Institut International d'Ingénierie de l'Eau et de l'Environnement

Machine Learning
Rapport

Élèves :
Aïman KONE
Héléna NIGNAN
Daniela ZOUNGRANA

Encadrants :
FABRICE SAWADOGO

28 juin 2025

Table des matières

1 Introduction Générale

Dans un contexte mondial marqué par les défis liés au changement climatique et à la gestion durable des ressources naturelles, l'agriculture reste l'un des secteurs les plus vulnérables, notamment dans les pays en développement. Au Burkina Faso, la dépendance à l'agriculture pluviale, la faible mécanisation, et le manque d'accès à des technologies modernes d'aide à la décision limitent considérablement la productivité agricole. Face à ces enjeux, l'intégration des technologies numériques, telles que l'Internet des objets (IoT) et l'Intelligence Artificielle (IA), offre de nouvelles perspectives pour améliorer les pratiques agricoles et optimiser les rendements.

Ce projet s'inscrit dans cette dynamique d'innovation au service de l'agriculture durable. Il vise à développer un système intelligent de gestion de l'arrosage basé sur la collecte de données environnementales via des capteurs Arduino et leur analyse à l'aide d'algorithmes de Machine Learning. L'objectif est de prédire de manière précise la quantité d'eau nécessaire et le moment optimal pour arroser une plante donnée. Ce système, conçu pour être simple, peu coûteux et facilement réplicable, a pour ambition de répondre aux besoins réels des petits exploitants burkinabè tout en valorisant les compétences techniques acquises durant notre formation.

2 Contexte et Problématique

Au Burkina Faso, les producteurs agricoles font face à des conditions climatiques difficiles, caractérisées par des sécheresses fréquentes, une répartition irrégulière des pluies et une forte évapotranspiration. Dans ce contexte, l'arrosage manuel constitue encore la méthode la plus courante d'irrigation. Cependant, cette méthode repose souvent sur l'intuition des cultivateurs, ce qui conduit à des excès ou à des insuffisances d'eau, compromettant la santé des plantes et gaspillant une ressource déjà précieuse.

Ce manque de précision dans l'arrosage impacte non seulement la croissance des cultures, mais aussi la durabilité des exploitations agricoles. Il devient donc urgent d'introduire des outils technologiques accessibles, capables d'aider les agriculteurs à prendre des décisions éclairées quant à l'utilisation de l'eau.

Dans cette optique, nous avons choisi de nous concentrer sur la culture de la salade, plante particulièrement sensible au stress hydrique et aux variations climatiques. En installant un système de capteurs (température, humidité, lumière, etc.) dans un espace contrôlé, nous allons récolter des données sur les conditions environnementales entourant la plante. Ces données permettront ensuite d'entraîner un modèle de Machine Learning capable de prédire deux éléments cruciaux :

— La quantité d'eau nécessaire pour assurer une croissance optimale de la salade.
— Le moment de la journée le plus approprié pour effectuer l'arrosage.

Notre problématique principale peut ainsi se formuler comme suit :

> **Comment prédire, à partir de données environnementales en temps réel, la quantité d'eau à fournir et le moment idéal pour arroser une plante, afin de garantir une irrigation intelligente, économe et adaptée aux conditions locales ?**

3 Objectifs du Projet

Ce projet a pour ambition de concevoir un système intelligent d'assistance à l'arrosage pour les cultures maraîchères, en particulier la salade, en se basant sur l'analyse de données environnementales. Il s'agit d'une approche à la fois technologique et pragmatique, pensée pour être utile aux exploitants agricoles dans un contexte local à faibles ressources.

Les objectifs spécifiques que nous visons sont les suivants :

— **Collecter des données pertinentes sur l'environnement immédiat des cultures** : température ambiante, humidité du sol, luminosité, hygrométrie de l'air, etc.

— **Concevoir un système de collecte automatisée** basé sur une carte Arduino et un ensemble de capteurs fiables et peu coûteux, capable d'enregistrer les données à des intervalles réguliers tout au long de la journée.

— **Exploiter ces données pour entraîner des modèles de Machine Learning**, afin de prédire à la fois la quantité d'eau optimale à fournir et le moment le plus opportun pour arroser les cultures.

— **Évaluer la performance des modèles** à l'aide de métriques d'évaluation telles que l'erreur quadratique moyenne (RMSE), la précision temporelle et la fiabilité des prédictions.

— **Proposer un prototype de solution d'irrigation intelligente**, facilement réplicable et adaptable à d'autres cultures ou environnements, tout en respectant les contraintes du terrain.

— **Développer une application web pour visualiser et gérer les données récoltées**, permettant aux utilisateurs d'accéder aux rapports générés par les modèles et de recevoir des alertes personnalisées pour optimiser l'irrigation.

À travers ces objectifs, notre projet souhaite démontrer que des solutions simples, basées sur l'intelligence artificielle et des composants open-source, peuvent réellement améliorer la productivité agricole de manière durable et économique.

4 Méthodologie

Notre démarche méthodologique s'articule en plusieurs phases, de la collecte de données à l'analyse prédictive, en passant par le traitement et la modélisation. Voici les grandes étapes prévues dans le cadre de ce projet :

1. **Installation du système de mesure** : nous avons mis en place un dispositif Arduino comportant plusieurs capteurs : un capteur DHT11 pour mesurer la température et l'humidité ambiantes, un capteur d'humidité du sol (YL-69), un capteur de luminosité (LDR), ainsi qu'un capteur de température du sol. Ces capteurs sont positionnés autour de la plante (salade) cultivée dans un espace test.

Début du Montage :

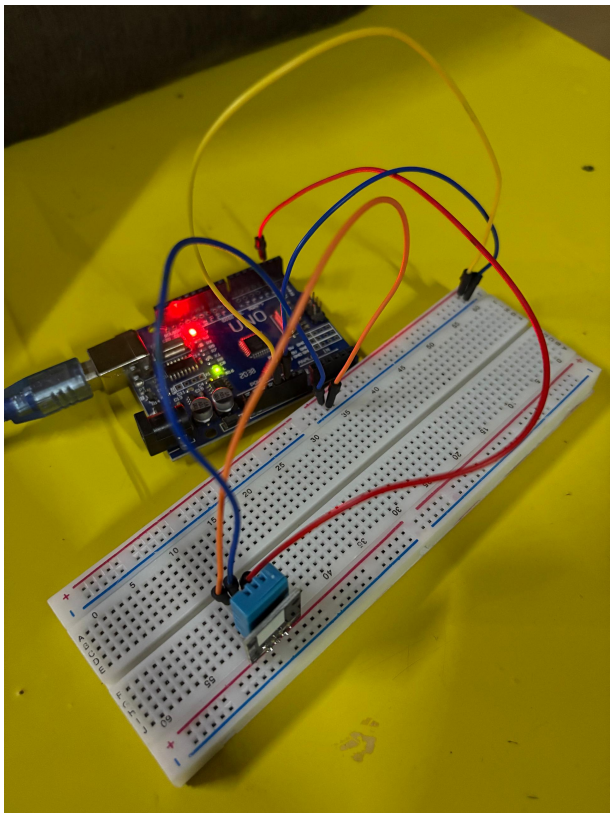

FIGURE 1 – Début du montage du système Arduino avec capteurs

Le montage une fois terminé :

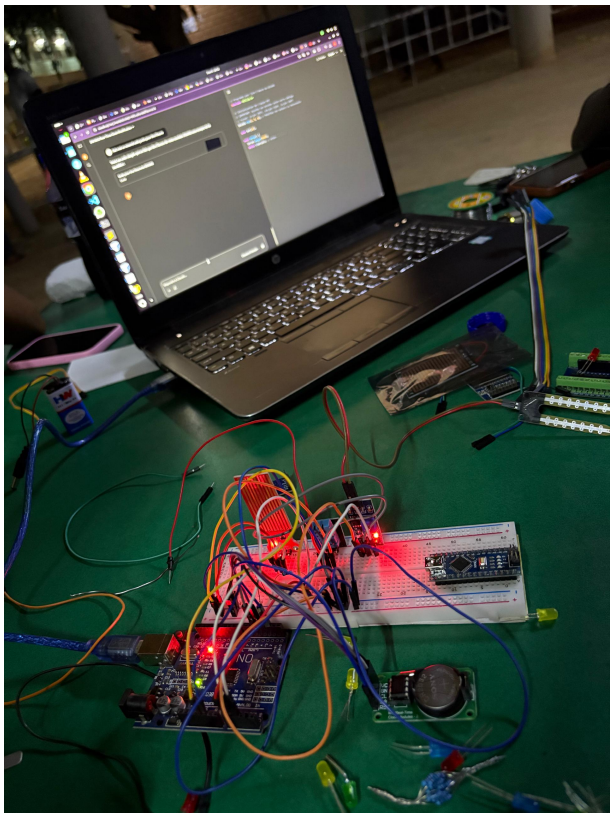

FIGURE 2 – Montage du système Arduino avec capteurs une fois terminé

2. **Collecte des données environnementales** : les données sont collectées automatiquement à intervalles réguliers (toutes les 10 minutes) durant toute la journée, afin de capturer les variations naturelles entre le matin, le midi et le soir. Chaque enregistrement contient l'heure, les valeurs des différents capteurs, et une indication de l'arrosage effectué.

FIGURE 3 – Dispositif installé dans le local avec une graine de salade plantée

3. **Nettoyage et préparation des données** : les données brutes seront exportées au format CSV puis nettoyées (suppression des valeurs manquantes ou aberrantes, standardisation des unités, traitement des doublons). Elles seront ensuite préparées pour l'apprentissage automatique.
4. **Exploration et visualisation des données (EDA)** : nous réaliserons une analyse exploratoire des données (histogrammes, courbes temporelles, corrélations entre variables) afin de mieux comprendre les relations entre les paramètres environnementaux et les besoins en eau de la plante.
5. **Choix et entraînement des modèles de Machine Learning** : nous allons tester différents algorithmes (régression linéaire, forêt aléatoire, LSTM) pour prédire la quantité d'eau et l'heure d'arrosage optimale, en divisant notre jeu de données en ensemble d'entraînement et de test.
6. **Évaluation des modèles et amélioration** : les performances des modèles seront évaluées à l'aide de métriques standard (MAE, RMSE, R2-score) et les hyperparamètres seront ajustés si nécessaire pour améliorer la précision des prédictions.

7. **Interprétation des résultats et conception d'une stratégie d'arrosage intelligente** : à partir des prédictions obtenues, nous proposerons un modèle d'aide à la décision pour l'agriculteur, basé sur un planning d'arrosage quotidien ou en temps réel.

5 Choix des algorithmes et des métriques d'évaluation

Dans le cadre de ce projet, le choix des algorithmes est une étape cruciale pour répondre efficacement à notre problématique, qui consiste à prédire la quantité d'eau nécessaire pour l'arrosage des plantes et à déterminer le moment optimal pour cet arrosage.

Nos données sont constituées de variables numériques continues (température, humidité du sol, humidité de l'air, luminosité), catégorielles (pluie, état de la pompe) et temporelles (heure). Ces caractéristiques nous ont conduits à opter pour des modèles capables de gérer la diversité des types de données, tout en restant robustes face aux contraintes d'un dataset de taille modeste, issu de mesures réelles avec des capteurs Arduino.

Notre choix repose sur plusieurs critères fondamentaux :

— **La robustesse aux données bruitées et incomplètes**, caractéristiques fréquentes dans les données environnementales collectées par des capteurs low-cost.
— **La capacité à modéliser des relations complexes**, qu'elles soient linéaires ou non linéaires, entre les différentes variables.
— **La complémentarité des approches** : des modèles supervisés pour la prédiction (régression), et des modèles non supervisés pour la segmentation (clustering des moments de la journée).
— **La facilité d'implémentation et l'optimisation pour l'intégration mobile**, notamment via Firebase et une interface React Native.

5.1 Régression Linéaire

La Régression Linéaire constitue le modèle de base de toute approche prédictive. Elle établit une relation mathématique directe entre une ou plusieurs variables indépendantes (*features*) et une variable dépendante (*target*). Dans notre cas, la variable cible est la **quantité d'eau nécessaire**.

Pourquoi la Régression Linéaire ?

— C'est un modèle rapide, facile à mettre en œuvre, et offrant une bonne capacité d'interprétation.
— Il permet de quantifier l'impact de chaque variable sur la quantité d'eau : par exemple, savoir si l'humidité du sol a un effet plus marqué que la température ou la luminosité.
— Il sert de référence pour comparer la performance des modèles plus complexes.

Ses limites : elle suppose que la relation entre les variables est linéaire. Or, en environnement agricole, les besoins en eau peuvent dépendre de relations non linéaires (par exemple : un effet de seuil avec l'humidité du sol ou la luminosité).

5.2 Forêt Aléatoire (Random Forest)

La Forêt Aléatoire est un algorithme d'ensemble basé sur la combinaison de plusieurs arbres de décision. Contrairement à la Régression Linéaire, elle permet de capturer des

relations non linéaires et des interactions complexes.

Pourquoi Random Forest ?

— Elle est très robuste aux valeurs aberrantes, au bruit et aux données incomplètes.
— Elle permet de gérer des variables fortement corrélées ou dont l'importance évolue selon les conditions environnementales.
— Elle fournit des indicateurs précieux comme l'importance des variables, ce qui aide à comprendre quels facteurs influencent le plus l'arrosage (exemple : humidité du sol > température > luminosité).
— Elle offre de bonnes performances même avec des datasets de taille réduite, ce qui est le cas dans notre projet.

Applications dans le projet :

— Prédiction de la **quantité d'eau nécessaire**.
— Aide potentielle à la classification des moments d'arrosage (matin, midi, soir) selon les patterns environnementaux.

5.3 K-Means

K-Means est un algorithme non supervisé de clustering, qui regroupe les observations en **k groupes** basés sur la similarité entre les valeurs des variables.

Pourquoi K-Means ?

— Pour segmenter les données selon des patterns naturels : température, humidité, luminosité, heure, etc.
— Pour identifier des groupes correspondant à des moments typiques de la journée : *matin*, *midi* et *soir*.
— Pour structurer la gestion de l'arrosage en fonction de ces groupes.

Intérêt dans notre projet :

— Fournir une aide à la décision temporelle : savoir si les besoins en eau sont plus critiques le matin, le midi ou le soir.
— Adapter dynamiquement l'arrosage en fonction des clusters détectés.

5.4 Métriques d'évaluation

L'évaluation des modèles est essentielle pour mesurer leur pertinence et guider leur sélection.

5.4.1 Pour les modèles de régression

Nous utilisons trois métriques principales :

— **MAE – Erreur Absolue Moyenne (Mean Absolute Error)** : Mesure l'erreur moyenne entre les valeurs prédites et les valeurs réelles. Elle donne une idée claire de la précision moyenne attendue du modèle, sans accorder trop de poids aux grandes erreurs.

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

Avantages : simple à comprendre, intuitive. **Limites :** traite toutes les erreurs de manière égale.

— **RMSE – Racine de l'Erreur Quadratique Moyenne (Root Mean Squared Error)** : Similaire au MAE mais pénalise plus fortement les grandes erreurs.

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

Avantages : sensible aux grosses erreurs, utile pour détecter les cas où le modèle échoue gravement. **Limites :** plus sensible aux valeurs extrêmes.

— **R^2 – Coefficient de Détermination** : Mesure la proportion de la variance des données expliquée par le modèle.

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

Interprétation :

— $R^2 = 1$: le modèle explique parfaitement la variance.
— $R^2 = 0$: le modèle ne fait pas mieux qu'une prédiction basée sur la moyenne.
— $R^2 < 0$: le modèle est pire qu'une prédiction aléatoire.

5.4.2 Pour le clustering (K-Means)

L'évaluation du clustering repose sur des critères qualitatifs et visuels :

— **La cohérence intra-cluster :** les données d'un même cluster doivent être similaires.
— **La séparation inter-cluster :** les clusters doivent être bien distincts les uns des autres.
— **L'interprétation métier :** les clusters doivent correspondre à des segments significatifs (exemple : matin, midi, soir).
— **Visualisation :** l'observation des clusters sur des graphiques (projection sur deux ou trois variables) permet de valider leur cohérence.

Ce choix d'algorithmes et de métriques nous permet d'aborder la problématique sous plusieurs angles, en combinant des approches supervisées pour la prédiction et non supervisées pour la segmentation temporelle. Cette démarche garantit à la fois robustesse, efficacité et interprétabilité dans un contexte agricole local, avec des données collectées en environnement réel.

6 Résultats expérimentaux

L'ensemble des étapes de prétraitement, d'entraînement des modèles, de calcul des performances, ainsi que les visualisations associées sont regroupés dans le fichier *rapport_ analyse_ jardin.ipynb* fourni en annexe.

Ce fichier contient :

— Le code complet de nettoyage et de transformation des données.
— Le processus d'entraînement des trois modèles sélectionnés : Régression Linéaire, Forêt Aléatoire (Random Forest) et K-Means.
— Le calcul et l'interprétation des métriques d'évaluation.
— Les courbes de performances et les visualisations des clusters.

Les performances des modèles de régression sont synthétisées dans le tableau suivant :

Modèle	MAE	RMSE	R^2
Régression Linéaire	509.973	805.632	0.680
Random Forest	122.818	469.109	0.891

TABLE 1 – Performances des modèles de régression sur le jeu de test

Concernant l'algorithme de clustering K-Means, les données ont été segmentées en trois groupes correspondant aux périodes de la journée : *matin*, *midi* et *soir*. Cette segmentation s'appuie sur les variables environnementales telles que l'heure, la température, l'humidité et la luminosité.

Les visualisations des clusters formés par K-Means, ainsi que les courbes de comparaison entre les valeurs réelles et les valeurs prédites pour les modèles de régression, sont disponibles dans le fichier `rapport_analyse_jardin.ipynb`.

L'ensemble des résultats détaillés, y compris le code source, les courbes, et les interprétations étape par étape, est disponible dans le fichier `rapport_analyse_jardin.ipynb` ajouté en annexe.

7 Analyse comparative des modèles

7.1 Comparaison des résultats obtenus

L'analyse des performances montre que le modèle **Random Forest** surpasse très nettement la **Régression Linéaire** sur notre jeu de données. En effet, la Forêt Aléatoire présente un **RMSE beaucoup plus faible (469.1 contre 805.6)** et un **R^2 bien plus élevé (0.891 contre 0.680)**, ce qui indique une bien meilleure capacité de prédiction.

Le **MAE** confirme cette tendance, avec une erreur absolue moyenne de **122.8 pour Random Forest** contre **509.9 pour la Régression Linéaire**, soit une amélioration très significative. Cela démontre que le modèle Random Forest est beaucoup plus robuste face aux variations des données et capable de capturer les relations complexes entre les différentes variables environnementales.

La **Régression Linéaire**, bien qu'elle soit simple et rapide à entraîner, montre ici ses limites. Elle n'est pas capable de modéliser les relations non linéaires présentes dans nos données. Elle reste néanmoins utile comme **modèle de référence**, pour donner une première approximation des besoins en eau et pour comprendre l'impact global des différentes variables de manière linéaire.

En ce qui concerne le clustering, l'algorithme **K-Means** a permis d'identifier de façon pertinente **trois groupes représentatifs correspondant aux moments de la journée : matin, midi et soir**. Cette segmentation est particulièrement utile pour adapter l'arrosage aux différentes phases de la journée, en tenant compte des variations naturelles de la température, de l'humidité du sol, de l'humidité de l'air et de la luminosité.

L'analyse visuelle des clusters montre des regroupements cohérents, qui peuvent être directement utilisés pour optimiser la planification de l'arrosage selon les périodes où les besoins en eau sont les plus importants.

7.2 Choix final du modèle

Au vu des performances obtenues, le modèle **Random Forest** a été retenu pour la prédiction de la **quantité d'eau nécessaire**. Sa robustesse face aux variations des données, sa résistance aux valeurs extrêmes et sa capacité à modéliser des relations non linéaires et complexes en font le candidat le plus adapté à notre problématique.

Comparativement à la Régression Linéaire, Random Forest offre une réduction significative de l'erreur (MAE et RMSE), ainsi qu'un coefficient de détermination $\mathbf{R^2}$ beaucoup plus élevé (**0.891**), ce qui atteste de sa capacité à expliquer la majeure partie de la variance des données.

Concernant la classification temporelle des moments propices à l'arrosage, l'algorithme de clustering **K-Means** a été jugé pertinent et suffisant. Il permet de segmenter efficacement la journée en **trois périodes distinctes : matin, midi et soir**, en tenant compte des variations naturelles de la température, de l'humidité et de la luminosité. Cette approche apporte une aide précieuse pour adapter les recommandations d'arrosage selon les périodes de la journée.

Ainsi, la combinaison de **Random Forest pour la prédiction de la quantité d'eau** et de **K-Means pour la segmentation temporelle** constitue une solution robuste, efficace et adaptée au contexte de notre projet.

8 Conclusion et perspectives

8.1 Conclusion

Ce projet avait pour objectif de concevoir un système intelligent d'assistance à l'arrosage, reposant sur l'analyse de données environnementales collectées en temps réel via une plateforme Arduino. Grâce aux mesures de température, d'humidité de l'air, d'humidité du sol, de luminosité et de l'heure, nous avons pu entraîner et évaluer plusieurs modèles de Machine Learning, afin de prédire de manière optimale les besoins en eau de la plante.

Les résultats obtenus démontrent que le modèle **Random Forest** fournit les meilleures performances pour la prédiction de la **quantité d'eau nécessaire**, avec un coefficient de détermination $\mathbf{R^2}$ **de 0.891**, indiquant une excellente capacité de généralisation. Le modèle de **Régression Linéaire**, bien qu'il soit simple et rapide, reste limité dans la capture des relations non linéaires, et ses performances sont inférieures. Par ailleurs, le **clustering avec K-Means** a permis de segmenter efficacement les conditions environnementales en **trois périodes clés de la journée** : *matin*, *midi* et *soir*, facilitant ainsi l'adaptation de la stratégie d'arrosage selon le moment.

L'intégration des prédictions, des visualisations et des recommandations au sein d'une **application mobile connectée à Firebase** rend notre solution pratique, accessible et directement exploitable pour les agriculteurs. Cette approche offre un outil intelligent pour optimiser l'irrigation, réduire le gaspillage d'eau, et améliorer la gestion des ressources hydriques, particulièrement dans un contexte local comme celui du Burkina Faso.

Ce projet prouve qu'une combinaison de technologies accessibles (Arduino, Firebase) et d'outils d'intelligence artificielle peut contribuer efficacement à moderniser les pratiques agricoles de manière durable et intelligente.

8.2 Perspectives

Pour améliorer ce projet, plusieurs axes de développement sont envisagés :

- **Collecte de données sur une plus longue période** pour enrichir le jeu de données et permettre l'entraînement de modèles plus performants.
- **Mise en œuvre d'un modèle LSTM** pour la gestion des séries temporelles afin d'anticiper les besoins en eau sur plusieurs jours.
- **Extension du système à d'autres types de cultures**, afin de généraliser l'approche au-delà de la salade.
- **Optimisation de l'application mobile** avec un système de notifications intelligentes basé sur les prédictions.
- **Intégration de capteurs supplémentaires** tels que des pluviomètres ou des capteurs de vent pour affiner la prise de décision.
- **Déploiement d'une API ou d'un modèle embarqué** pour rendre la solution totalement autonome, même hors connexion internet.

En conclusion, ce projet démontre qu'il est possible, avec des moyens accessibles et des compétences en data science, de développer une solution intelligente et adaptée aux réalités agricoles du Burkina Faso, contribuant ainsi à une gestion plus durable de l'eau et à l'amélioration des rendements agricoles.

FIGURE 4 – L'intelligence, c'est la capacité de s'adapter au changement. Et parfois, c'est aussi savoir quand et comment arroser une graine d'avenir.