1. Introducción y fundamentos conceptuales

El desarrollo de modelos de crecimiento o productividad de cultivos basados en variables agroclimáticas y sistemas IoT representa una de las áreas más prometedoras de la agricultura de precisión. Sin embargo, la implementación práctica de estos sistemas enfrenta limitaciones importantes, especialmente en contextos donde la infraestructura de sensores in situ es limitada o inexistente.

En este proyecto, se buscó simular un sistema inteligente de monitoreo agrícola que permitiera predecir el crecimiento o la productividad de las plantas a partir de variables agroclimáticas. Aunque la idea original contemplaba el uso de datos generados por sensores IoT reales, se identificaron varios obstáculos técnicos:

- Acceso restringido a sensores agroclimáticos instalados en campo, lo cual impide obtener variables continuas como humedad del suelo, temperatura foliar, conductancia estomática u otras métricas fisiológicas específicas.
- Falta de Datasets extensos y multitemporales, necesarios para entrenar modelos de aprendizaje automático con capacidad de generalización.
- Costos y tiempos asociados a la instalación, calibración y monitoreo de sensores físicos, que dificultan la ejecución experimental en el corto plazo.

Ante esta situación, se adoptó un enfoque alternativo y metodológicamente sólido: la simulación del crecimiento y rendimiento de cultivos mediante modelos fisiológicos como WOFOST, utilizando como entradas datos climáticos diarios obtenidos de fuentes confiables como NASA POWER. Estos modelos permiten replicar el comportamiento fisiológico de una planta en función de variables ambientales clave como radiación solar, temperatura, precipitación, evapotranspiración y viento.

Además, para acelerar el proceso de entrenamiento y evaluación de modelos, se implementó un esquema de procesamiento paralelo que permite ejecutar múltiples experimentos de aprendizaje automático de forma simultánea. Esta estrategia no solo reduce significativamente el tiempo de cómputo en comparación con un flujo secuencial, sino que también facilita la comparación eficiente entre distintos algoritmos y configuraciones. La utilización de bibliotecas como Prefect con ConcurrentTaskRunner hizo posible esta ejecución paralela, optimizando el uso de los recursos computacionales y permitiendo iteraciones más rápidas en el desarrollo del modelo.Esta estrategia presenta ventajas notables:

- Permite generar datasets sintéticos realistas y completamente etiquetados, con diferentes escenarios de manejo o clima.
- Proporciona una base consistente para entrenar modelos de Machine Learning, incluso en ausencia de datos experimentales directos.
- Facilita la evaluación de variables clave para la productividad, identificando los factores más influyentes y apoyando decisiones de manejo optimizado.
- Reduce los tiempos de procesamiento mediante la ejecución paralela, agilizando el análisis exploratorio y la validación cruzada.

En resumen, el uso de modelos fisiológicos como generadores de datos, combinado con algoritmos de aprendizaje automático y procesamiento paralelo, constituye una solución efectiva para simular un sistema IoT inteligente cuando no se dispone de sensores en campo, garantizando la viabilidad técnica del proyecto y su aplicabilidad a escenarios reales.

Objetivo general

Desarrollar un sistema de simulación inteligente para predecir el crecimiento y la productividad de cultivos mediante la integración de datos climáticos diarios obtenidos por fuentes remotas, modelos fisiológicos de simulación y técnicas de aprendizaje automático, con el fin de identificar los factores agroclimáticos más relevantes para la optimización del rendimiento agrícola.

Objetivos específicos

Implementar el modelo fisiológico WOFOST mediante la librería PCSE en Python, simulando el desarrollo diario de un cultivo genérico bajo diferentes condiciones ambientales y obteniendo variables como biomasa acumulada en los frutos.

Entrenar modelos de aprendizaje automático utilizando los datos simulados, con el objetivo de predecir la productividad del cultivo a partir de variables climáticas agregadas o derivadas.

Evaluar la importancia relativa de las variables climáticas en el rendimiento del cultivo simulado, identificando los factores con mayor influencia sobre la productividad como base para estrategias de manejo y optimización.

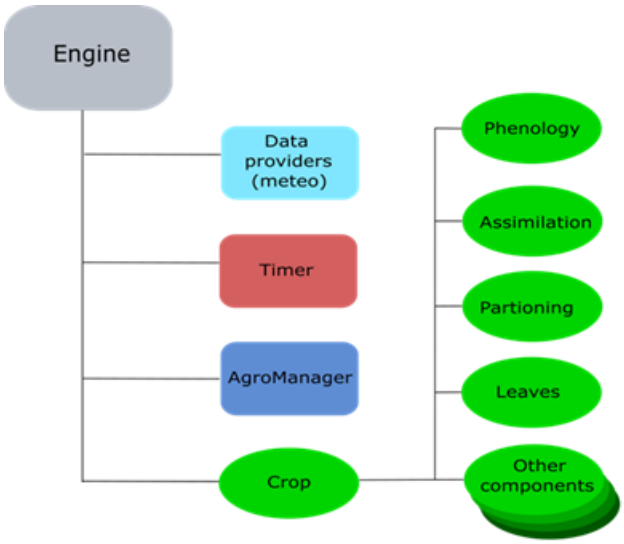
Incorporar un esquema de procesamiento paralelo en el flujo de entrenamiento de modelos, con el fin de reducir los tiempos de ejecución, facilitar la validación cruzada eficiente y permitir la comparación rápida entre diferentes algoritmos de aprendizaje automático.

Marco teórico

Modelo fisiológico de cultivos (WOFOST)

WOFOST (World Food Studies) es un modelo matemático que simula el crecimiento y desarrollo de los cultivos a partir de variables agroclimáticas, fisiológicas y edafológicas. Este modelo permite estimar variables como la biomasa total, la producción de órganos comestibles (como raíces o frutos) y la demanda hídrica del cultivo. Su estructura interna se organiza de forma modular, como se muestra en la Figura 1.

Figura 1. Workflow WOFOST.



El Engine actúa como núcleo del sistema, coordinando el flujo de información entre los distintos componentes. Entre sus entradas clave se encuentran los Data Providers, que suministran los datos meteorológicos diarios, ya sea a partir de archivos propios o conectándose directamente a la API de NASA POWER para extraer datos históricos y actualizados de variables como temperatura, radiación, precipitación y evapotranspiración. Además, el Timer controla el avance del tiempo en la simulación, el AgroManager gestiona las prácticas agrícolas como la siembra y el riego, y el módulo Crop define las características fisiológicas del cultivo.

Estos elementos alimentan una serie de módulos funcionales, como Phenology (desarrollo del cultivo), Assimilation (asimilación de carbono), Partitioning (distribución de biomasa), Leaves (dinámica del área foliar) y Other components, que agrupan otros procesos biológicos. Esta arquitectura modular permite representar de forma detallada y flexible el comportamiento fisiológico del cultivo bajo distintas condiciones ambientales y de manejo.

Contenido volumétrico de humedad del suelo
Es la cantidad de agua contenida en un volumen dado de suelo, expresada como cm^3 de agua por cm^3 de suelo. Este valor varía entre tres puntos clave: saturación (SM0), capacidad de campo (SMFCF) y punto de marchitez (SMW), los cuales determinan la disponibilidad de agua para las plantas.

- **Saturación (SM0)**
  Es el contenido máximo de agua que puede contener el suelo cuando todos los poros están completamente llenos de agua**.** En este punto no hay espacio para aire, y el exceso de agua comienza a drenar por gravedad. No toda el agua en este estado está disponible para las plantas.

- **Capacidad de campo (SMFCF)**
  Es la cantidad de agua que permanece en el suelo después del drenaje gravitacional, es decir, cuando el agua libre ya ha salido y solo queda la que está retenida en los microporos. Representa el nivel óptimo de humedad para el crecimiento vegetal, ya que es fácilmente aprovechable por las raíces.
- **Punto de marchitez permanente (SMW)**
  Es el nivel de humedad en el que las plantas ya no pueden extraer agua suficiente para mantener sus funciones vitales. Aunque el suelo aún contiene algo de agua, está retenida con una fuerza tan alta que las raíces no pueden absorberla, provocando el marchitamiento irreversible.

pF y curva de retención de agua
El pF representa el logaritmo de la succión que el suelo ejerce sobre el agua (en cm de columna de agua). A medida que aumenta el pF, el agua está más retenida y es más difícil de extraer por las raíces. La curva pF-humedad permite visualizar la disponibilidad hídrica del suelo en función de su textura.

Conductividad hidráulica
Es la capacidad del suelo para transmitir agua a través de sus poros. Esta conductividad disminuye conforme el suelo se seca, y afecta directamente el drenaje, la percolación y la eficiencia del riego.

Estado de desarrollo de la planta (DVS)
Es un índice que representa el avance del cultivo desde la emergencia hasta la madurez, en una escala que va de 0 (emergencia) a 2 (cosecha). Este valor permite relacionar la respuesta del cultivo a factores como el riego en cada etapa fenológica.

Índice de Área Foliar (LAI)
El LAI (Leaf Area Index) representa el área total de hojas por unidad de superficie del suelo (m^2/m^2). Este índice es crucial para estimar la capacidad fotosintética de la planta y su demanda de agua, ya que influye en la intercepción de radiación solar, la transpiración y el crecimiento.

Random Forest Regressor
Random Forest es un algoritmo de aprendizaje supervisado basado en ensamblaje de árboles de decisión. Su principio se basa en construir múltiples árboles independientes (cada uno entrenado sobre subconjuntos aleatorios de los datos y características), y promediar sus predicciones para mejorar la precisión y reducir el sobreajuste. Es especialmente robusto frente a ruido y variables irrelevantes, y funciona bien incluso sin escalar los datos.

XGBoost (Extreme Gradient Boosting)
XGBoost es un algoritmo basado en boosting de árboles, donde los árboles se construyen secuencialmente. Cada nuevo árbol corrige los errores cometidos por los anteriores minimizando una función de pérdida mediante gradiente descendente. Es conocido por su eficiencia computacional y alto rendimiento en problemas de regresión y clasificación. A diferencia de Random Forest, que promedia múltiples árboles independientes, XGBoost construye árboles en cadena, cada uno aprendiendo de los errores del anterior.

2. Descripción y exploración del Dataset

Para simular de manera precisa el crecimiento y desarrollo de los cultivos en el modelo WOFOST, es necesario contar con información detallada sobre las condiciones ambientales que rodean al cultivo. En particular, el modelo requiere como entrada datos meteorológicos diarios —como temperatura, radiación solar, precipitación, humedad del aire y velocidad del viento— así como información edafológica, relacionada con las propiedades físicas e hídricas del suelo. Estos datos permiten al modelo calcular procesos fisiológicos clave como la fotosíntesis, la transpiración y la disponibilidad de agua para las plantas. Por lo tanto, antes de proceder con la implementación de la simulación, se realizará una exploración detallada tanto del conjunto de datos meteorológicos como de los datos del suelo, con el fin de comprender su comportamiento, detectar posibles inconsistencias y garantizar su adecuación al modelo.

2.1 Datos Meteorológicos

Inicialmente, se empleó el conjunto de datos meteorológicos provenientes de la estación de Wageningen, ubicada en los Países Bajos, con registros diarios que abarcan el periodo comprendido entre los años 2004 y 2008. Este archivo contiene información en formato .xlxs estructurada en varias secciones, incluyendo metadatos, unidades y los datos numéricos correspondientes.

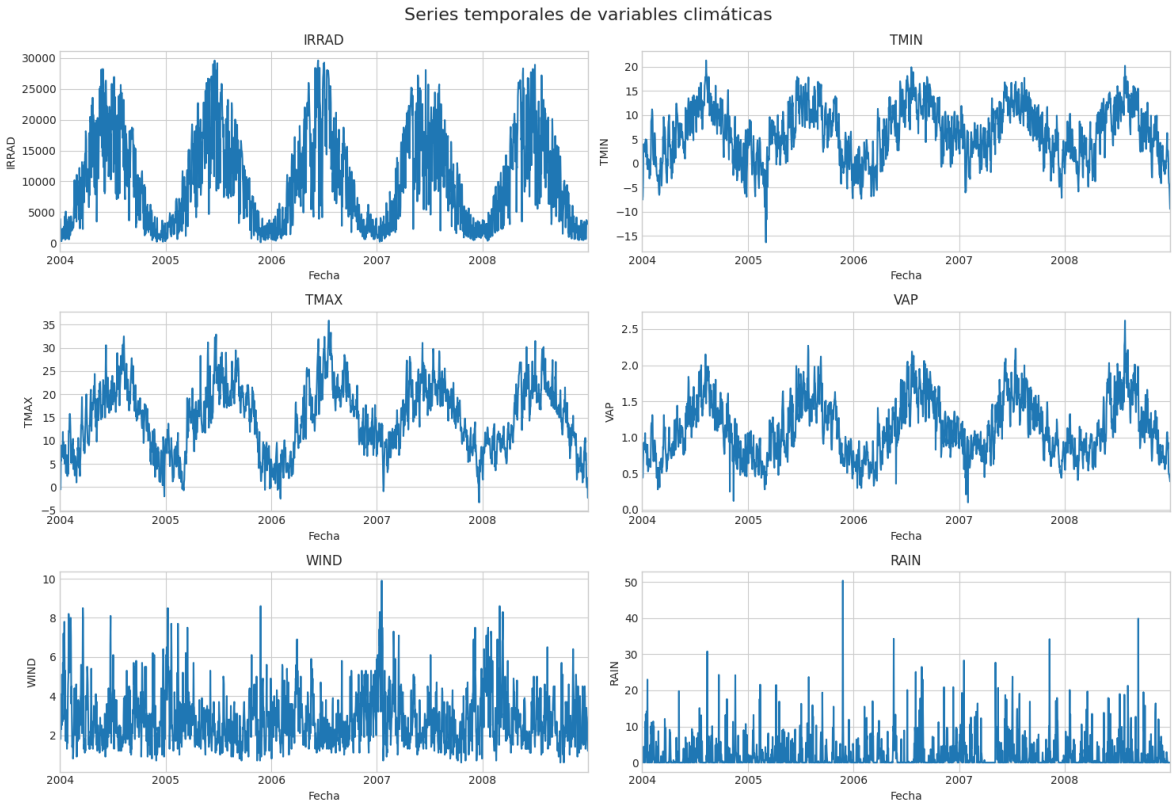
Las variables climáticas consideradas en el análisis incluyen:

- **IRRAD**: radiación solar diaria (kJ/m²/día),
- **TMIN**: temperatura mínima (°C),
- **TMAX**: temperatura máxima (°C),
- **VAP**: presión de vapor (kPa),
- **WIND**: velocidad del viento (m/s),
- **RAIN**: precipitación (mm),
- **SNOWDEPTH**: profundidad de nieve acumulada (cm).

Dado que el modelo WOFOST no requiere la variable de profundidad de nieve para las simulaciones y esta además presenta valores faltantes en la mayor parte del dataset, la columna SNOWDEPTH fue descartada del conjunto de datos antes de continuar con el análisis. Asimismo, se hizo la cuenta de datos NA y solo hay un valor faltante en IRRAD en la primera fila. Por lo cual podemos contar con un Dataset muy robusto.

Como primer paso del análisis exploratorio, se realiza la visualización de cada una de las variables restantes mediante series temporales (Figura 2). Esto permite observar el comportamiento general de los datos a lo largo del tiempo, identificar patrones estacionales y detectar posibles inconsistencias o registros atípicos.

Figura 2. Series temporales de variables climáticas.



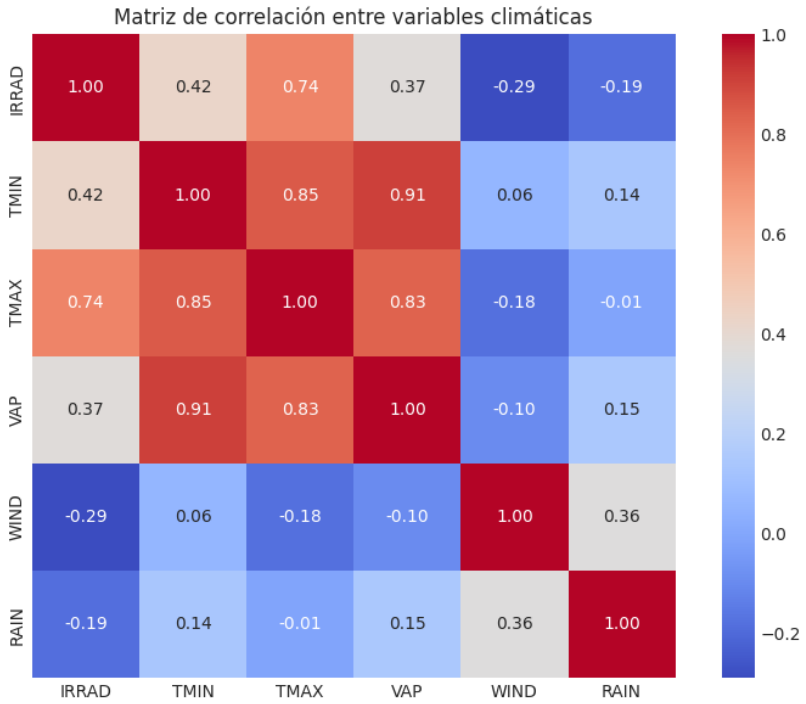
Haciendo un breve análisis a la distribución temporal se observa que:

- **IRRAD (radiación solar):** presenta un patrón estacional claro, con máximos en verano y mínimos en invierno de forma consistente cada año.
- **TMIN y TMAX (temperaturas mínima y máxima):** siguen una oscilación anual típica del clima templado, con temperaturas negativas en invierno y superiores a 25 °C en verano.
- **VAP (presión de vapor):** muestra una tendencia estacional alineada con la temperatura, aumentando en los meses cálidos y disminuyendo en los fríos.
- **WIND (velocidad del viento):** presenta alta variabilidad diaria sin un patrón estacional definido, aunque se observan periodos con mayor intensidad media.
- **RAIN (precipitación):** se distribuye de manera irregular a lo largo del año, con eventos puntuales de alta intensidad y sin una estacionalidad clara.

Posteriormente, se construye una matriz de correlación con el objetivo de identificar posibles relaciones lineales entre las variables climáticas. Esta herramienta permite detectar

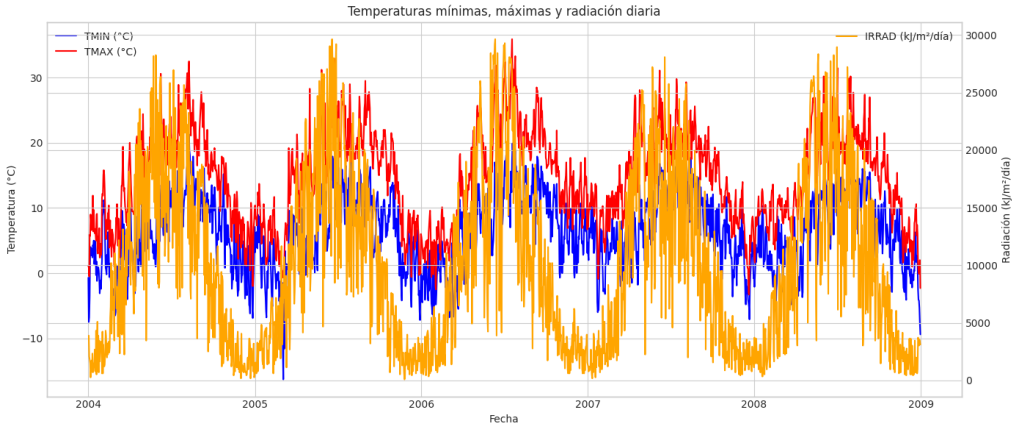
tendencias, asociaciones directas o inversas, y posibles redundancias en los datos que podrían influir en el comportamiento del modelo.

Figura 3. Matriz de correlación.



Se observa una alta correlación positiva entre las variables TMIN y TMAX (~0.90), lo cual resulta coherente desde el punto de vista físico, ya que los días con temperaturas máximas elevadas suelen estar asociados a temperaturas mínimas igualmente altas. La variable IRRAD (radiación solar) presenta una correlación moderada positiva con TMAX (~0.64), lo que indica que, a mayor radiación diaria, tienden a registrarse temperaturas máximas más elevadas. Asimismo, la variable VAP muestra una correlación positiva tanto con TMIN como con TMAX, reflejando que la presión de vapor aumenta con la temperatura, lo cual está respaldado tanto teóricamente como empíricamente. Esta relación entre las variables térmicas y la radiación puede apreciarse de forma más clara al representar TMIN, TMAX e IRRAD en una misma gráfica de series temporales (Figura 4).

Figura 4. TMIN, TMAX e IRRAD.



2.2 Datos del suelo

Los datos requeridos por el modelo WOFOST en relación con el suelo son altamente específicos, incluyendo propiedades físicas e hídricas como la capacidad de retención de agua, la densidad aparente y la profundidad de las capas del perfil edáfico. Este tipo de información no siempre está disponible de forma directa o accesible para todas las regiones.

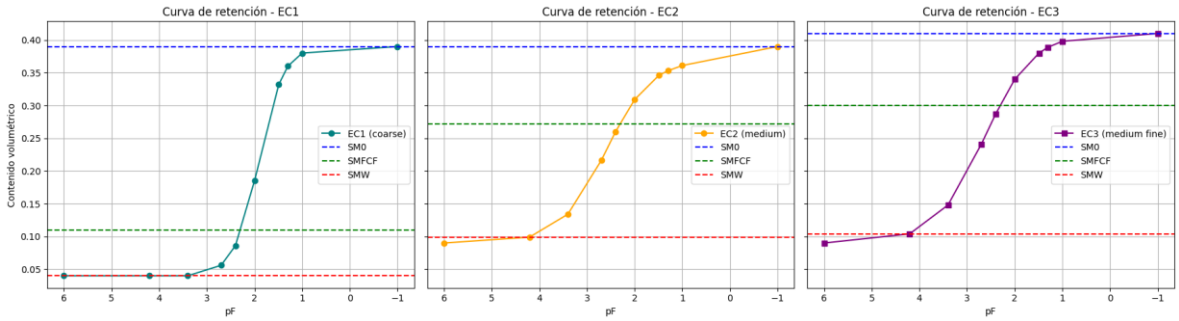
Sin embargo, WOFOST incluye por defecto tres archivos con extensión ".soil", cada uno correspondiente a un tipo de suelo representativo. Estos archivos pueden utilizarse como referencia o punto de partida para realizar simulaciones en ausencia de información local detallada. Los tres tipos de suelo disponibles se resumen en la Tabla 1.

Tabla 1. Clasificación de suelos.

Término	Traducción	Características físicas
Coarse	Suelo grueso	Mucha arena, baja retención de agua, drena rápido, aireado
Medium	Suelo medio	Mezcla equilibrada de arena, limo y arcilla
Medium fine	Medio fino	Más limo o arcilla, retiene más agua, drena más lento

A continuación, se analiza la curva de retención de agua para cada uno de los tipos de suelo incluidos en WOFOST (Figura 5), comparando el contenido volumétrico de humedad en función de la succión del suelo, representada en escala logarítmica. Este análisis permite visualizar las diferencias en la capacidad de retención hídrica entre suelos de distinta textura. Donde EC1 corresponde a Coarse, EC2 a Medium y Ec3 Medium fine.

Figura 5. Curvas de retención para los 3 tipos de suelos.



En las tres curvas se incluyen además los umbrales característicos definidos por el modelo:

- **SM0:** contenido de humedad a saturación.
- **SMFCF:** capacidad de campo.
- **SMW:** punto de marchitez permanente.

Se observa que el suelo EC1, de textura gruesa, tiene una curva más vertical y un contenido de agua disponible más reducido, con una menor capacidad de retención a succión media. En contraste, el suelo EC3, de textura media-fina, presenta una curva más extendida, indicando una mayor capacidad de retención de agua a distintos niveles de succión. El suelo EC2 se comporta de forma intermedia entre ambos extremos, tanto en forma de la curva como en valores de humedad.

Estas diferencias son clave al momento de simular el crecimiento del cultivo, ya que afectan directamente la disponibilidad de agua en el perfil del suelo y, por tanto, la respuesta del modelo ante condiciones de estrés hídrico.

Ya con los datos de entrada preparados —tanto meteorológicos como del suelo y del cultivo—, el siguiente paso consiste en seleccionar una clase de cultivo, definir la fecha de siembra y ejecutar el modelo WOFOST para simular el comportamiento fisiológico del cultivo a lo largo de su ciclo de crecimiento. Esta simulación permitirá generar un conjunto de datos que será considerado como referencia o "realidad" para el entrenamiento posterior del modelo de Machine Learning.

Como parte inicial del análisis, se evaluará si distintos regímenes de riego tienen un efecto significativo sobre el desarrollo de las plantas. Para ello, se aprovechará la funcionalidad de programación de riegos que ofrece WOFOST, la cual permite establecer umbrales de contenido de humedad en el suelo que determinan cuándo debe aplicarse riego. En este caso, se realizaron diferentes simulaciones variando dicho umbral entre los valores 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 y 0.8 (contenido volumétrico de humedad), con el fin de observar cómo responde el cultivo bajo diferentes niveles de disponibilidad hídrica.

Las variables de salida generadas por WOFOS se agrupan en tres categorías principales:

Variables de desarrollo del cultivo

- DVS: Etapa fenológica del cultivo (emergencia, floración, madurez).
- LAI: Índice de área foliar (m^2 de hoja por m^2 de suelo).
- RD: Profundidad efectiva de raíces (m).

Biomasa acumulada

- TAGP: Biomasa total aérea acumulada (kg/ha).
- TWLV: Biomasa acumulada en hojas (kg/ha).

- TWST: Biomasa acumulada en tallos (kg/ha).
- TWSO: Biomasa en órganos de almacenamiento (ej. frutos o raíces cosechables) (kg/ha).
- TWRT: Biomasa acumulada en raíces (kg/ha), que puede estimarse directamente o de forma indirecta según el modelo.

Variables hídricas

- TRA: Transpiración actual (mm/día), influenciada por la disponibilidad de agua en el suelo.
- SM: Contenido volumétrico de humedad en la zona radicular (cm^3/cm^3).
- WWLOW: Agua disponible en capas más profundas del suelo (mm), relevante para la extracción radicular profunda.

Sin embargo, solo vamos a utilizar de respuesta TWSO que representa el rendimiento del fruto. Los resultados se pueden observar en las Figura 6 y 7.

Figura 6. TWSO a diferentes regímenes de riego para Sugarbeet_603 con suelo EC3.

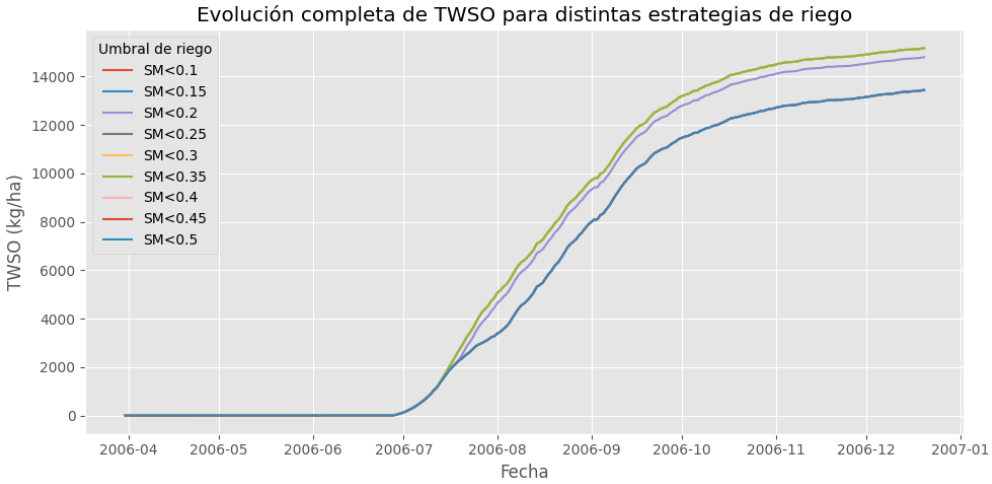
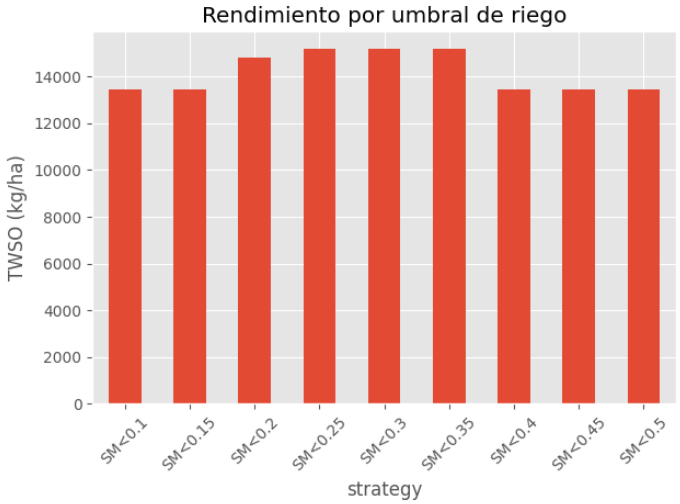


Figura 7. Rendimiento por umbral para Sugarbeet_603 con suelo EC3.



Se observa que los umbrales entre SM < 0.2 y SM < 0.35 generan los mayores rendimientos, alcanzando valores superiores a 14.500 kg/ha. En este rango, el cultivo se mantiene bien abastecido de agua sin incurrir en condiciones de exceso o déficit severo.

En contraste, los umbrales más extremos —tanto muy bajos (SM < 0.1, SM < 0.15) como más altos (SM < 0.4, SM < 0.45, SM < 0.5)— muestran una reducción en el rendimiento. En el primer caso, probablemente por aplicar el riego con demasiada anticipación, reduciendo la eficiencia fisiológica del cultivo; y en el segundo caso, por permitir que el suelo alcance niveles de humedad muy bajos antes de activar el riego, generando estrés hídrico. Este comportamiento sugiere que existe un rango óptimo de umbral de riego entre 0.2 y 0.35 donde se maximiza la producción sin incurrir en limitaciones hídricas ni en condiciones de exceso.

Se hizo la comparación con los otros dos tipos de suelo y se tuvieron resultados muy similares para el soil 2 (Figura 8), sin embargo, para el soil 1 los mejores resultados estuvieron en los umbrales más bajos (Figura 9).

Figura 8. Rendimiento por umbral para Sugarbeet_603 con suelo EC1.

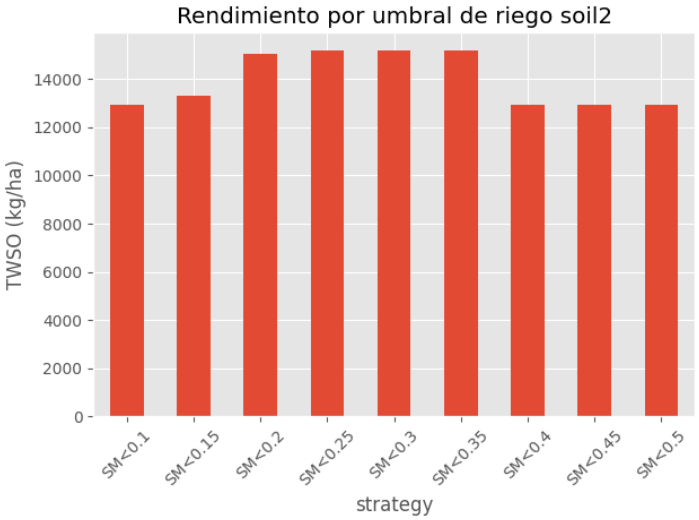
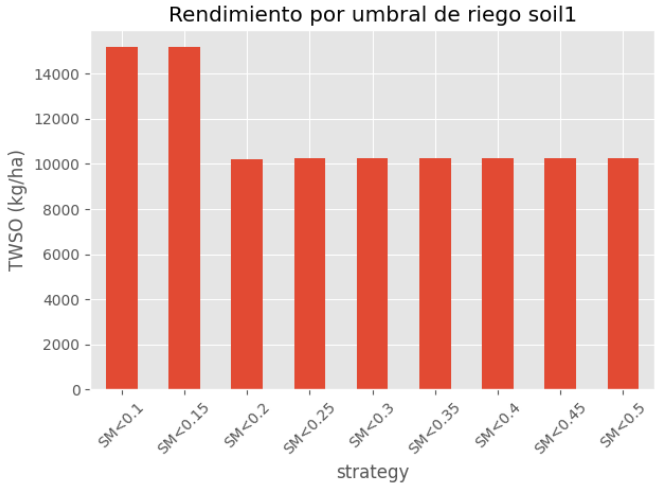


Figura 9. Rendimiento por umbral para Sugarbeet_603 con suelo EC1.



3. Metodología

Una vez obtenidos los resultados de las simulaciones en WOFOST bajo diferentes umbrales de riego y tipos de suelo, se procedió a la construcción del Dataset que serviría como base para el entrenamiento del modelo de aprendizaje automático.

Primero, se ejecutaron simulaciones fisiológicas con WOFOST para distintas combinaciones de umbral de humedad del suelo (strategy) y tipo de suelo (soil). En cada caso, se almacenaron las variables fisiológicas clave, incluyendo la biomasa acumulada en los órganos cosechables (TWSO), así como otras variables como DVS, LAI, RD, TRA, SM, WWLOW, entre otras. Además, se añadió una columna days_since_emergence, la cual representa el número de días transcurridos desde la emergencia del cultivo. Esto permite capturar la dimensión temporal del desarrollo fenológico sin utilizar directamente la fecha.

Paralelamente, se utilizó un conjunto de datos meteorológicos diarios correspondiente al mismo período y ubicación geográfica, el cual contenía variables ambientales como IRRAD, TMIN, TMAX, VAP, WIND y RAIN. Este conjunto fue cargado como un archivo separado y conservaba su índice de fechas.

Para unificar ambas fuentes de información, se realizó una fusión (*merge*) de los datos fisiológicos con las variables ambientales utilizando la fecha como llave común. Antes de unirlos, se filtraron los resultados para conservar únicamente las observaciones a partir de la fecha de emergencia del cultivo, eliminando las entradas con valores nulos previos al inicio del ciclo fenológico.

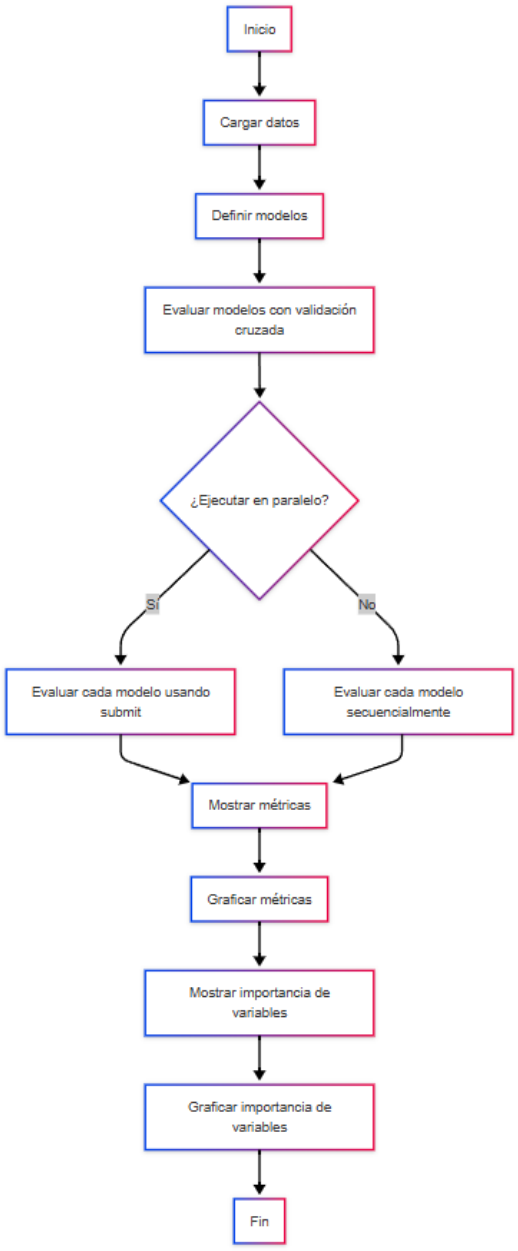
Una vez combinadas ambas fuentes, se seleccionaron como variables de entrada las variables ambientales (IRRAD, TMIN, TMAX, VAP, WIND, RAIN), el tipo de suelo (soil), la estrategia de riego (strategy) y el contador days_since_emergence. La variable objetivo utilizada fue TWSO.

Finalmente, para permitir que los algoritmos de Machine Learning trabajen con las variables categóricas (soil y strategy), estas se transformaron utilizando codificación *One-Hot Encoding*, generando una columna binaria para cada categoría. Tras eliminar cualquier fila con valores faltantes, se obtuvo un conjunto de datos limpio y completamente numérico, listo para ser utilizado en el entrenamiento del modelo.

Descripción del pipeline implementado con Prefect

Para automatizar y estructurar el proceso de entrenamiento, validación y comparación de modelos de regresión, se implementó un flujo de trabajo (flow) usando la librería Prefect. Este flujo está compuesto por varias tareas (tasks), definidas y ejecutadas en orden lógico. Se desarrollaron dos versiones del flujo: una secuencial y otra paralela usando ConcurrentTaskRunner, con el fin de comparar eficiencia y tiempos de ejecución. Los tiempos de ejecución fueron medidos y comparados usando el módulo time. El workflow se puede observar en la Figura 10.

Figura 10. Workflow con prefect.



Tareas del flujo (@flow)

1. **cargar_datos()**
    - Esta tarea lee el conjunto de datos desde un archivo .csv.
    - Se separan las variables independientes X y la variable objetivo y (TWSO), que representa el crecimiento del cultivo.
2. **definir_modelos()**
    - Se crean dos pipelines de aprendizaje supervisado con scikit-learn:
        - RandomForestRegressor con 100 árboles.

        - XGBRegressor (XGBoost) con tasa de aprendizaje 0.1 y profundidad máxima de 6.
    - Ambos modelos están integrados en pipelines que incluyen una etapa de escalado (StandardScaler).
3. **evaluar_un_modelo()** *(usada internamente para paralelizar)*
    - Esta tarea evalúa un modelo mediante **validación cruzada K-Fold (k=10)**.
    - Se calculan tres métricas:
        - **MAE** (Mean Absolute Error)
        - **RMSE** (Root Mean Squared Error)
        - **R²** (Coeficiente de determinación)
    - La validación puede realizarse de forma **paralela** para múltiples modelos usando .submit() y ConcurrentTaskRunner.
4. **mostrar_resultados()**
    - Recibe los resultados agregados de cada modelo y los imprime en consola.
    - Se presentan la media y desviación estándar de cada métrica.
5. **graficar_resultados()**
    - Genera una visualización con subplots comparando **MAE, RMSE y R²** entre modelos.
    - El gráfico se guarda en disco para su análisis posterior.

Las métricas para validar el aprendizaje de máquina que se utilizaron fueron el MAE (Mean Absolute Error (Error Absoluto Medio), RMSE (Raíz del Error Cuadrático Medio) y R^2 (Coeficiente de Determinación).

- **MAE**

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

Donde y_i es el valor real, $\hat{y}_i$ el valor predicho y n el número de observaciones

- **RMSE**

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

- $\boldsymbol{R^2}$

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

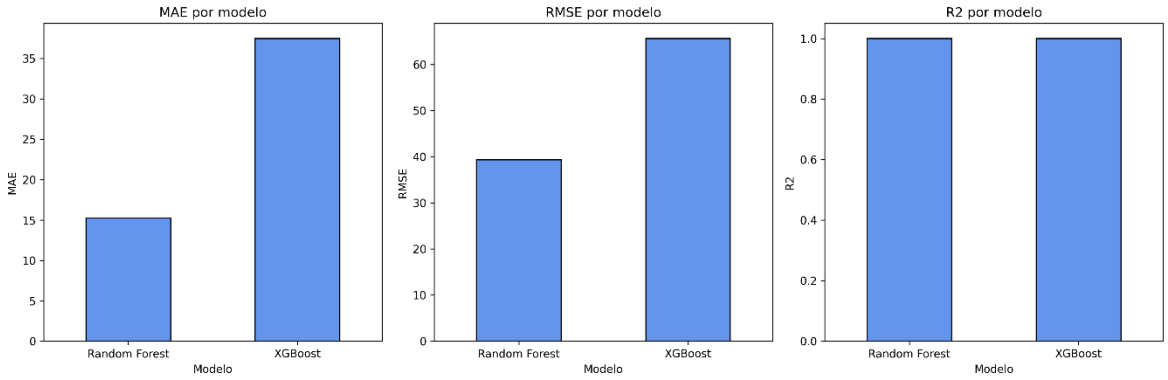
Donde $\bar{y}$ es la media de los valores reales, el numerador representa el error del modelo y el denominador es la varianza total del objetivo.

4. Resultados

Se compararon dos enfoques de ejecución para el flujo de entrenamiento: uno secuencial y otro paralelo. En el enfoque secuencial, cada modelo fue evaluado uno tras otro, lo cual resultó en un tiempo total de ejecución de 13.15 segundos. En contraste, al aplicar paralelización mediante ConcurrentTaskRunner de la librería Prefect, se logró ejecutar la evaluación de los modelos de forma simultánea, reduciendo el tiempo total a 7.03 segundos. Esta mejora evidencia la eficiencia del procesamiento paralelo en tareas independientes y repetitivas como la validación cruzada por modelo, especialmente en flujos de trabajo escalables.

La figura muestra la comparación del desempeño de los modelos Random Forest y XGBoost en términos de tres métricas: MAE, RMSE y R^2. Se observa que Random Forest obtiene un menor error absoluto medio (MAE ≈ 15.3) y un menor error cuadrático medio (RMSE ≈ 39.3), lo que indica una mayor precisión en sus predicciones. En contraste, XGBoost presenta errores significativamente más altos (MAE ≈ 37.5, RMSE ≈ 65.7). A pesar de esto, ambos modelos alcanzan un coeficiente de determinación R^2 cercano a 1, lo que implica que ambos ajustan bien a los datos, aunque Random Forest lo hace con menor dispersión del error. Esta comparación permite concluir que, bajo las condiciones de entrenamiento y validación utilizadas, Random Forest ofrece un mejor rendimiento general para el conjunto de datos analizado.

Figura 11. Comparación del modelo por métricas.



Los gráficos muestran la importancia relativa de las variables utilizadas por los modelos Random Forest y XGBoost para predecir el crecimiento del cultivo. En ambos casos, la variable days_since_emergence (días desde la emergencia) domina con una contribución muy superior al resto, lo que indica que es la principal determinante en la predicción de la variable objetivo. En el modelo XGBoost, sin embargo, se observa una mayor dispersión de la importancia entre otras variables secundarias, como strategy_SM_0_15, strategy_SM_0_1, y soil_EC1, lo que sugiere una utilización más amplia del conjunto de datos. En contraste, el modelo Random Forest tiende a concentrar su aprendizaje en una única variable, lo que puede implicar mayor simplicidad, pero también mayor dependencia de esa característica (Figura 12 y Figura 13). Esta comparación permite concluir que, aunque ambos modelos coinciden en la variable más influyente, XGBoost distribuye mejor la importancia entre múltiples atributos, lo que puede aportar mayor robustez frente a variaciones en los datos de entrada.

Figura 12. Importancia de variables Random Forest.

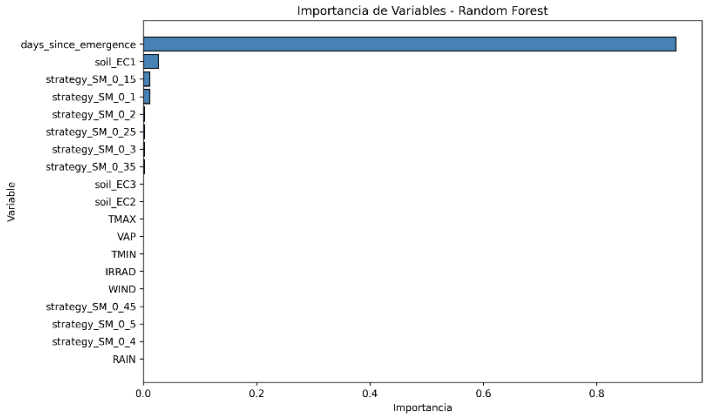
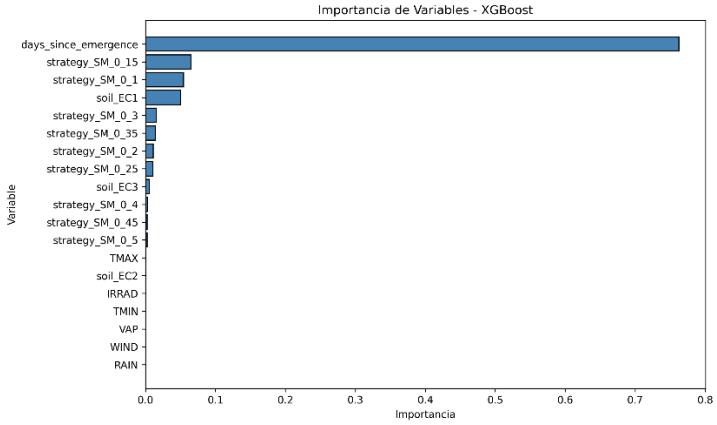


Figura 13. Importancia de variables XGBoost.



Cabe destacar que se esperaba una mayor relevancia de las variables relacionadas con el tipo de suelo (soil_EC), dado que se conoce la existencia de interacción entre el tipo de suelo y las estrategias de riego (strategy_SM). Sin embargo, ninguno de los modelos capturó esta interacción de forma significativa.

Interpretación, Limitaciones y Trabajo Futuro

Una posible explicación del comportamiento observado es la limitación del conjunto de datos, el cual incluye únicamente tres tipos de suelo y está basado exclusivamente en las condiciones climáticas de Wageningen. Esta escasa variabilidad en los datos reduce la capacidad del modelo para identificar interacciones más complejas y generalizar su comportamiento a otros entornos agroecológicos.

Como trabajo futuro, se propone ampliar la base de datos incorporando información proveniente de distintas regiones con climas diversos y una mayor variedad de tipos de suelo. Esto permitiría al modelo capturar relaciones más representativas del mundo real, mejorar su capacidad de generalización y hacerlo más robusto y escalable para aplicaciones en agricultura de precisión en distintos contextos productivos.

5. Conclusiones

Dada la limitada disponibilidad de sensores agroclimáticos en campo, la estrategia de utilizar el modelo fisiológico WOFOST con datos climáticos resultó ser una solución efectiva para generar datos sintéticos representativos del crecimiento y productividad de cultivos.

La combinación de datos simulados y técnicas de Machine Learning permitió predecir con alta precisión el rendimiento del cultivo (TWSO), destacando especialmente el modelo Random Forest, que obtuvo los mejores resultados en términos de MAE y RMSE.

La implementación de un flujo de trabajo paralelo mediante Prefect y ConcurrentTaskRunner redujo el tiempo de ejecución de 13.15 a 7.03 segundos, demostrando una mejora sustancial en eficiencia computacional al entrenar y validar múltiples modelos en simultáneo.

En ambos modelos, la variable days_since_emergence fue identificada como la más influyente en la predicción del rendimiento, superando ampliamente al resto. Sin embargo, se observó que XGBoost tendió a distribuir más equitativamente la importancia entre varias variables, a diferencia de Random Forest, que concentró su aprendizaje principalmente en una sola.

A pesar del buen desempeño predictivo, los modelos no lograron capturar adecuadamente la interacción entre tipo de suelo y estrategia de riego. Esto se atribuye a la escasa variabilidad del Dataset, que incluía solo tres tipos de suelo y un único entorno climático (Wageningen), limitando la capacidad del modelo para generalizar a otras condiciones.

Para mejorar la robustez y escalabilidad del sistema, se recomienda ampliar el conjunto de datos con simulaciones o mediciones reales provenientes de múltiples zonas agroecológicas. Esto permitiría capturar relaciones más complejas y entrenar modelos más generalizables, aptos para diversas condiciones climáticas, edafológicas y de manejo agrícola.