissance. C'est dans ce cadre que le projet de prediction de la production de l'electricité dans un champ de panneaux solaires a été développé.

figure2.1: logo Digimytch

3.Objectifs visés (cahier des charges)

3.1 Objectifs et problématiques à résoudre:

L'objectif principal de ce projet est de développer un modèle de machine learning capable de résoudre des problématiques spécifiques et complexes dans le cadre des services personnalisés offerts par **Digimytch**. Le projet vise à tirer parti des techniques avancées d'apprentissage automatique pour analyser, prédire, et automatiser divers processus, contribuant ainsi à l'efficacité et à la compétitivité des solutions proposées.

Les problématiques clés à résoudre incluent :

Préparation et traitement des données : Collecter, nettoyer et organiser efficacement les données disponibles afin qu'elles soient exploitables pour entraîner les modèles de machine learning.

Choix et optimisation des algorithmes : Identifier les algorithmes les plus adaptés aux problèmes spécifiques abordés (classification, régression, ou clustering) et les ajuster pour maximiser leur performance.

Déploiement du modèle : Intégrer le modèle dans un environnement opérationnel pour qu'il soit utilisable par les équipes ou les clients de Digimytch de manière simple et efficace.

Visualisation et interprétation des résultats : Développer des tableaux de bord intuitifs pour présenter les insights générés par le modèle, permettant une prise de décision éclairée.

Scalabilité et généralisation : Assurer que le modèle puisse traiter des volumes croissants de données et s'adapter à des scénarios variés sans perte significative de performance.

Grâce à ce projet, Digimytch pourra renforcer son expertise en machine learning, étendre son portefeuille de services technologiques et fournir des solutions toujours plus innovantes à ses clients.

3.2. Technologies et outils utilisés

Pour réaliser ce projet de machine learning, plusieurs technologies et outils logiciels ont été utilisés pour garantir une analyse précise et efficace des données. Ces outils ont été sélectionnés en raison de leur robustesse, flexibilité et large adoption dans le domaine de la data science et de l'apprentissage automatique.

Python : Langage principal utilisé pour le développement, connu pour sa syntaxe simple et sa vaste bibliothèque de modules dédiés à la science des données et au machine learning.

Pandas : Librairie utilisée pour la manipulation et l'analyse de données. Elle permet de nettoyer, transformer et explorer les ensembles de données grâce à ses structures puissantes telles que les DataFrames.

NumPy : Utilisée pour le calcul scientifique et la manipulation de données numériques. Elle facilite le traitement rapide des matrices et des tableaux, indispensables dans les algorithmes de machine learning.

scikit-learn : Librairie essentielle pour le machine learning, offrant une implémentation facile et performante de nombreux algorithmes tels que la régression, la classification et le clustering. Elle a également été utilisée pour le prétraitement des données et l'évaluation des modèles.

TensorFlow : Framework de deep learning qui a permis de concevoir, entraîner et optimiser les modèles d'apprentissage profond utilisés dans ce projet. TensorFlow est particulièrement adapté pour travailler avec de grandes quantités de données et des réseaux neuronaux complexes.

Matplotlib et Seaborn : Librairies de visualisation de données utilisées pour créer des graphiques et visualiser les résultats obtenus à différentes étapes du projet, facilitant ainsi l'interprétation des modèles et la présentation des résultats.

Ces outils ont été combinés pour créer un pipeline complet de machine learning, allant de la collecte et du nettoyage des données à l'entraînement et au déploiement des modèles, en

passant par l'analyse et la visualisation des résultats. Leur intégration fluide a permis de garantir la fiabilité et la précision des solutions développées au sein de Digimytch.

figure3.1: logo Python

figure3.2: logo Pandas

figure3.3: logo NumPy

figure3.4: logo scikit-learn

figure3.5: logo TensorFlow

3.3. Critères de performance et de validation

Pour garantir que le modèle de machine learning développé réponde aux objectifs fixés, plusieurs critères de performance et de validation ont été définis. Ces critères permettent d'évaluer la qualité des résultats, la robustesse du modèle et son adéquation avec les besoins de l'entreprise.

Précision des prédictions :
Le modèle doit atteindre un niveau de précision adéquat en termes de métriques telles que le *Mean Absolute Error* (MAE), le *Root Mean Square Error* (RMSE), ou d'autres métriques adaptées aux données et aux objectifs du projet. Une attention particulière est portée à la minimisation des erreurs sur les cas critiques.

Fiabilité du modèle :
Le modèle doit être stable et robuste face à des données variées, notamment en cas de légères anomalies ou variations dans les données d'entrée. Les performances ne doivent pas se dégrader de manière significative lorsqu'il est testé sur des ensembles de données hors échantillon (out-of-sample).

Temps de traitement :
Les temps de prétraitement des données, d'entraînement et d'inférence doivent être optimisés pour garantir une exécution fluide. Idéalement, le modèle doit être utilisable en quasi-temps réel dans un contexte opérationnel.

Interprétabilité des résultats :
Les résultats du modèle doivent être compréhensibles et exploitables par les parties prenantes. Cela inclut des visualisations claires, des explications sur l'importance des différentes caractéristiques (*feature importance*), et des graphiques explicatifs.

Flexibilité et scalabilité :
Le système doit être modulaire et facilement extensible pour intégrer de nouvelles données ou modèles, ou pour répondre à d'autres problématiques liées à la machine learning. La

scalabilité doit permettre une augmentation du volume de données ou une amélioration des algorithmes sans refonte complète.

4. Journal de stage

4.1. Diagramme de Gantt et suivi du projet:

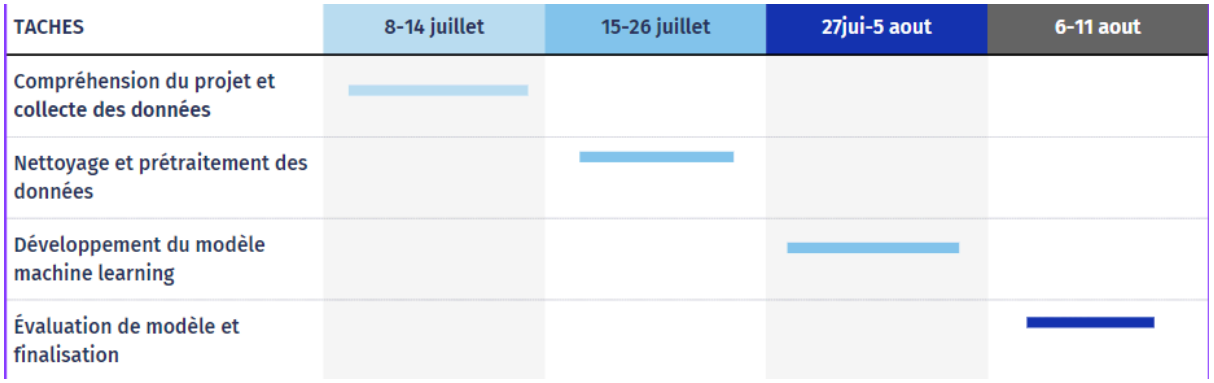

figure4.1: Diagramme de Gantt

5.Travail réalisé

Le travail effectué dans le cadre de ce projet s'est articulé autour de plusieurs étapes clés, visant à transformer les données brutes en un modèle prédictif performant capable d'estimer la production solaire en fonction de divers facteurs. Voici une description détaillée des différentes étapes :

5.1. Préparation et nettoyage des données

La première étape a consisté à préparer les données pour garantir leur qualité et leur pertinence dans le processus d'entraînement du modèle.

- **Gestion des valeurs manquantes :**

  Les données brutes contenaient des valeurs manquantes, notamment pour des variables météorologiques et temporelles. Ces valeurs ont été remplacées par des méthodes

adaptées, telles que l'interpolation ou l'utilisation de moyennes, afin de conserver la cohérence du dataset.

- **Création de nouvelles variables (Feature Engineering) :**

  Une étape importante a été l'ajout de nouvelles variables, comme la production quotidienne d'énergie (daily production), calculée à partir des données existantes. Cette variable a simplifié l'entraînement du modèle en condensant les informations sur la production totale par jour, ce qui a permis de mieux capturer les tendances.

- **Normalisation des données temporelles :**

  Les heures et dates ont été transformées pour être utilisées efficacement par le modèle. Par exemple, des indicateurs comme la saison (hiver, été) ou la fraction de la journée (matin, après-midi) ont été extraits pour enrichir l'analyse.

- **Encodage des variables catégoriques :**

  Les variables qualitatives, telles que les conditions météorologiques (ensoleillé, nuageux, pluvieux), ont été converties en valeurs numériques, facilitant ainsi leur utilisation dans les modèles de Machine Learning.

	Date	temp	wind	humidity	barometer	Sunrise	Sunset	Daily Production	weather_cond
0	2012-01-01	12.520000	19.640000	89.440000	1008.920000	05:30:39.872985	18:29:20.127015	0.5	Cloudy
1	2012-01-02	7.880000	17.200000	78.640000	1008.560000	05:30:46.619181	18:29:13.380819	0.8	Passing clouds
2	2012-01-03	9.230769	38.730769	81.769231	1005.307692	05:30:53.884941	18:29:06.115059	2.9	Partly sunny
3	2012-01-04	7.461538	26.384615	67.730769	1016.000000	05:31:01.668112	18:28:58.331888	0.8	Scattered clouds
4	2012-01-05	8.863636	40.500000	75.227273	994.727273	05:31:09.966389	18:28:50.033611	2.7	Partly sunny

figure5.1: aperçu du dataset après la phase de nettoyage et prétraitement des données

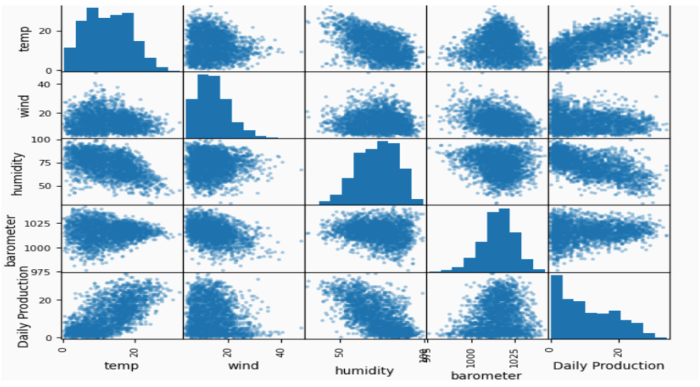

figure5.2: correlation entre les variables

5.2. Sélection et entraînement des modèles

Plusieurs modèles ont été testés afin de trouver celui qui répondait le mieux aux objectifs du projet.

- **Exploration de différents algorithmes :**
  Divers algorithmes de Machine Learning, dont les forêts aléatoires (*Random Forest*), les régressions linéaires, et le modèle **XGBoost**, ont été comparés. Chaque modèle a été évalué selon sa capacité à minimiser les erreurs de prédiction, mesurées par des métriques telles que le **RMSE (Root Mean Squared Error)** et le **MSE (Mean Squared Error)**.
- **Optimisation des hyperparamètres :**
  Pour chaque modèle, les paramètres ont été ajustés afin d'obtenir les meilleures performances possibles. Cela a permis d'améliorer la précision des prédictions en adaptant le comportement du modèle aux spécificités des données.
- **Sélection finale du modèle :**
  Parmi tous les modèles testés, l'algorithme **XGBoost** a été retenu, car il a donné les meilleurs résultats, comme indiqué dans la figure d'évaluation des performances (figure 5.2). Sa capacité à capturer des relations complexes entre les variables, tout en maintenant une précision élevée, a fait de XGBoost le choix optimal pour ce projet.

```
Tuning RandomForest...
RandomForest - MAE: 4.3882, MSE: 34.6885, RMSE: 5.8897
Tuning LinearRegression...
LinearRegression - MAE: 4.4401, MSE: 32.8513, RMSE: 5.7316
Tuning DecisionTree...
DecisionTree - MAE: 5.1714, MSE: 48.9189, RMSE: 6.9942
Tuning GradientBoosting...
GradientBoosting - MAE: 4.3913, MSE: 34.0401, RMSE: 5.8344
Tuning XGBoost...
XGBoost - MAE: 4.3340, MSE: 33.6249, RMSE: 5.7987
```

figure5.3: Évaluation des performances des différents modèles

5.3. Validation et test du modèle

Une fois le modèle sélectionné, il a été soumis à des tests rigoureux pour évaluer sa robustesse et sa généralisation.

- **Validation croisée :**

  La validation croisée a permis de vérifier que le modèle restait performant sur différents sous-ensembles du dataset. Cela garantit qu'il ne s'agit pas d'un modèle trop spécifique (*overfitting*), mais qu'il est capable de généraliser à de nouvelles données.

- **Évaluation sur le dataset de test :**

  Le modèle a été testé sur des données non utilisées lors de l'entraînement pour évaluer sa capacité à faire des prédictions précises dans un contexte réel. Les résultats ont confirmé que le modèle retenu est fiable et adapté au problème.

5.4. Visualisation et interprétation des résultats

L'analyse finale des résultats a permis de tirer des conclusions utiles et exploitables.

- **Graphiques de performance :**

  Des graphiques ont été générés pour comparer les prédictions du modèle avec les données réelles. Ces visualisations ont mis en évidence la précision des prédictions et ont aidé à identifier d'éventuels écarts ou tendances inattendues.

- **Importance des variables :**

  Une analyse des facteurs les plus influents dans les prédictions a montré que des variables comme les conditions météorologiques, les heures de la journée, et les saisons jouaient un rôle clé. Cette analyse aide à mieux comprendre le comportement du modèle et son adéquation avec les données.

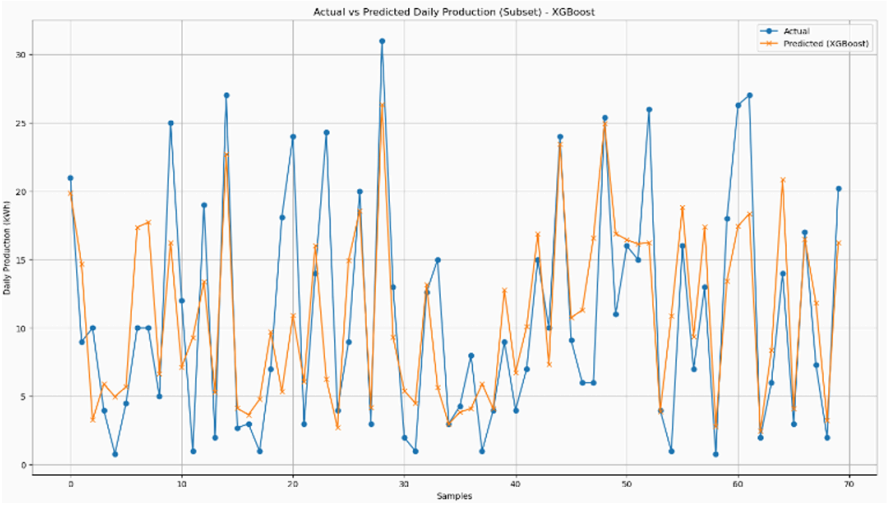

figure 5.4: visualisation de la performance du modéle choisi (xgboost)

6.Consolidation des acquis

Compétences acquises a l'INSAT	Compétences acquises lors du stage
Probabilités et statistiques algorithmiques et structures de données resolution de problèmes travail en équipe	Python Machine Learning et Data Science gestion du temps et respect des deadlines

7. Conclusion générale

7.1. Bilan du projet et résultats obtenus

Le projet de prédiction de la production des panneaux solaires a permis d'atteindre les objectifs fixés, démontrant l'efficacité des approches de machine learning dans le domaine des énergies renouvelables. Les principales étapes, de la préparation des données à l'entraînement des modèles prédictifs, ont été réalisées avec succès.

- **Nettoyage et préparation des données** : Les données brutes ont été nettoyées et enrichies grâce au feature engineering, permettant de créer des variables comme la production journalière, essentielles pour améliorer la performance des modèles.
- **Développement des modèles** : Plusieurs algorithmes, dont XGBoost, Random Forest et Régression Linéaire, ont été testés et comparés. XGBoost s'est révélé être le modèle le plus performant en termes de précision, avec des scores optimaux sur des métriques comme le RMSE et le MSE.
- **Validation des résultats** : Les prédictions obtenues ont démontré une corrélation significative avec les données réelles, prouvant la fiabilité du modèle dans un contexte industriel.

Ce projet a ainsi permis de concevoir un système prédictif robuste, offrant des solutions concrètes pour optimiser l'exploitation des panneaux solaires.

7.2. Points forts et points faibles du projet

Points forts :

- **Précision du modèle** : L'utilisation de XGBoost a permis d'obtenir des prédictions précises, répondant aux attentes du projet.
- **Automatisation complète** : Le pipeline mis en place permet un traitement automatisé des données, de la collecte à l'analyse, réduisant les interventions humaines.
- **Compétences développées** : Le projet a permis de renforcer des compétences en machine learning, en gestion de données et en optimisation de modèles.

Points faibles :

- **Problèmes initiaux liés aux données** : La qualité des données collectées a nécessité un travail important de nettoyage et de préparation, ce qui a ralenti les premières étapes du projet.
- **Limitation des modèles explorés** : Le temps imparti n'a pas permis d'explorer davantage d'approches innovantes, comme des modèles neuronaux ou des techniques avancées d'ensemble learning.
- **Impact des données manquantes** : Certaines valeurs manquantes ont été imputées, mais cela pourrait avoir influencé légèrement les performances du modèle.

7.3. Perspectives d'avenir et améliorations possibles

Ce projet ouvre la voie à de nombreuses améliorations pour renforcer encore plus la précision et l'efficacité du système de prédiction de la production solaire :

1. **Extension des données collectées** : Intégrer davantage de variables météorologiques ou techniques pour enrichir les données d'entrée et affiner les prédictions.
2. **Intégration de l'intelligence artificielle** : Explorer des architectures avancées, comme les réseaux neuronaux profonds (Deep Learning), pour modéliser des relations plus complexes entre les variables.

3. **Amélioration de la visualisation** : Développer des tableaux de bord interactifs pour offrir une meilleure compréhension des prédictions et faciliter la prise de décisions.
4. **Analyse prédictive sur le long terme** : Étendre le modèle pour fournir des prévisions à long terme, permettant une meilleure planification des ressources et de la maintenance.

En conclusion, ce projet a démontré la pertinence des outils de machine learning dans le domaine des énergies renouvelables, et a contribué à la mise en place d'une solution innovante pour optimiser la gestion des installations solaires. Les compétences acquises et les résultats obtenus serviront de base solide pour relever de futurs défis dans la data science et l'énergie durable.