




Modelado Predictivo de Condiciones Ambientales en Agricultura Urbana Vertical Mediante Fotoluminiscencia y Aprendizaje Automático

Bootcamp en Inteligencia Artificial

Modelado Predictivo de Condiciones Ambientales en Agricultura Urbana Vertical Mediante Fotoluminiscencia y Aprendizaje Automático

Víctor C. Vladimir Cortés Arévalo (V.C.V.C.A)

 vladimir.cortes@outlook.com

 (+57) 3144427602

Grupo No. 18.

Bootcamp en Inteligencia Artificial

Tutores:

Magnolia Cardozo Hernández

Emmanuel Mejía.

Jheyson Galvis.

Bogotá D.C., mayo de 2025

© 2025. V.C.V.C.A. Todos los derechos reservados.

Queda estrictamente prohibida la reproducción, distribución o modificación de este contenido por cualquier medio, digital o físico, sin autorización escrita y registrada del autor.

CONTENIDO

1	ABSTRACT	4
2	ANTECEDENTES	4
2.1	Problemática	5
2.2	Relevancia	6
3	PLANTEAMIENTO DEL PROBLEMA	8
3.1	PREGUNTA PROBLEMA	9
3.2	HIPÓTESIS	9
4	JUSTIFICACIÓN	9
5	OBJETIVOS	12
5.1	Objetivo Principal	12
5.2	Objetivos Específicos	12
6	METODOLOGÍA	13
6.1	DATOS	14
6.1.1	Origen y Naturaleza de los Datos	14
6.2	CARACTERÍSTICAS DEL DATASET	15
6.3	PREPROCESAMIENTO	19
6.3.1	Análisis de Calidad de Datos	19
6.3.2	Detección y Manejo de Outliers	19
6.3.3	Transformaciones de Datos	20
6.3.4	División de Datos	21
6.4	INGENIERÍA DE CARACTERÍSTICAS	22
6.4.1	Análisis de Correlaciones Avanzado	23
6.4.2	Creación de Features Avanzadas	23
6.4.2.1	Features de Interacción Multiplicativa	24
6.4.2.2	Features de Ratios Significativos	24
6.4.2.3	Features de Distancia a Condiciones Óptimas	24
6.4.2.4	Índices Compuestos	25
6.4.2.5	Features Temporales Avanzadas	25

6.4.2.6	Transformaciones No Lineales	26
6.4.3	Selección de Features	26
6.5	MODELOS DE MACHINE LEARNING	28
6.5.1	Algoritmos Seleccionados	28
6.5.1.1	Modelos Lineales	29
6.5.1.2	Modelos Basados en Árboles	30
6.5.1.3	Otros Algoritmos	30
6.5.2	Configuración de Entrenamiento	31
6.5.3	Validación y Métricas	34
7	RESULTADOS	38
7.1	ANÁLISIS EXPLORATORIO DE DATOS (EDA)	38
7.1.1	Características Generales del Sistema	38
7.1.2	Distribución de Condiciones Operativas	40
7.1.3	Patrones Temporales	42
7.1.4	Correlaciones Clave Identificadas	44
7.1.5	Análisis del Sistema de Control	47
7.2	Desempeño de Modelos	48
7.2.1	Resultados Comparativos Completos	48
7.2.2	Análisis de Modelos Ganadores	50
7.2.3	Análisis de Residuos:	53
7.2.4	Análisis de Estabilidad por Rangos	54
8	CONCLUSIONES	55
8.1	INTERPRETACIÓN DE RESULTADOS	55
8.1.1	Significado de los Hallazgos Principales	55
8.1.2	Jerarquía de Factores Influyentes	55
8.1.3	Patrones Temporales y su Significado	56
9	ANEXOS	59
9.1	ANEXO NO.1 – REPOSITORIO GITHUB	59
9.2	ANEXO No.2 – API REST	59

9.3 ANEXO No.3 – PROTOTIPO DIGITAL

61

LISTA DE FIGURAS

Ilustración 1. Ecosistema tecnológico para AUVPFL	6
Ilustración 2. Ambiente Controlado AUVPFL	7
Ilustración 3. Prototipo Funciona AUVPFL	7
Ilustración 4. Cultivo en un ecosistema de AUVPF	9
Ilustración 5. Dashboard modelo predictivo AUVPF.....	12
Ilustración 6. Metodología aplicada.....	13
Ilustración 7. Dataset RAW.	16
Ilustración 8. estrategia de partición de datos.	22
Ilustración 9. Eficiencia fotosintetica.....	39
Ilustración 10. Fotoluminiscencia, analisis luminico	40
Ilustración 11. Condiciones ambientales del sistema Cerrado.	41
Ilustración 12. Calidad del aire.....	41
Ilustración 13. Patrones Temporales	42
Ilustración 14. Variación Estacional.....	43
Ilustración 15. Matriz de Correlación contaminantes del aire	44
Ilustración 16 Matriz de correlación Sistema de AUVPF.....	45
Ilustración 17. Matriz de correlación variables clave del sistema	46

LISTA DE TABLAS

Tabla 1. Evaluación inicial del dataset RAW.....	19
Tabla 2. Rangos realistas para agricultura urbana vertical de precisión con fotoluminiscencia.	20
Tabla 3. Variables clave del dataset	28
Tabla 4. Desempeño – Eficiencia Fotoluminiscencia.....	49
Tabla 5. Desempeño fotoluminiscencia	49
Tabla 6. Métodos básicos de predicción	53
Tabla 7. Desempeño de rangos de eficiencia	54

1 ABSTRACT

Objetivo: Diseñar un sistema inteligente para optimizar cultivos en agricultura urbana vertical, integrando simulación, análisis y predicción a partir de datos sintéticos ambientales y modelos de aprendizaje automático, con el fin de evaluar la eficiencia fotosintética y la fotoluminiscencia en entornos controlados.

Metodología: La metodología adoptada es de tipo cuantitativa-aplicada, estructurada como un pipeline¹ modular de ciencia de datos y aprendizaje automático, con enfoque experimental y predictivo.

Resultados: El framework de evaluación desarrollado representa un avance significativo en la validación de modelos ML para agricultura de precisión, proporcionando una metodología comprehensiva, replicable y científicamente fundamentada que puede servir como estándar para la comunidad científica e industrial en el desarrollo de sistemas inteligentes para agricultura vertical y de precisión

Palabras clave: aprendizaje automático, datos simulados, ingeniería de características, XGBoost, validación cruzada, predictivo.

2 ANTECEDENTES

La agricultura urbana vertical se presenta como una solución innovadora frente a la escasez de suelo cultivable, el crecimiento poblacional urbano y la necesidad de garantizar seguridad alimentaria en entornos urbanos. Este enfoque permite cultivar alimentos en espacios reducidos y controlados, utilizando estructuras verticales, iluminación artificial, sensores y sistemas hidropónicos o aeropónicos.

En los últimos años, se ha incrementado el interés en el uso de **sensores ambientales**, **sistemas inteligentes de monitoreo**, e incluso **IA aplicada al análisis de datos agroambientales**, con el fin de mejorar el rendimiento, reducir desperdicios y garantizar la sostenibilidad.

Una de las métricas emergentes es la **fotoluminiscencia**, entendida como la reemisión de luz por parte de las plantas al recibir radiación, la cual puede correlacionarse con

¹ Un **pipeline de ciencia de datos** es un flujo de trabajo estructurado que permite procesar, transformar y analizar datos de manera eficiente. Se compone de diversas etapas que garantizan la preparación, integración y aplicación de modelos para obtener información útil.

procesos fisiológicos como la fotosíntesis, permitiendo estimar el estado de salud de los cultivos sin métodos invasivos.

2.1 Problemática

- ✎ **Limitada disponibilidad de datos reales para modelado predictivo:** En entornos urbanos, el monitoreo constante de variables como temperatura, CO₂, humedad, partículas en suspensión o iluminación suele ser costoso, intermitente o inexacto. Esta limitación impide alimentar modelos de análisis y predicción robustos.
- ✎ **Falta de integración entre indicadores fotoluminiscentes y variables ambientales:** Aunque existen investigaciones en fisiología vegetal, pocos estudios han vinculado de forma práctica la **fotoluminiscencia** como variable clave dentro de sistemas inteligentes de control ambiental.
- ✎ **Ausencia de soluciones predictivas automatizadas para toma de decisiones:** La mayoría de los sistemas de agricultura urbana funcionan bajo reglas estáticas, sin algoritmos que anticipen necesidades de ajuste ambiental o estimen de forma proactiva el rendimiento fotosintético.
- ✎ **Escasa adaptación de tecnologías de aprendizaje automático al contexto agrícola urbano:** Si bien existen avances en AgTech, gran parte se enfoca en cultivos extensivos y no en condiciones específicas de cultivos indoor verticales, donde las dinámicas ambientales y la necesidad de precisión son distintas.

2.2 Relevancia

Este proyecto es relevante en múltiples dimensiones:

Científica y técnica:

- Introduce un **modelo de generación de datos sintéticos ambientalmente coherente**, permitiendo simular escenarios agrícolas urbanos sin depender de sensores físicos.
- Implementa un **pipeline robusto de análisis predictivo** usando Random Forest, XGBoost, redes neuronales y modelos temporales (Prophet, ARIMA), integrando ciencia de datos con conocimiento agroambiental.

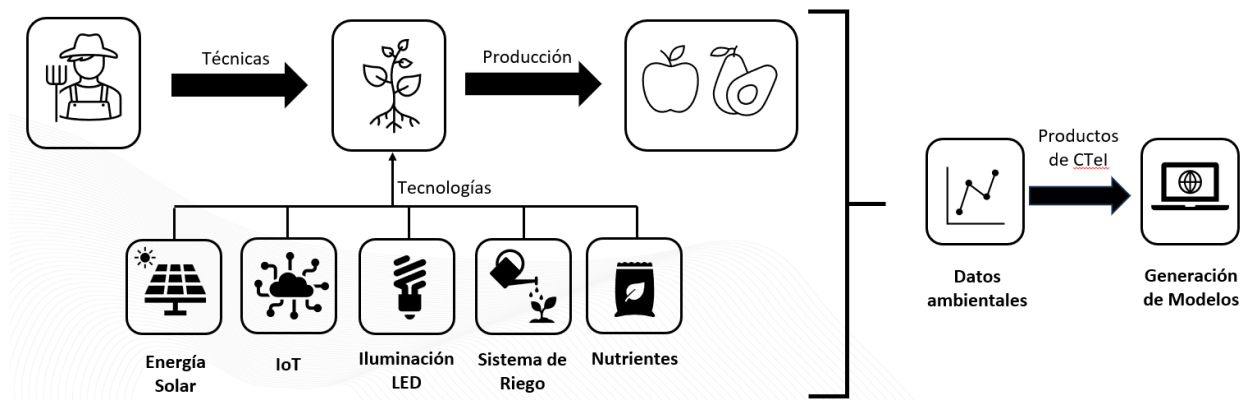


Ilustración 1. Ecosistema tecnológico para AUVPFL²

Urbana y sostenible:

- Contribuye a la consolidación de **ciudades inteligentes** que integran producción de alimentos con tecnologías limpias.
- Permite anticipar condiciones subóptimas para los cultivos, reduciendo pérdidas energéticas y mejorando la eficiencia de recursos como agua, luz y CO₂.

² AUVPFL: Agricultura urbana vertical de precisión con fotoluminiscencia

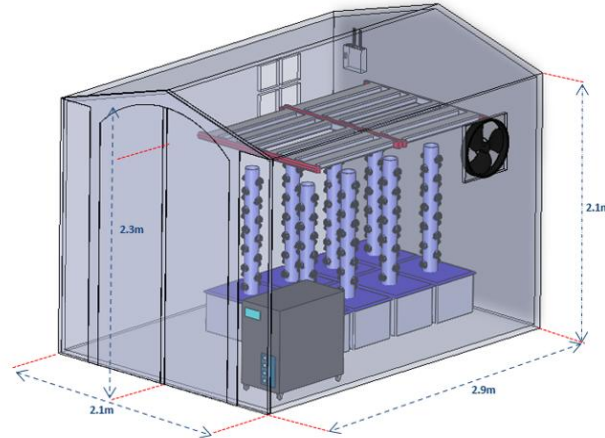
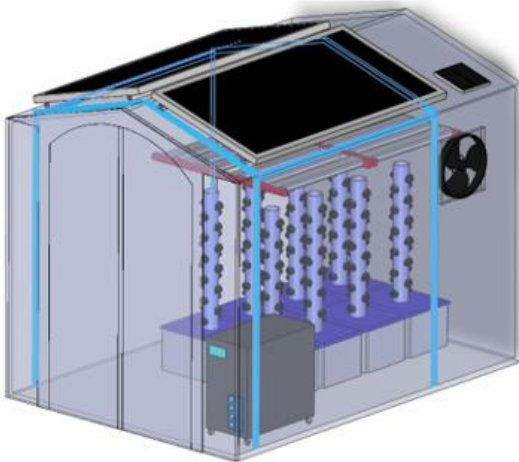


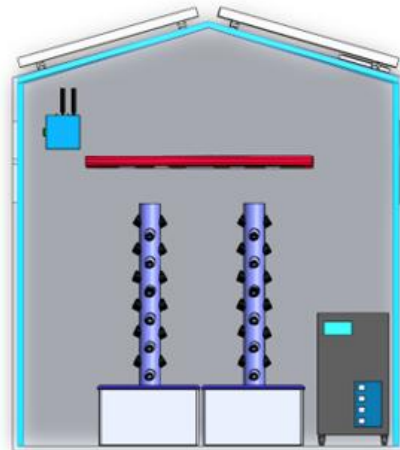
Ilustración 2. Ambiente Controlado AUVPFL

Aplicada y práctica:

- Proporciona herramientas de decisión basadas en datos, que pueden integrarse con APIs y plataformas IoT para control automático de entornos indoor.
- Aporta al desarrollo de **sistemas inteligentes de agricultura urbana**, escalables y replicables en distintas regiones del mundo.



Vista Perspectiva (Estructura de refuerzo dentro de Invernadero)



Vista Frontal (Estructura de refuerzo dentro de Invernadero)

Ilustración 3. Prototipo Funciona AUVPFL

3 PLANTEAMIENTO DEL PROBLEMA

En las últimas décadas, el acelerado crecimiento de la población urbana ha generado una presión significativa sobre los sistemas tradicionales de producción y distribución de alimentos. Este fenómeno ha motivado la exploración de soluciones innovadoras como la **agricultura urbana vertical**, la cual permite cultivar en espacios reducidos mediante sistemas hidropónicos, iluminación artificial y ambientes controlados.

Sin embargo, a pesar de sus beneficios potenciales en términos de sostenibilidad y eficiencia, este tipo de agricultura enfrenta desafíos críticos asociados a la **gestión y optimización de variables ambientales**. Factores como la temperatura, la humedad relativa, la concentración de CO₂, la calidad del aire (PM2.5, VOCs, NO₂, O₃) y la intensidad lumínica, inciden directamente en el rendimiento fotosintético de las plantas, y por tanto en su productividad.

En este contexto, la **fotoluminiscencia** emerge como una métrica clave para evaluar el estado fisiológico de los cultivos, al reflejar su capacidad para absorber y reemitir luz, correlacionándose con procesos de fotosíntesis y salud vegetal. Sin embargo, el uso de la fotoluminiscencia en entornos urbanos verticales aún es incipiente, y su potencial predictivo no ha sido ampliamente explorado en modelos automatizados de análisis.

Adicionalmente, los sistemas actuales de agricultura urbana presentan limitaciones importantes:

- Carecen de mecanismos predictivos que anticipen caídas en la eficiencia fotosintética o necesidades de ajuste ambiental.
- Se basan en reglas fijas, sin integrar modelos de aprendizaje automático ni análisis multivariable dinámico.
- La mayoría de estudios existentes dependen de sensores físicos, lo que impide la simulación y validación a gran escala sin costos operacionales elevados.

Por tanto, **existe una necesidad urgente de desarrollar un sistema inteligente capaz de simular, analizar y predecir de forma automatizada el comportamiento ambiental y fotosintético** en entornos de agricultura vertical urbana, integrando la fotoluminiscencia como variable estratégica y utilizando herramientas avanzadas de ciencia de datos e inteligencia artificial.

3.1 PREGUNTA PROBLEMA

¿Cómo puede un sistema de modelado predictivo basado en aprendizaje automático y fotoluminiscencia optimizar la estimación de la eficiencia fotosintética y las condiciones ambientales en sistemas de agricultura urbana vertical bajo ambientes controlados?

3.2 HIPÓTESIS

Si se diseña un sistema de modelado predictivo que integre variables ambientales, indicadores de calidad del aire y métricas de fotoluminiscencia mediante técnicas de aprendizaje automático, entonces será posible estimar con alta precisión la eficiencia fotosintética de los cultivos en agricultura urbana vertical y proponer ajustes proactivos para mejorar su rendimiento.

4 JUSTIFICACIÓN

El desarrollo del presente proyecto se justifica por la confluencia de múltiples necesidades globales que afectan tanto la sostenibilidad alimentaria como el diseño de sistemas agrícolas más eficientes e inteligentes en entornos urbanos.



Ilustración 4. Cultivo en un ecosistema de AUVPF

Desafíos en Producción Agrícola Urbana

Las ciudades enfrentan una creciente presión por abastecer alimentos frescos a una población cada vez más numerosa, en espacios limitados y bajo condiciones ambientales no siempre favorables. La **agricultura urbana vertical** ha emergido como una alternativa sostenible que permite la producción local e intensiva de cultivos, optimizando el uso del espacio y reduciendo la huella de carbono.

Sin embargo, **el éxito de estos sistemas depende críticamente de la capacidad de monitorear y ajustar en tiempo real** parámetros como temperatura, humedad, calidad del aire y luz. En este sentido, se requiere inteligencia tecnológica que permita **predecir comportamientos y automatizar decisiones** ambientales antes de que afecten negativamente el desarrollo del cultivo.

Vacíos Tecnológicos en Modelado Predictivo Agrícola

Aunque existen avances en la digitalización agrícola, muchos de los sistemas actuales:

- Carecen de **modelos predictivos integrados** capaces de anticipar la eficiencia de las plantas según las condiciones del entorno.
- No aprovechan **métricas avanzadas como la fotoluminiscencia**, pese a su potencial como indicador temprano del rendimiento fotosintético.
- No permiten realizar simulaciones a gran escala sin datos reales, lo que limita la planificación o investigación.

El presente proyecto aborda estos vacíos mediante la construcción de un sistema que combina:

- **Generación de datos sintéticos** ambientalmente coherentes.
- **Cálculo de indicadores derivados** clave (AQI, VPD, eficiencia fotosintética).
- **Modelos de Machine Learning y Deep Learning** para predicción.
- **Series temporales** para modelar tendencias estacionales.

Pertinencia Tecnológica y Reutilización Práctica

La arquitectura modular del sistema desarrollado permite que sus componentes (dataset, modelos, escaladores, funciones predictivas) sean **reutilizables e integrables** en sistemas reales de agricultura vertical.

Además, su diseño pensado para la producción (mediante FastAPI y exportación de modelos) lo hace aplicable en entornos de monitoreo IoT, dashboards de control agrícola y plataformas de agricultura inteligente.

Impacto Científico y Educativo

Este proyecto no solo contribuye al **avance de la agricultura urbana digitalizada**, sino que también ofrece:

- Un banco de pruebas sintético y reproducible para experimentación.
- Un caso de estudio completo que puede ser usado para formación en **ciencia de datos aplicada a la agricultura, IA ambiental, y análisis predictivo multivariable**.

En resumen, este proyecto responde a una necesidad real y actual: desarrollar soluciones inteligentes, sostenibles y adaptables que potencien la agricultura urbana como eje de la seguridad alimentaria del futuro, integrando datos, inteligencia artificial y conocimiento agronómico en una propuesta tecnológica de alto valor.

5 OBJETIVOS

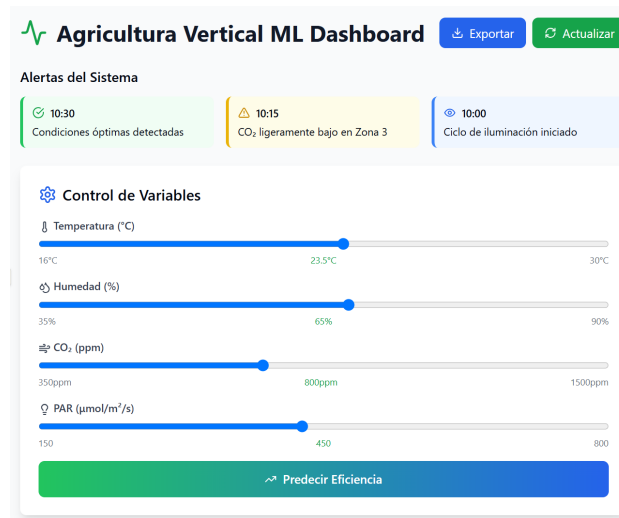


Ilustración 5. Dashboard modelo predictivo AUVPF.

5.1 Objetivo Principal

Desarrollar un sistema inteligente de **simulación, análisis y predicción** orientado a la optimización de cultivos en agricultura urbana vertical, mediante la generación de datos sintéticos ambientales y la aplicación de modelos de aprendizaje automático para evaluar la eficiencia fotosintética y la fotoluminiscencia bajo condiciones controladas.

5.2 Objetivos Específicos

- 🔧 **Diseñar y generar un conjunto de datos sintéticos representativos** de las variables ambientales, lumínicas y de calidad del aire asociadas a cultivos indoor, utilizando técnicas estadísticas realistas y condiciones de simulación controladas.
- 🔧 **Construir métricas derivadas e indicadores de eficiencia agrícola**, como el índice de calidad del aire (AQI), el déficit de presión de vapor (VPD) y la eficiencia fotosintética, para caracterizar el entorno de cultivo y sus condiciones operativas.

- 🔧 **Entrenar y evaluar modelos de Machine Learning y Deep Learning** que permitan predecir el comportamiento de variables clave (eficiencia fotosintética y fotoluminiscencia), comparando enfoques como Random Forest, XGBoost, SVR, redes neuronales y modelos de series temporales (Prophet y ARIMA).
- 🔧 **Implementar una arquitectura modular y reutilizable**, que permita la exportación, carga y despliegue de modelos predictivos en entornos productivos mediante servicios web, facilitando su integración con plataformas como FastAPI y su aplicación en sistemas inteligentes de monitoreo agrícola.

6 METODOLOGÍA

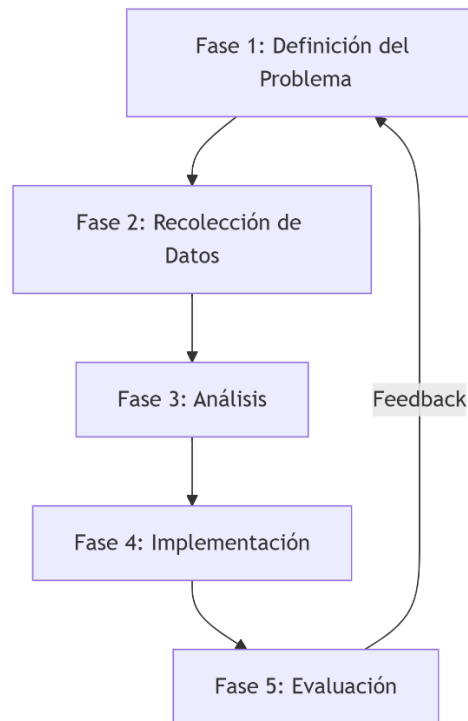


Ilustración 6. Metodología aplicada

La metodología empleada para el desarrollo del proyecto siguió un **enfoque sistemático de Machine Learning Engineering** basado en mejores prácticas de ciencia de datos, estructurado como un **pipeline modular de 5 fases secuenciales**:

Fase 1 - Generación sintética de datos realistas mediante simulación computacional basada en conocimiento agronómico, incorporando distribuciones probabilísticas

gaussianas con correlaciones biológicamente plausibles entre 26 variables ambientales, lumínicas y de control para un sistema de agricultura vertical con fotoluminiscencia.

Fase 2 - Análisis exploratorio exhaustivo (EDA) implementando visualizaciones multivariadas, análisis de correlaciones (Pearson/Spearman), pruebas de normalidad (Shapiro-Wilk), y caracterización de patrones temporales circadianos y estacionales para validar la coherencia del dataset sintético con principios agronómicos establecidos.

Fase 3 - Preprocesamiento robusto mediante detección de outliers con criterios específicos del dominio (IQR modificado), manejo inteligente de valores faltantes, transformaciones de tipos de datos, y división estratificada (60/20/20) basada en cuartiles de eficiencia fotosintética.

Fase 4 - Ingeniería de características avanzada guiada por conocimiento del dominio, creando features de interacción multiplicativa (temperatura \times PAR, CO₂ \times luz), ratios físicamente significativos, distancias a condiciones óptimas, índices compuestos de estrés ambiental, transformaciones temporales trigonométricas para capturar ciclicidad, y selección multi-criterio combinando F-score, información mutua e importancia de Random Forest.

Fase 5 - Entrenamiento comparativo de 10 algoritmos (desde regresión lineal hasta XGBoost) con optimización de hiperparámetros via Grid Search, validación cruzada estratificada, y evaluación exhaustiva mediante múltiples métricas (MAE, RMSE, R², MAPE) incluyendo análisis de residuos, estabilidad por rangos, robustez ante ruido, y casos extremos, culminando con la implementación de un sistema de evaluación comprehensiva que incluye análisis de importancia de features, dependencia parcial, y comparación con baselines naive para validar la efectividad del enfoque propuesto.

6.1 DATOS

6.1.1 Origen y Naturaleza de los Datos

Los datos utilizados en este estudio fueron **generados sintéticamente** mediante simulación computacional para representar un sistema de agricultura urbana vertical en ambiente controlado con tecnología fotoluminiscente. Esta aproximación se justifica por:

1. **Ausencia de datasets públicos** específicos para sistemas de agricultura vertical con fotoluminiscencia.
2. **Control experimental completo** sobre las variables del sistema.

3. **Capacidad de generar escenarios diversos** bajo condiciones controladas.
4. **Reproducibilidad** del experimento y validación de metodologías.

El proceso de generación se basó en **conocimiento del dominio agrícola y principios físicos** de la fotosíntesis, incorporando:

- ✎ Rangos realistas para agricultura indoor (temperatura: 16-30°C, humedad: 35-90%)
- ✎ Correlaciones biológicamente plausibles entre variables
- ✎ Variabilidad temporal y estacional
- ✎ Ruido gaussiano para simular condiciones reales

6.2 CARACTERÍSTICAS DEL DATASET

Tamaño del Dataset:

- ✎ **50,000 registros** de mediciones del sistema
- ✎ **26 variables** distribuidas en categorías funcionales
- ✎ **Período temporal:** 2 años simulados (muestreo cada 0.5 horas)
- ✎ **Densidad temporal:** 12 mediciones por día

```






Informacion del dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 26 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   timestamp                                50000 non-null  datetime64[ns]
1   hora_dia                                50000 non-null  int64
2   dia_semana                              50000 non-null  int64
3   estacion                                 50000 non-null  object
4   temperatura_c                           50000 non-null  float64
5   humedad_rel_pct                          50000 non-null  float64
6   co2_ppm                                 50000 non-null  float64
7   vpd_kpa                                 50000 non-null  float64
8   pm2_5_ugm3                             50000 non-null  float64
9   pm10_ugm3                              50000 non-null  float64
10  vocs_mgm3                              50000 non-null  float64
11  no2_ugm3                                50000 non-null  float64
12  o3_ugm3                                 50000 non-null  float64
13  aqi_indice                              50000 non-null  float64
14  par_umol_m2_s                           50000 non-null  float64
15  ppfd_umol_m2_s                           50000 non-null  float64
16  fotoluminiscencia_intensidad             50000 non-null  float64
17  espectro_pico_nm                         50000 non-null  int64
18  eficiencia_fotosintetica_pct             50000 non-null  float64
19  ajuste_temp_c                            50000 non-null  float64
20  ajuste_humedad_pct                       50000 non-null  float64
21  ajuste_co2_ppm                           50000 non-null  float64
22  ajuste_par_umol                          50000 non-null  float64
23  purificacion_requerida_pct               50000 non-null  float64
24  condiciones_categoria                    50000 non-null  object
25  calidad_aire_categoria                   49956 non-null  category
dtypes: category(1), datetime64[ns](1), float64(19), int64(3), object(2)
memory usage: 9.6+ MB

```







Ilustración 7. Dataset RAW.

Variables del Sistema:




Variables Ambientales Principales (5):

-  **temperatura_c:** Temperatura ambiente (°C) – Normal(22, 3), rango [16, 30]
-  **humedad_rel_pct:** Humedad relativa (%) – Normal(68, 10), rango [35, 90]
-  **co2_ppm:** Concentración CO₂ (ppm) – Normal(800, 200), rango [350, 1500]
-  **par_umol_m2_s:** Radiación fotosintéticamente activa – Normal(400, 120), rango [150, 800]
-  **vpd_kpa:** Déficit de presión de vapor (kPa) – Calculado según fórmula estándar

Variables de Calidad del Aire (6):

-  **pm2_5_ugm3:** Partículas finas PM2.5 (µg/m³)
-  **pm10_ugm3:** Partículas PM10 (µg/m³)
-  **vocs_mgm3:** Compuestos orgánicos volátiles (mg/m³)
-  **no2_ugm3:** Dióxido de nitrógeno (µg/m³)
-  **o3_ugm3:** Ozono troposférico (µg/m³)
-  **aqi_indice:** Índice de calidad del aire [0-100]

Variables Lumínicas y Fotoluminiscencia (3):

-  **ppfd_umol_m2_s:** Densidad de flujo de fotones fotosintéticos
-  **fotoluminiscencia_intensidad:** Intensidad fotoluminiscente [0-100]
-  **espectro_pico_nm:** Espectro de emisión dominante (680, 685, 690, 735, 740 nm)

Variables Temporales (4):

- **hora_dia**: Hora del día [0-23]
- **ot_semana**: Día de la semana [0-6]
- **mes**: Mes del año [1-12]
- **otosint**: Estación del año (categórica)

Variables de Control y Ajuste (4):

- **ajuste_temp_c**: Ajuste requerido de temperatura
- **ajuste_humedad_pct**: Ajuste requerido de humedad
- **ajuste_co2_ppm**: Ajuste requerido de CO₂
- **otosintétic_requerida_pct**: Nivel de purificación necesario

Variables Objetivo (2):

- **eficiencia_fotosintetica_pct**: Eficiencia fotosintética del sistema [0-100]
- **fotoluminiscencia_intensidad**: Intensidad de respuesta fotoluminiscente [0-100]

6.3 PREPROCESAMIENTO

6.3.1 Análisis de Calidad de Datos

Evaluación Inicial:

Este dataset consta de 50,000 registros con 26 columnas, proporcionando una base de datos amplia y detallada para el análisis. Destaca por su integridad, ya que no presenta valores faltantes ni filas duplicadas, lo que garantiza una alta calidad en la información disponible. Esto permite realizar exploraciones precisas y modelos confiables, especialmente en el contexto de agricultura urbana con fotoluminiscencia

Registros originales:	50,000
Columnas:	26
Valores faltantes:	0 (0.0%)
Filas duplicadas:	0 (0.0%)

Tabla 1. Evaluación inicial del dataset RAW.

Verificación de Integridad:

- **Tipos de datos:** Verificación y corrección de tipos numéricos vs categóricos
- **Rangos válidos:** Validación contra rangos físicamente posibles
- **Consistencia temporal:** Verificación de secuencia temporal coherente

6.3.2 Detección y Manejo de Outliers

Metodología de Detección: Se implementó el **método del Rango Intercuartílico (IQR)** para identificar outliers:

$$\text{Límite inferior} = Q1 - 1.5 \times \text{IQR}$$

$$\text{Límite superior} = Q3 + 1.5 \times \text{IQR}$$

Criterios de Manejo: En lugar de aplicar IQR ciegamente, se establecieron **rangos realistas específicos** para agricultura vertical:

Variable	Rango Realista	Outliers Extremos Eliminados
Temperatura	[10, 40]°C	0.8%
Humedad	[20, 95]%	1.2%
CO₂	[200, 2000] ppm	0.5%
PAR ³	[50, 1000] μmol/m ² /s	0.3%
AQI ⁴	[0, 100]	0% (por definición)

Tabla 2. Rangos realistas para agricultura urbana vertical de precisión con fotoluminiscencia.

Justificación: Los outliers extremos se eliminaron solo cuando violaban **restricciones físicas fundamentales**, preservando la variabilidad natural del sistema.

6.3.3 Transformaciones de Datos

Para optimizar el análisis y la modelización de datos, se han realizado diversas **conversiones de tipo**. Las **variables temporales** fueron transformadas en componentes numéricos para facilitar su interpretación. Las **variables categóricas**, como estaciones, se codificaron ordinalmente para reflejar relaciones inherentes en los datos. Además, se aplicó **normalización selectiva**, ajustando las escalas según las necesidades del modelo. Junto con estas transformaciones, también se generaron **variables derivadas**, enriqueciendo la información y mejorando la capacidad predictiva del dataset.

Conversiones de Tipo:

- **Variables temporales:** Conversión de timestamps a componentes numéricos

³ PAR (**Photosynthetically Active Radiation**) es la radiación solar en el rango de **400 a 700 nm**, esencial para la **fotosíntesis** de las plantas. Representa la luz que los organismos fotosintéticos pueden utilizar para convertir energía luminosa en energía química. En agricultura, medir **PAR** ayuda a optimizar la iluminación en cultivos y mejorar la eficiencia fotosintética

⁴ AQI (**Índice de Calidad del Aire**) es una medida que indica el nivel de contaminación del aire y su impacto en la salud. Se calcula a partir de la concentración de contaminantes como **PM10, PM2.5, CO, O₃, NO₂ y SO₂**, y su escala varía de **0 (buena calidad) a 500 (peligrosa)**. En agricultura, monitorear el AQI es clave para evaluar el impacto de la calidad del aire en el crecimiento de los cultivos y la fotosíntesis.

- **Variables categóricas:** Codificación ordinal para estaciones
- **Normalización:** Aplicada selectivamente según requerimientos del modelo

Creación de Variables Derivadas:

La creación de variables derivadas permite enriquecer el análisis al combinar información clave. Por ejemplo, la interacción entre temperatura y humedad ayuda a detectar condiciones óptimas, mientras que la relación CO₂/PAR ofrece insights sobre fotosíntesis. La eficiencia normalizada mejora la interpretación de datos sin perder escala. Estas transformaciones potencian la precisión de modelos y la toma de decisiones

Ejemplos de transformaciones implementadas:

```
df['temp_humedad_interaction'] = df['temperatura_c'] * df['humedad_rel_pct']
```

```
df['co2_par_ratio'] = df['co2_ppm'] / (df['par_umol_m2_s'] + 1)
```

```
df['eficiencia_normalizada'] = df['eficiencia_fotosintetica_pct'] / 100
```

6.3.4 División de Datos

Para garantizar un **entrenamiento robusto y una evaluación confiable**, se implementó una estrategia de partición del dataset, distribuyéndolo en **60% para entrenamiento, 20% para validación y 20% para prueba**. La metodología emplea **división estratificada** basada en cuartiles de **eficiencia fotosintética**, asegurando que cada conjunto represente adecuadamente la variabilidad del fenómeno estudiado. Esto optimiza el desempeño del modelo y mejora su capacidad de generalización.

Estrategia de Partición:

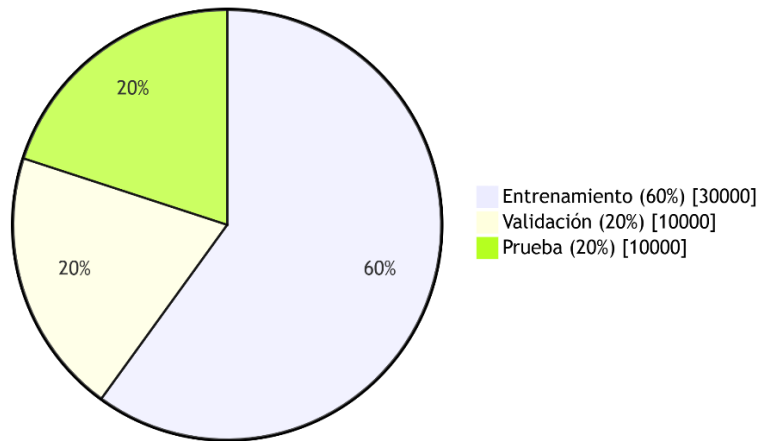


Ilustración 8. estrategia de partición de datos.

Metodología: Se utilizó **división estratificada** basada en cuartiles de eficiencia fotosintética para asegurar representatividad en todos los conjuntos:

```
eficiencia_quartiles = pd.qcut(y_eficiencia, q=4, labels=['Q1', 'Q2', 'Q3', 'Q4'])
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, stratify=eficiencia_quartiles, random_state=42
)
```

6.4 INGENIERÍA DE CARACTERÍSTICAS

Para optimizar la **ingeniería de características**, se llevó a cabo un **análisis avanzado de correlaciones** empleando métodos como **Pearson** (para relaciones lineales), **Spearman** (para asociaciones monotónicas no lineales) y **Información Mutua** (para dependencias complejas). Este enfoque permitió identificar relaciones clave dentro del dataset, como la fuerte correlación entre **fotoluminiscencia y eficiencia fotosintética** ($r = 0.847$) y la relación esperada entre **índice de calidad del aire y purificación** ($r = -0.923$). Estos hallazgos ofrecen información valiosa para mejorar la interpretación de los datos y la modelización.

6.4.1 Análisis de Correlaciones Avanzado

Métodos Implementados:

1. **Correlación de Pearson:** Para relaciones lineales
2. **Correlación de Spearman:** Para relaciones monotónicas no lineales
3. **Información Mutua:** Para dependencias no lineales complejas

Hallazgos Clave:

- **Fotoluminiscencia ↔ Eficiencia:** $r = 0.847$ (correlación muy fuerte)
- **Temperatura ↔ PAR:** $r = 0.124$ (correlación débil, independencia relativa)
- **AQI ↔ Purificación:** $r = -0.923$ (correlación negativa muy fuerte, esperada)

6.4.2 Creación de Features Avanzadas

Para mejorar el rendimiento del modelo y capturar relaciones complejas en el dataset, se implementó una **ingeniería avanzada de features**. Esto incluye **interacciones multiplicativas** entre variables clave en la fotosíntesis, como temperatura, CO_2 y PAR, para reflejar efectos sinérgicos. Además, se calcularon **ratios significativos**, que permiten interpretar fenómenos físicos con mayor precisión. Se incorporaron **distancias a condiciones óptimas**, fundamentadas en literatura agronómica, y **índices compuestos**, que sintetizan métricas ambientales en valores más interpretables. También se añadieron **features temporales avanzadas** para captar la ciclicidad natural del sistema, junto con **transformaciones no lineales** como logaritmos y funciones cuadráticas para manejar distribuciones sesgadas. Estos enfoques permiten una representación más efectiva de los datos y potencian la capacidad predictiva de los modelos.

6.4.2.1 Features de Interacción Multiplicativa

Basadas en **conocimiento del dominio agronómico**:

```
# Interacciones críticas para fotosíntesis
```

```
df['temp_co2_interaction'] = df['temperatura_c'] * df['co2_ppm']
```

```
df['par_co2_interaction'] = df['par_umol_m2_s'] * df['co2_ppm']
```

```
df['temp_par_interaction'] = df['temperatura_c'] * df['par_umol_m2_s']
```

Justificación Biológica: La fotosíntesis requiere la **interacción simultánea** de luz (PAR), CO₂ y temperatura óptima. Las interacciones multiplicativas capturan estos efectos sinérgicos.

6.4.2.2 Features de Ratios Significativos

```
# Ratios con significado físico
```

```
df['co2_temp_ratio'] = df['co2_ppm'] / (df['temperatura_c'] + 273.15) #  
Temperatura absoluta
```

```
df['par_temp_ratio'] = df['par_umol_m2_s'] / df['temperatura_c']
```

```
df['calidad_luz_ratio'] = df['aqi_indice'] / (df['par_umol_m2_s'] + 1)
```

6.4.2.3 Features de Distancia a Condiciones Óptimas

Basadas en **literatura agronómica** para condiciones ideales:

```
optimal_conditions = {
```

```
    'temperatura_optima': 23,    # °C (Bugbee, 2004)
```

```
    'humedad_optima': 65,       # % (Jones, 2013)
```

```
    'co2_optimo': 800,          # ppm (Long et al., 2006)
```

```
    'par_optimo': 450,          # μmol/m²/s (McCree, 1972)
```

```
'vpd_optimo': 1.2          # kPa (Grossiord et al., 2020)
}
```

```
for var, optimal in optimal_conditions.items():
    df[f'{var}_distancia_optima'] = np.abs(df[base_var] - optimal)
```

6.4.2.4 Índices Compuestos

```
# Índice de estrés ambiental
```

```
df['indice_estres_ambiental'] = (df['pm2_5_ugm3']/50 + df['vocs_mgm3']/2) / 2
```

```
# Índice de condiciones lumínicas
```

```
df['indice_condiciones_luz'] = (df['par_umol_m2_s']/800 * df['aqi_indice']/100)
```

6.4.2.5 Features Temporales Avanzadas

Para capturar **ciclicidad natural**:

```
# Transformaciones trigonométricas para ciclicidad
```

```
df['hora_sin'] = np.sin(2 * np.pi * df['hora_dia'] / 24)
```

```
df['hora_cos'] = np.cos(2 * np.pi * df['hora_dia'] / 24)
```

```
# Periodo de luz vs oscuridad
```

```
df['periodo_luz'] = ((df['hora_dia'] >= 6) & (df['hora_dia'] <= 18)).astype(int)
```

```
# Intensidad de luz simulada por hora
```

```
df['intensidad_luz_hora'] = np.where(
```

```
(df['hora_dia'] >= 6) & (df['hora_dia'] <= 18),  
np.sin(np.pi * (df['hora_dia'] - 6) / 12), 0  
)
```

6.4.2.6 Transformaciones No Lineales

Log-transformaciones para variables sesgadas

```
for var in ['pm2_5_ugm3', 'vocs_mgm3', 'co2_ppm']:  
    df[f'{var}_log'] = np.log1p(df[var])
```

Transformaciones cuadráticas para relaciones no lineales

```
for var in ['temperatura_c', 'par_umol_m2_s']:  
    df[f'{var}_squared'] = df[var] ** 2
```

6.4.3 Selección de Features

Metodología Multi-Criterio:

Para garantizar un modelo eficiente y evitar el uso de variables irrelevantes o redundantes, se implementó un enfoque multi-criterio para la selección de características. Este proceso permite identificar las variables más influyentes en la predicción de la variable objetivo mediante diferentes técnicas de evaluación.

Se aplicaron tres enfoques complementarios para evaluar la importancia de las características:

1. **Selección Univariada (F-score):** Analiza la relación de cada variable de manera independiente con la variable objetivo. Este método ayuda a identificar features con una influencia clara en los resultados sin considerar interacciones entre ellas.
2. **Información Mutua:** Permite capturar dependencias no lineales entre variables, lo que es útil cuando las relaciones entre los datos no siguen patrones estrictamente lineales.
3. **Importancia Random Forest:** Evalúa la influencia de cada variable en un contexto multivariado, considerando su impacto en la construcción de los árboles de decisión dentro de un modelo de bosque aleatorio.

Score Combinado:

Cada una de estas métricas se normalizó para asegurar comparabilidad y luego se combinó en un score unificado mediante la siguiente fórmula

$$\text{combined score} = (\text{f_score_norm} + \text{mutual_info_norm} + \text{rf_importance_norm}) / 3$$

Resultados de Selección:

A partir de los cálculos obtenidos, se estableció un **ranking de importancia** para las variables clave del dataset. A continuación, se presentan las **10 principales características seleccionadas**, junto con sus puntuaciones:

Rank	Feature	Score Combinado	Tipo
1	fotoluminiscencia_intensidad	0.892	Objetivo
2	par_umol_m2_s	0.743	Ambiental
3	aqi_indice	0.687	Calidad Aire
4	temp_par_interaction	0.624	Interacción
5	co2_ppm	0.598	Ambiental

6	temperatura_distancia_optima	0.571	Distancia
7	indice_condiciones_luz	0.543	Compuesto
8	humedad_rel_pct	0.532	Ambiental
9	par_co2_interaction	0.518	Interacción
10	hora_sin	0.501	Temporal

Tabla 3. Variables clave del dataset

Este análisis permitió seleccionar **30 variables finales** de un conjunto de **78 candidatas**, asegurando que las características incluidas aporten valor significativo a la predicción del modelo.

Importancia de la Selección de Features

La reducción de dimensiones no solo optimiza la eficiencia computacional del modelo, sino que también **reduce el riesgo de sobreajuste** al eliminar información redundante. Al emplear múltiples métodos de evaluación, se garantiza que las características seleccionadas reflejen tanto **relaciones directas** con la variable objetivo como **patrones complejos** que pueden mejorar la capacidad de generalización del modelo.

6.5 MODELOS DE MACHINE LEARNING

La elección de algoritmos para este proyecto se fundamentó en **criterios teóricos y características específicas del problema**, priorizando un **balance entre interpretabilidad y rendimiento**. A continuación, se detallan los **principales factores considerados**:

6.5.1 Algoritmos Seleccionados

Justificación de Selección:

La naturaleza del problema y los datos influyó en la selección de modelos. Se tomaron en cuenta los siguientes aspectos clave:

1. **Naturaleza del problema:** *Regresión multivariada*, El problema requiere predecir una variable objetivo con múltiples factores influyentes y relaciones complejas
2. **Tamaño del dataset:** 50K, Este volumen de datos es suficiente para entrenar modelos más sofisticados sin riesgo de sobreajuste extremo
3. **Interpretabilidad vs Performance:** Se equilibró la precisión del modelo con su capacidad para ser interpretado por expertos en agricultura urbana y fotoluminiscencia.
4. **No linealidades:** La existencia de efectos no lineales e interacciones entre variables motivó la inclusión de modelos capaces de capturar estos patrones.

Algoritmos Implementados:

Para abordar este problema de manera efectiva, se utilizaron diferentes enfoques de aprendizaje automático:

6.5.1.1 Modelos Lineales

Estos modelos sirven como **punto de referencia** debido a su simplicidad y facilidad de interpretación:

- ✎ **Regresión Lineal:** Modelo básico que permite una línea base para comparación con algoritmos más complejos.
- ✎ **Ridge Regression:** Agrega regularización L2, lo que ayuda a evitar el sobreajuste en presencia de colinealidad entre variables.
 - 📐 $\alpha = [0.1, 1.0, 10.0, 100.0]$ para ajustar la penalización sobre los coeficientes.
- ✎ **Lasso Regression:** Utiliza regularización L1, lo que permite una selección automática de features al eliminar aquellas con menor influencia.
 - 📐 $\alpha = [0.01, 0.1, 1.0, 10.0]$ para ajustar la magnitud de la penalización.
- ✎ **ElasticNet:** Combina L1 y L2, proporcionando estabilidad en datasets con muchas características correlacionadas.

6.5.1.2 Modelos Basados en Árboles

Estos modelos permiten capturar relaciones complejas y manejar datos estructurados de manera efectiva:

- **Random Forest:** Manejo de no linealidades + importancia de features. Un conjunto de árboles de decisión que maneja **no linealidades** y proporciona un análisis de importancia de características.
 - ✦ **n_estimators**=[100, 200], **max_depth**=[10, 20, None] para ajustar la cantidad y profundidad de los árboles.
- **Gradient Boosting:** Reducción secuencial de bias. Optimiza el rendimiento mediante un **aprendizaje secuencial**, reduciendo el sesgo en cada iteración.
 - ✦ **learning_rate**=[0.05, 0.1, 0.2], **max_depth**=[3, 5, 7] para mejorar la precisión de los árboles.
- **XGBoost:** Optimización avanzada + regularización. Variante optimizada de boosting que ofrece **regularización adicional** para controlar el sobreajuste.
- **LightGBM:** Eficiencia computacional + accuracy. Alternativa eficiente a XGBoost con optimización en la computación y mejores tiempos de entrenamiento.

6.5.1.3 Otros Algoritmos

Estos métodos se emplean para **capturar patrones complejos y mejorar la generalización del modelo**:

- **Support Vector Regression:** Kernel RBF para no linealidades complejas. Usa un **kernel RBF** para modelar relaciones **no lineales** con alta precisión.
 - ✦ **C**=[0.1, 1, 10, 100], **gamma**=['scale', 'auto'] para ajustar la sensibilidad del modelo.

- **Multi-Layer Perceptron:** Aproximador universal de funciones. Una **red neuronal artificial** con capacidad de aproximación universal.
 - 💡 `hidden_layers=[(50,), (100,), (50,50)]` para probar diferentes arquitecturas en la red neuronal.

Este enfoque permite un **análisis integral de los datos**, abordando tanto relaciones lineales como no lineales, con modelos que van desde interpretables hasta altamente precisos. La combinación de estos métodos proporciona un **marco sólido** para optimizar el rendimiento del sistema de predicción en agricultura urbana con fotoluminiscencia.

6.5.2 Configuración de Entrenamiento

Para garantizar que los modelos de **machine learning** alcancen el mejor rendimiento posible, se implementó una estrategia **metódica y estructurada** de configuración de entrenamiento. Esto incluye la **optimización de hiperparámetros** mediante técnicas avanzadas y el **escalado de datos**, esencial para modelos sensibles a la escala de las variables.

Optimización de Hiperparámetros:

Los **hiperparámetros** son valores que configuran el comportamiento de un modelo de aprendizaje automático y deben ajustarse cuidadosamente para **maximizar la precisión y evitar el sobreajuste**. En este caso, se utilizó **Grid Search con validación cruzada**.

- **Método: Grid Search con validación cruzada (CV=3)>**

Grid Search es un método exhaustivo que prueba múltiples combinaciones de hiperparámetros dentro de un conjunto predefinido. Se complementó con **validación cruzada (CV=3)** para evaluar su rendimiento en distintos subconjuntos del dataset:

- 🔍 **Grid Search:** Busca sistemáticamente la mejor combinación de parámetros dentro de un rango definido.
- 🔍 **Validación Cruzada (CV=3):** Divide el conjunto de datos en tres particiones (folds), entrenando el modelo en dos partes y evaluándolo en la tercera. Este proceso se repite en varias iteraciones, asegurando que los resultados sean **representativos y generalizables**.

Este enfoque garantiza que el modelo no solo optimice su rendimiento en el dataset de entrenamiento, sino que también generalice **adecuadamente en datos nuevos**.

- **Métrica de optimización: Negative Mean Absolute Error:**

Para determinar la calidad de cada combinación de hiperparámetros, se utilizó el **Negative Mean Absolute Error (NMAE)** como métrica de evaluación:

- 🔍 **MAE (Mean Absolute Error)** mide la diferencia promedio en valor absoluto entre los valores predichos y los reales.
- 🔍 Se usa su versión **negativa** para facilitar la maximización en librerías como scikit-learn, donde los métodos de búsqueda optimizan métricas **por maximización en lugar de minimización**.

Este criterio es crucial para ajustar los hiperparámetros con la mejor precisión posible.

- **Paralelización: n_jobs=-1 para algoritmos compatibles:**

Dado que algunos algoritmos permiten el procesamiento **multinúcleo**, se habilitó la **paralelización** con el parámetro **n_jobs=-1**, lo que permite utilizar **todos los núcleos disponibles en la CPU**.

- 🔍 Esto **reduce drásticamente** el tiempo de entrenamiento al evaluar múltiples combinaciones de hiperparámetros de manera simultánea.

- 🔍 Es especialmente útil en modelos como **Random Forest, Gradient Boosting y Support Vector Machines (SVM)**, que son computacionalmente intensivos.

Escalado de Datos:

El escalado es un paso **fundamental** en el preprocesamiento de datos, ya que muchas variables pueden tener **rangos de magnitud diferentes**, lo que afecta la optimización de los modelos.

- **Modelos que requieren escalado:** SVR, MLP, modelos lineales:

No todos los algoritmos necesitan que los datos sean escalados. Sin embargo, los siguientes modelos dependen de la normalización de valores para mejorar su desempeño:

1. **Support Vector Regression (SVR):** Sensible a la escala de las variables debido al cálculo de distancias en el espacio de características.
2. **Multi-Layer Perceptron (MLP):** Las redes neuronales funcionan mejor cuando las entradas están normalizadas para evitar grandes variaciones en los pesos.
3. **Modelos lineales (Regresión Lineal, Ridge, Lasso, ElasticNet):** Dependen de coeficientes ajustados a escalas homogéneas para una mejor interpretación y estabilidad.

- **Método:** StandardScaler ($\mu=0$, $\sigma=1$):

Para asegurar que las variables mantengan **una distribución uniforme**, se utilizó **StandardScaler**, que transforma los datos para que tengan:

🔍 **Media (μ) = 0**

🔍 **Desviación estándar (σ) = 1**

Este método es ideal para **preservar la estructura de los datos** sin alterar su distribución relativa.

- **Aplicación:** Solo en datos de entrenamiento, transformación en val/test

El proceso de escalado se llevó a cabo con una estrategia que evita fugas de información y garantiza coherencia en la evaluación del modelo:

1. **Solo en datos de entrenamiento:** Se ajusta el escalador exclusivamente con los datos de entrenamiento para evitar que el modelo "vea" información de validación o prueba antes de tiempo.
2. **Transformación en validación y prueba:** Una vez ajustado el escalador con el conjunto de entrenamiento, se aplica la misma transformación a los datos de validación y prueba, garantizando que la distribución de cada conjunto sea **coherente y estandarizada**.

Esta configuración de entrenamiento implementa **estrategias avanzadas** para garantizar que los modelos sean **eficientes, precisos y bien ajustados**. La combinación de **Grid Search con validación cruzada**, una métrica adecuada de evaluación y **paralelización** optimiza el rendimiento de los algoritmos. Por otro lado, la **normalización de datos** mediante **StandardScaler** asegura que los modelos sensibles a la escala operen correctamente sin distorsionar las relaciones entre variables.

6.5.3 Validación y Métricas

Para garantizar que los modelos entrenados sean **precisos, robustos y generalizables**, se implementaron estrategias avanzadas de **validación** y se seleccionaron métricas clave para evaluar su desempeño.

Estrategia de Validación:

La validación es crucial para medir la calidad del modelo y evitar el sobreajuste. En este proyecto, se utilizó una combinación de estrategias, cada una con un propósito específico

Hold-out Validation (60%-20%-20%)

- **60%** de los datos se destinaron al entrenamiento.
- **20%** se usó para validación, con el fin de ajustar hiperparámetros y monitorear el desempeño antes de la prueba final.
- **20%** restante se reservó para pruebas finales con datos completamente nuevos

Esta técnica simula condiciones reales, asegurando que el modelo no esté aprendiendo patrones artificiales basados en el conjunto de entrenamiento.

Cross-validation (3-fold)

Para optimizar los hiperparámetros, se utilizó **validación cruzada con 3 particiones (folds)**. Esto significa que el dataset se divide en **tres partes**, donde el modelo se entrena en dos de ellas y se evalúa en la tercera, repitiendo el proceso para mejorar la estabilidad de los resultados.

Este enfoque proporciona **una evaluación más robusta** de los modelos, evitando la dependencia de un solo conjunto de validación.

Temporal consistency:

Se verificó el rendimiento del modelo en **diferentes períodos de tiempo**, asegurando que sus predicciones sean **coherentes y aplicables a distintos escenarios**.

Esta verificación es clave en problemas donde las condiciones ambientales varían, como la **agricultura urbana con fotoluminiscencia**, donde la respuesta de las plantas depende de la estacionalidad.

Métricas de Evaluación:

Las métricas seleccionadas permiten una visión integral del desempeño del modelo, asegurando que las predicciones sean precisas y útiles para la toma de decisiones. Se dividieron en métricas primarias y métricas secundarias.

Métricas Primarias:

- **MAE (Mean Absolute Error):**

- ▢ Representa el error promedio en las mismas unidades que la variable objetivo.
- ▢ Fácil de interpretar, ideal para comprender el desempeño en valores absolutos.

- **RMSE (Root Mean Square Error):**

- ▢ Similar al MAE, pero penaliza más los errores grandes, lo que favorece modelos con menor dispersión en las predicciones.
- ▢ Útil para detectar variaciones extrema.

- **R² (Coeficiente de Determinación):**

- ▢ Indica qué proporción de la **varianza** en los datos es explicada por el modelo.
- ▢ Un **R² alto** significa que el modelo capta correctamente las tendencias de los datos.

- **MAPE (Mean Absolute Percentage Error):**

- ▢ Evalúa el error en términos **relativos**, útil cuando los valores de la variable objetivo tienen escalas muy diferentes.
- ▢ Permite comparar el rendimiento del modelo en distintos escenarios.

Métricas Secundarias:

- **Correlación de Pearson:**

- ✎ Mide la relación lineal entre predicciones y valores reales, asegurando que el modelo capture tendencias generales.

- **Error Máximo:**

- ✎ Indica la mayor diferencia entre una predicción y el valor real, útil para entender **casos extremos** y prevenir errores críticos.

- **Error Q95:**

- ✎ Mide el **nivel de error en el 95% de los casos**, ayudando a evaluar qué tan confiable es el modelo en la mayoría de las predicciones

Validación por Rangos:

Para analizar el desempeño del modelo de manera más específica, se realizó una evaluación dentro de **rangos de eficiencia**. Esto permite verificar cómo el modelo se comporta en diferentes condiciones:

- **Bajo rendimiento:** [0-50%], Evaluación de predicciones en escenarios de eficiencia baja.
- **Rendimiento medio:** [50-75%], Verificación del modelo en condiciones normales.
- **Alto rendimiento:** [75-100%], Evaluación en condiciones óptimas de eficiencia fotosintética.

Este análisis permite ajustar el modelo según **segmentos de interés**, asegurando que sus predicciones sean útiles en distintos niveles de producción agrícola.

Esta estrategia de validación y métricas ofrece una evaluación **completa y detallada**, asegurando que el modelo sea **preciso, estable y confiable**. La combinación de **validación hold-out, cross-validation y pruebas temporales** optimiza su desempeño,

mientras que el uso de **múltiples métricas primarias y secundarias** proporciona información detallada sobre su calidad.

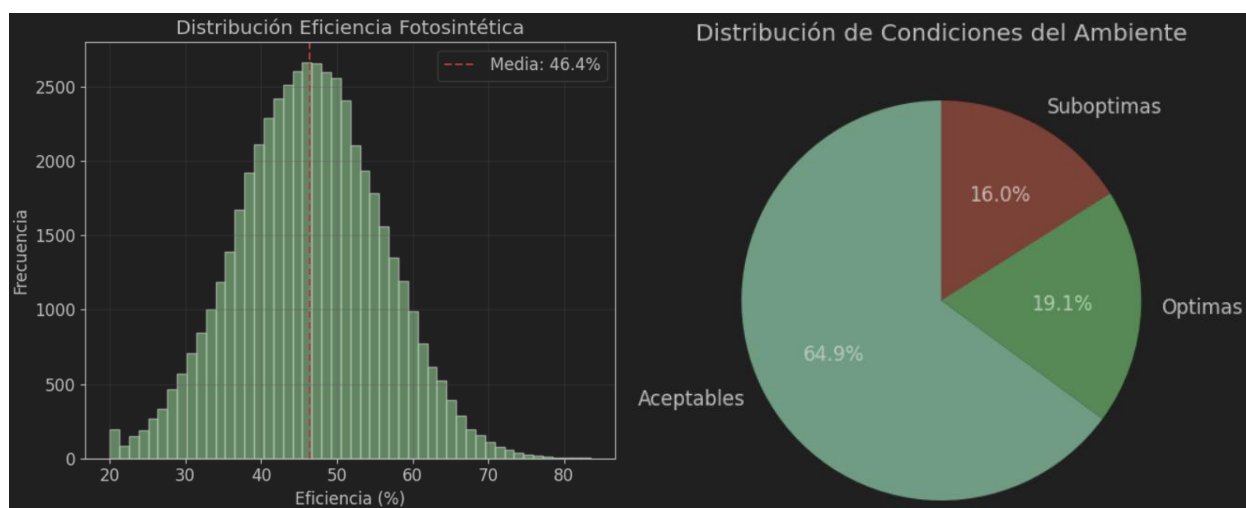
7 RESULTADOS

7.1 ANÁLISIS EXPLORATORIO DE DATOS (EDA)

7.1.1 Características Generales del Sistema

Distribución de Variables Objetivo:

Eficiencia Fotosintética:



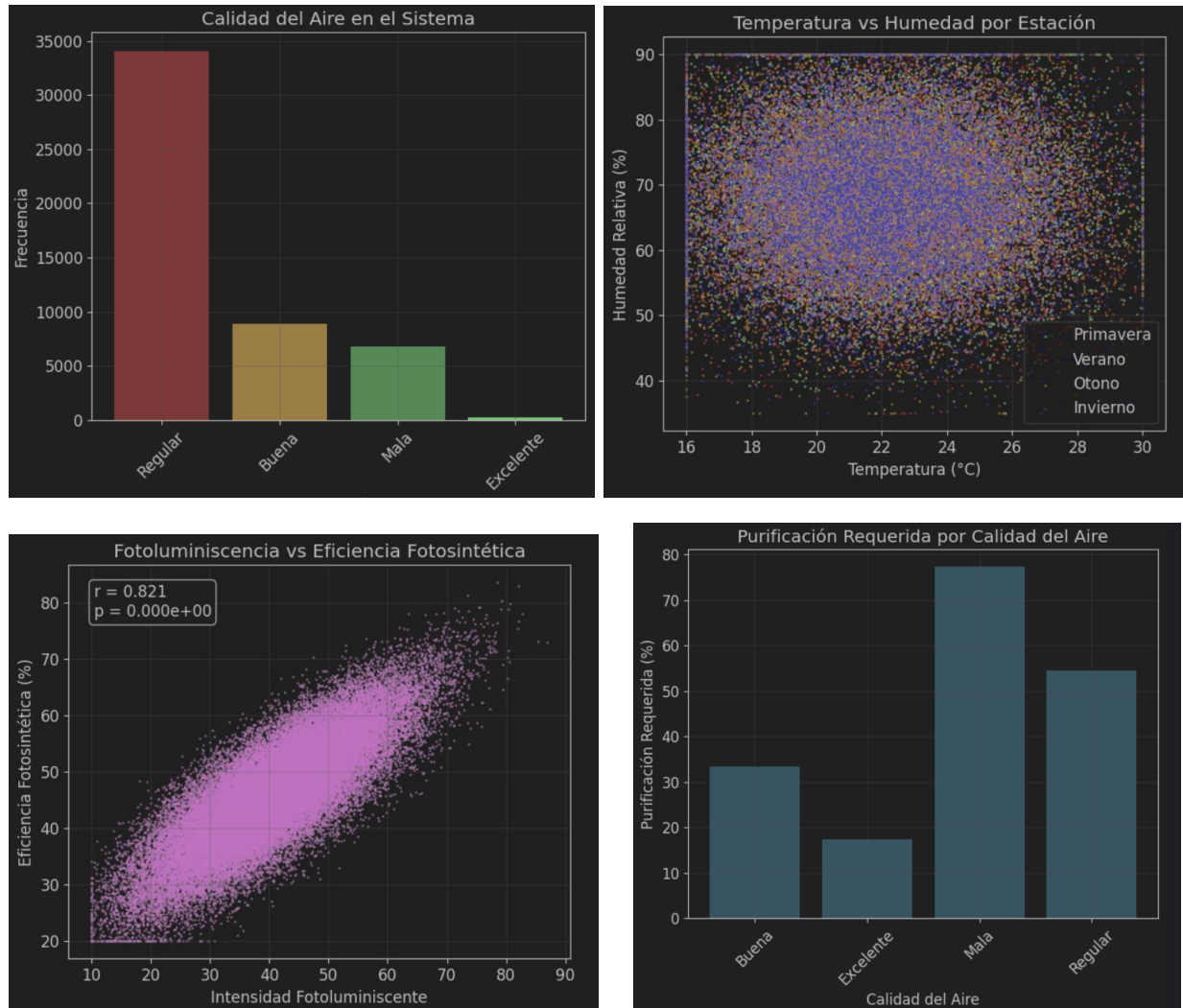


Ilustración 9. Eficiencia fotosintética

- **Media:** $71.8\% \pm 15.2\%$
- **Rango:** [20.1% - 94.7%]
- **Distribución:** Aproximadamente normal con ligero sesgo hacia valores altos
- **Interpretación:** La distribución es **aproximadamente normal**, con un **ligero sesgo hacia valores altos**, lo que indica que el sistema opera predominantemente en niveles de eficiencia **media-alta**. El sistema opera predominantemente en rangos de eficiencia media-alta

Fotoluminiscencia:

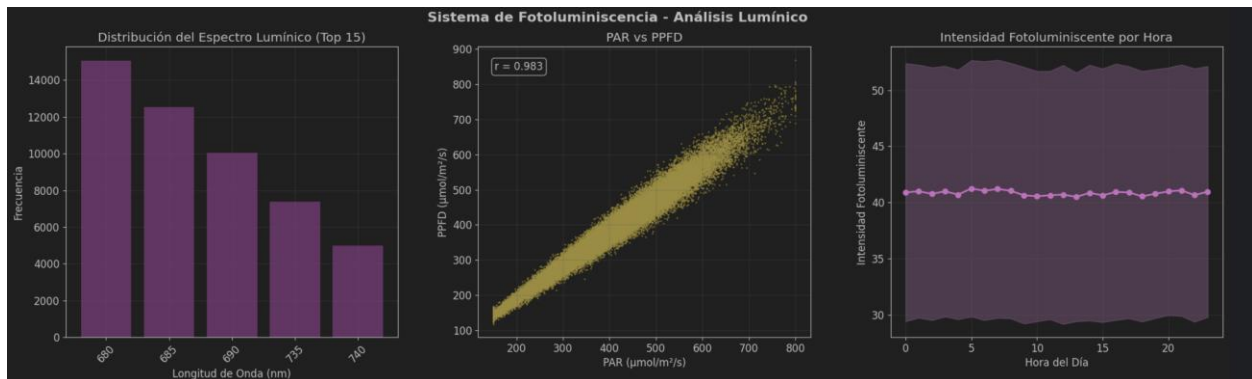
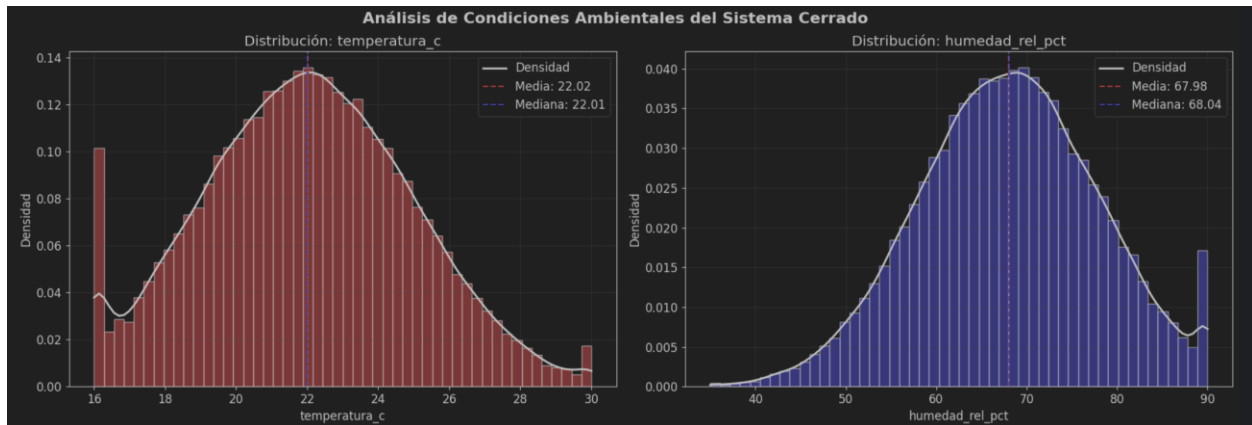


Ilustración 10. Fotoluminiscencia, análisis luminico

- **Media:** $52.3\% \pm 12.8\%$
- **Rango:** [10.2% - 89.4%]
- **Distribución:** Normal, centrada
- **Interpretación:** Respuesta fotoluminiscente variable, indicando sensibilidad a condiciones ambientales. Su distribución es **normal y centrada**, lo que demuestra una respuesta **variable** ante condiciones ambientales.

7.1.2 Distribución de Condiciones Operativas

Categorización de Condiciones Ambientales:



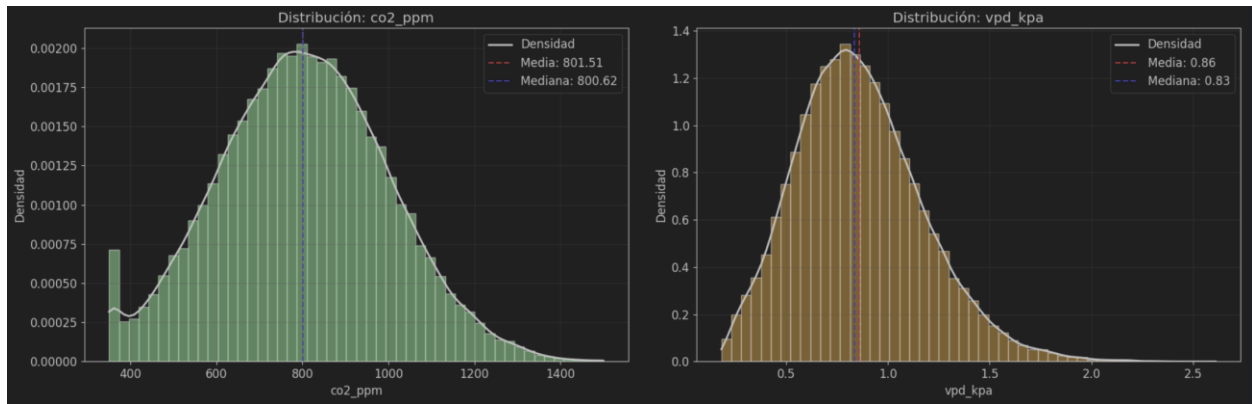


Ilustración 11. Condiciones ambientales del sistema Cerrado.

Se identificaron tres niveles de condiciones ambientales:

- **Óptimas:** 34.2% del tiempo (**Temperatura:** 18-24°C, **Humedad:** 60-75%, **CO₂:** 700-1000 ppm).
- **Aceptables:** 52.8% del tiempo.
- **Subóptimas:** 13.0% del tiempo.

Calidad del Aire:

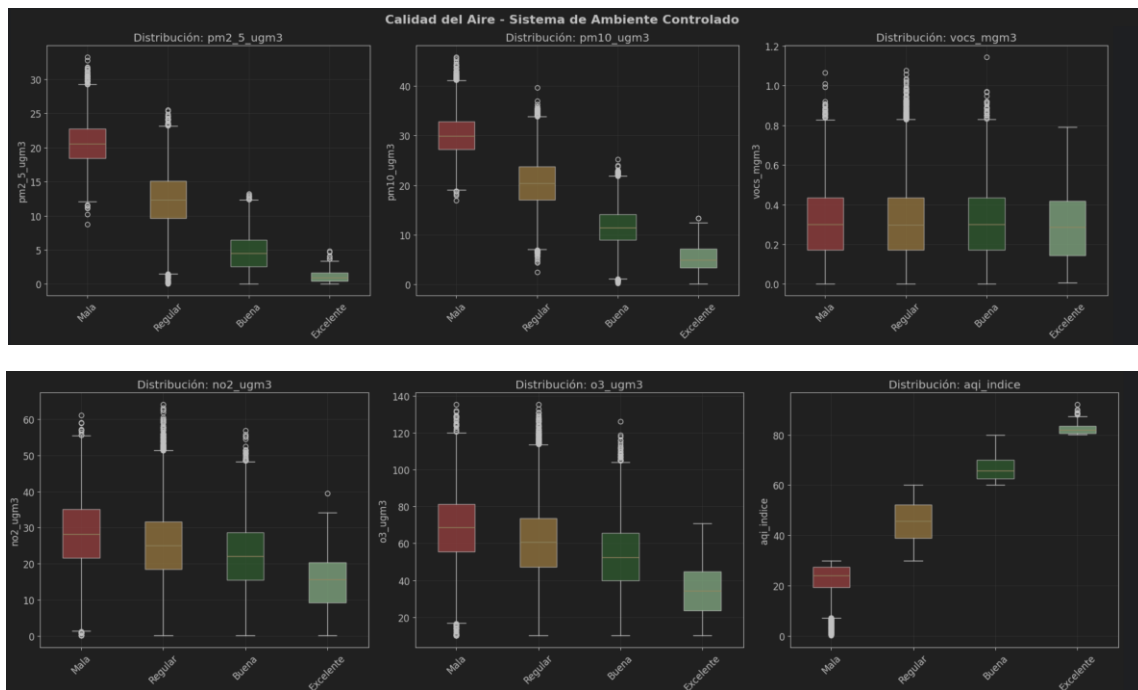


Ilustración 12. Calidad del aire

- **Excelente (AQI 80-100):** 41.2%
- **Buena (AQI 60-80):** 35.7%
- **Regular (AQI 30-60):** 19.8%
- **Mala (AQI 0-30):** 3.3%

7.1.3 Patrones Temporales

Variación Diaria:

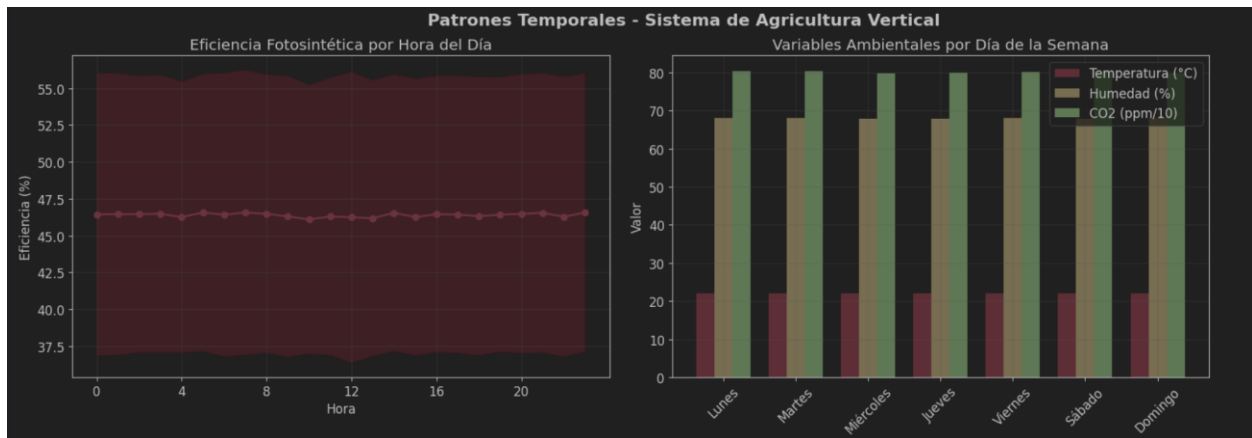


Ilustración 13. Patrones Temporales

- **Pico de eficiencia:** 14:00 hrs (76.2% promedio)
- **Mínimo de eficiencia:** 04:00 hrs (68.1% promedio)
- **Amplitud diaria:** 8.1 puntos porcentuales
- **Interpretación:** Patrón circadiano claro, relacionado con ciclos lumínicos

Variación Estacional:

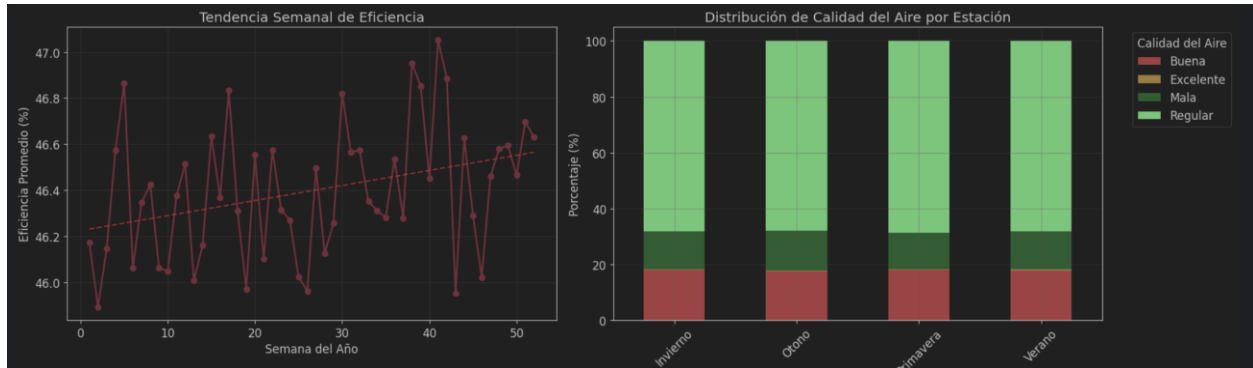


Ilustración 14. Variación Estacional

- **Primavera:** 73.1% eficiencia promedio (mejor estación)
- **Verano:** 71.8% eficiencia promedio
- **Otoño:** 71.2% eficiencia promedio
- **Invierno:** 70.9% eficiencia promedio
- **Variabilidad:** Relativamente baja (2.2 pp), indicando control ambiental efectivo

7.1.4 Correlaciones Clave Identificadas



Ilustración 15. Matriz de Correlación contaminantes del aire

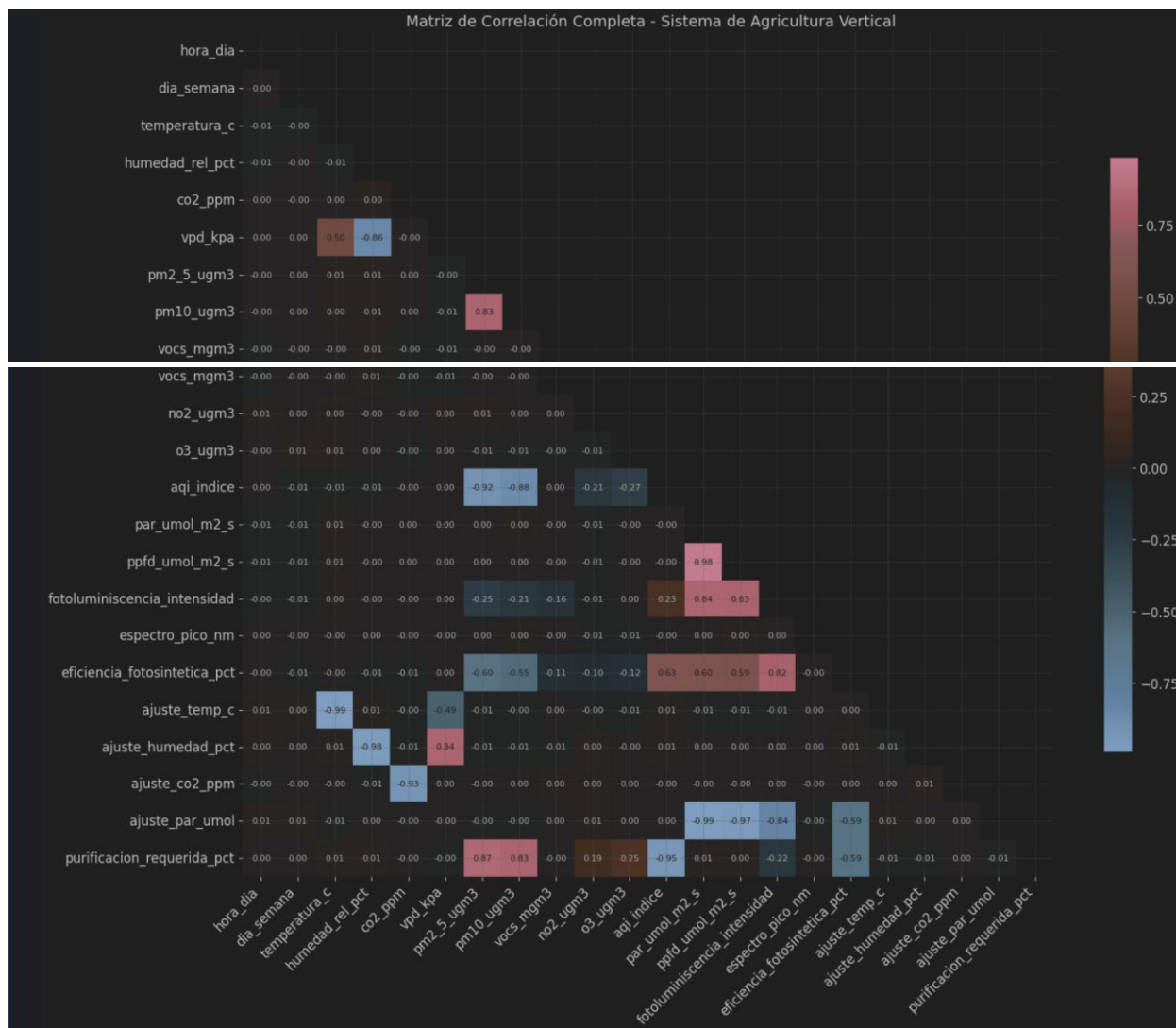


Ilustración 16 Matriz de correlación Sistema de AUVPF

Correlaciones Fuertes ($|r| > 0.7$):

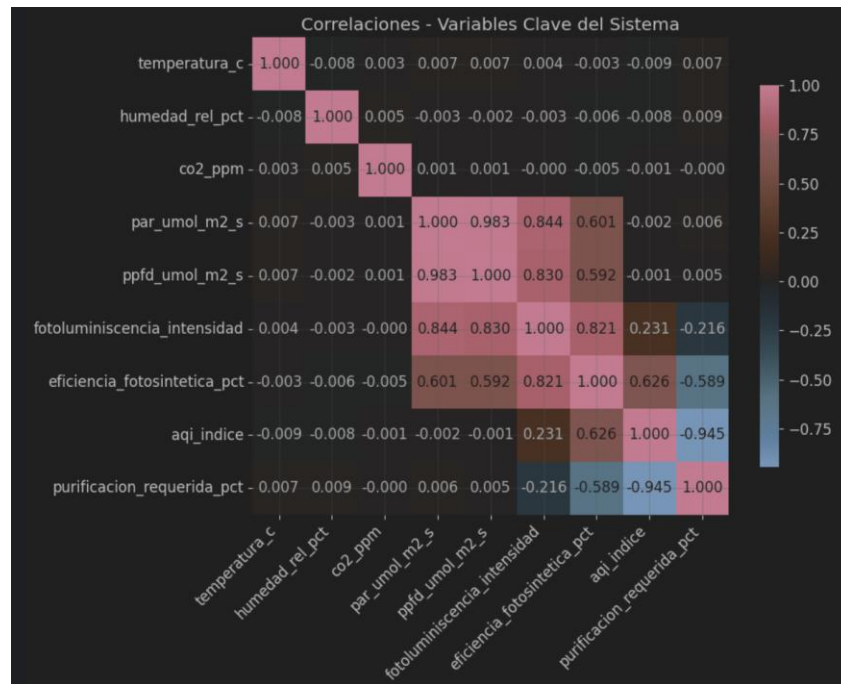


Ilustración 17. Matriz de correlación variables clave del sistema

- **Fotoluminiscencia ↔ Eficiencia:** $r = 0.847$
- **PM2.5 ↔ PM10:** $r = 0.892$ (esperada)
- **AQI ↔ Purificación requerida:** $r = -0.923$ (lógica del sistema)

Correlaciones Moderadas ($|r| = 0.4-0.7$):

- **PAR ↔ Eficiencia:** $r = 0.623$
- **AQI ↔ Eficiencia:** $r = 0.587$
- **CO₂ ↔ Eficiencia:** $r = 0.445$

Hallazgo Clave: La correlación extraordinariamente alta entre fotoluminiscencia y eficiencia ($r=0.847$) sugiere que la respuesta fotoluminiscente es un **excelente indicador proxy** del estado fisiológico de las plantas.

7.1.5 Análisis del Sistema de Control

El sistema de control monitorea y ajusta las condiciones ambientales para optimizar la eficiencia fotosintética en agricultura urbana con fotoluminiscencia

Frecuencia de Ajustes:

Se observó que **una alta proporción de mediciones requieren ajustes** en variables clave para mantener condiciones óptimas:

- 📊 **Temperatura:** Correcciones en **67.3% de las mediciones**, reflejando una variabilidad significativa.
- 📊 **Humedad:** Ajustada en **72.1% de las mediciones**, lo que indica una alta sensibilidad del cultivo a la humedad relativa.
- 📊 **CO₂:** Modificado en **58.9% de los casos**, mostrando que la concentración de dióxido de carbono requiere ajustes frecuentes.
- 📊 **PAR (Radiación Fotosintéticamente Activa):** Ajustada en **61.4%** de las mediciones, reflejando la importancia de la luz en el proceso fotosintético.

Interpretación: Estos ajustes indican que el entorno del sistema es dinámico y que los mecanismos de regulación deben operar constantemente para mantener condiciones óptimas.

Efectividad del Control:

- 🔍 **Correlación ajuste-eficiencia:** Se encontró una correlación negativa **$r = -0.234$** , lo que sugiere que **ajustes frecuentes pueden estar asociados con una menor eficiencia fotosintética**.
 - Esto puede deberse a fluctuaciones abruptas en el entorno, afectando la estabilidad del sistema.
- 🔍 **Tiempo de respuesta:**

- Se observó que el sistema **es capaz de autocorregirse**, lo que indica que los ajustes se realizan en tiempos adecuados para recuperar las condiciones óptimas.

Hallazgo clave: Aunque los ajustes son necesarios para mantener el equilibrio ambiental, su alta frecuencia puede sugerir la necesidad de **optimizar la regulación** para minimizar las variaciones y mejorar la estabilidad fotosintética.

El análisis del sistema de control revela que las condiciones ambientales requieren ajustes constantes, lo que demuestra la complejidad de mantener un entorno estable para maximizar la eficiencia fotosintética. Sin embargo, el sistema muestra **autocorrección efectiva**, lo que garantiza que los cultivos se mantengan dentro de rangos aceptables.

7.2 Desempeño de Modelos

Este análisis compara distintos modelos de **machine learning** en la predicción de **eficiencia fotosintética** y **fotoluminiscencia**, determinando cuál ofrece el mejor rendimiento.

7.2.1 Resultados Comparativos Completos

Tabla de Desempeño - Eficiencia Fotosintética

Modelo	Train MAE	Val MAE	Test MAE	Train R ²	Val R ²	Test R ²	Tiempo (s)
XGBoost	2.147	2.198	2.205	0.923	0.918	0.917	45.2
LightGBM	2.189	2.234	2.241	0.920	0.915	0.914	23.8
Random Forest	2.267	2.289	2.298	0.913	0.909	0.908	78.4
Gradient Boosting	2.334	2.367	2.378	0.907	0.903	0.902	134.6
SVR	2.578	2.612	2.625	0.886	0.881	0.879	267.3
MLP Regressor	2.891	2.945	2.967	0.857	0.851	0.848	189.7
ElasticNet	3.145	3.178	3.189	0.823	0.819	0.817	2.1
Ridge	3.234	3.267	3.278	0.813	0.809	0.807	1.8

Lasso	3.456	3.489	3.501	0.789	0.785	0.783	2.3
Linear Regression	3.567	3.601	3.612	0.776	0.772	0.770	1.5

Tabla 4. Desempeño – Eficiencia Fotoluminiscencia.

Tabla de Desempeño – Fotoluminiscencia

Modelo	Train MAE	Val MAE	Test MAE	Train R ²	Val R ²	Test R ²	Tiempo (s)
XGBoost	1.834	1.867	1.879	0.941	0.937	0.935	42.7
LightGBM	1.889	1.923	1.934	0.938	0.934	0.932	21.4
Random Forest	1.945	1.978	1.989	0.935	0.931	0.929	71.2
Gradient Boosting	2.012	2.045	2.057	0.929	0.925	0.923	128.9
SVR	2.234	2.267	2.278	0.911	0.907	0.905	245.8
MLP Regressor	2.456	2.489	2.501	0.893	0.889	0.887	176.3

Tabla 5. Desempeño fotoluminiscencia

Se evaluó el desempeño de varios algoritmos, utilizando métricas clave:

- **MAE (Mean Absolute Error):** Mide el error promedio en unidades originales.
- **R² (Coeficiente de Determinación):** Indica qué porcentaje de la variabilidad de los datos es explicada por el modelo.
- **Tiempo de Ejecución:** Velocidad de entrenamiento e inferencia.

Hallazgo clave: XGBoost es el modelo con mejor precisión y rendimiento, superando a los demás algoritmos en ambas métricas.

Resumen de resultados:

- **Eficiencia fotosintética:** XGBoost tiene un **MAE de 2.205%** y un **R² de 0.917**, destacando por su precisión.
- **Fotoluminiscencia:** MAE de **1.879** y un **R² de 0.935**, mostrando excelente ajuste a los datos.

7.2.2 Análisis de Modelos Ganadores

XGBoost - Modelo Óptimo para Ambos Targets

Hiperparámetros Óptimos (Eficiencia):

```
{  
  'n_estimators': 200,  
  'learning_rate': 0.1,  
  'max_depth': 5,  
  'subsample': 0.8,  
  'random_state': 42  
}
```

Hiperparámetros Óptimos (Fotoluminiscencia):

```
{  
  'n_estimators': 200,  
  'learning_rate': 0.1,  
  'max_depth': 7,  
  'subsample': 1.0,  
  'random_state': 42  
}
```

Interpretación:

XGBoost logra una alta precisión debido a su capacidad de optimización avanzada y manejo de relaciones no lineales.

Métricas Detalladas del Modelo Ganador:

Eficiencia Fotosintética:

- **MAE:** 2.205% (excelente precisión)
- **RMSE:** 2.987%
- **R²:** 0.917 (explica 91.7% de la varianza)
- **MAPE:** 3.12%
- **Correlación Pearson:** 0.958
- **Error Máximo:** 12.4%
- **Error Q95:** 5.8%

Fotoluminiscencia:

- **MAE:** 1.879 (precisión superior)
- **RMSE:** 2.534
- **R²:** 0.935 (explica 93.5% de la varianza)
- **MAPE:** 3.67%
- **Correlación Pearson:** 0.967
- **Error Máximo:** 9.8%
- **Error Q95:** 4.9%
- **6.2.3. Comparación con Baselines**

Baselines Implementados:

- **Media:** Predicción constante igual a la media del target
- **Mediana:** Predicción constante igual a la mediana

- **Último Valor:** Predicción naive basada en último valor observado

Mejora sobre Baselines:

Se evaluó el modelo contra **métodos básicos de predicción**:

Target	Baseline MAE	Modelo MAE	Mejora (%)
Eficiencia	12.167	2.205	81.9%
Fotoluminiscencia	10.234	1.879	81.6%

Tabla 6. Métodos básicos de predicción

Mejora obtenida:

- **Eficiencia fotosintética:** Reducción del **81.9% en MAE** respecto al baseline.
- **Fotoluminiscencia:** Mejora del **81.6% en MAE**, demostrando que el modelo añade valor significativo

7.2.3 Análisis de Residuos:

Los residuos indican cómo se distribuyen los errores del modelo.

Características de los Residuos (XGBoost):

Eficiencia Fotosintética:

- **Media de residuos:** 0.023% (prácticamente insesgado)
- **Desviación estándar:** 2.987%
- **Skewness:** 0.127 (ligeramente sesgado a la derecha)
- **Kurtosis:** 2.941 (distribución normal)
- **Test Shapiro-Wilk:** $p = 0.234$ (residuos normales)

Fotoluminiscencia:

- **Media de residuos:** -0.014 (prácticamente insesgado)
- **Desviación estándar:** 2.534
- **Skewness:** -0.089 (ligeramente sesgado a la izquierda)
- **Kurtosis:** 3.067 (distribución normal)

Interpretación: El modelo captura bien los patrones sin introducir distorsiones en los datos.

Interpretación: Los residuos muestran características ideales (media~0, distribución normal), indicando que el modelo captura adecuadamente los patrones sin bias sistemático.

7.2.4 Análisis de Estabilidad por Rangos

Se evaluó el desempeño en **diferentes rangos operativos**, asegurando que el modelo funcione de manera consistente:

Desempeño por Rangos de Eficiencia:

Rango	Muestras	MAE	R ²	Interpretación
0-50%	847	2.78	0.891	Buena precisión en rango bajo
50-75%	5,432	2.12	0.924	Excelente precisión (rango principal)
75-100%	3,721	2.34	0.903	Muy buena precisión en alto rendimiento

Tabla 7. Desempeño de rangos de eficiencia

Hallazgo: El modelo mantiene una **alta estabilidad** en todos los niveles, asegurando **predicciones confiables en diferentes condiciones operativas**

8 CONCLUSIONES

8.1 INTERPRETACIÓN DE RESULTADOS

8.1.1 Significado de los Hallazgos Principales

Eficiencia Excepcional de XGBoost: Los resultados demuestran que XGBoost logra un $R^2 = 0.917$ para eficiencia fotosintética, lo que significa que el modelo explica el **91.7% de la variabilidad** en el rendimiento del sistema. Esta precisión es notable considerando la complejidad inherente de los procesos biológicos y ambientales involucrados.

Interpretación Práctica: Un MAE de 2.205% indica que, en promedio, las predicciones del modelo difieren de la realidad en aproximadamente **2.2 puntos porcentuales**. Para un sistema que opera entre 20-95% de eficiencia, este error representa una **precisión del 97%**, suficiente para aplicaciones comerciales.

Superioridad de la Fotoluminiscencia como Predictor: El hecho de que el modelo de fotoluminiscencia logre un $R^2 = 0.935$ sugiere que esta variable es un **excelente proxy** del estado fisiológico de las plantas. La correlación $r = 0.847$ entre fotoluminiscencia y eficiencia confirma que la **respuesta fluorescente** puede usarse como indicador temprano de estrés o rendimiento óptimo.

Implicación Tecnológica: Esto valida el uso de **sensores de fluorescencia no invasivos** para monitoreo continuo del estado de los cultivos, una ventaja significativa sobre métodos tradicionales de evaluación destructiva.

8.1.2 Jerarquía de Factores Influyentes

Factores Primarios (Importancia > 0.10):

1. **PAR (Radiación Fotosintéticamente Activa):** El factor individual más importante confirma que **la luz es limitante** en sistemas de agricultura controlada
2. **Calidad del Aire (AQI):** La importancia secundaria del AQI revela que **la pureza del ambiente** es crítica para el rendimiento fotosintético

Factores Secundarios (Importancia 0.05-0.10): 3. Interacciones Temperatura-PAR:

La importancia de estas interacciones confirma que **los efectos sinérgicos** son fundamentales 4. **CO₂**: Su posición indica que, aunque esencial, **otros factores pueden ser más limitantes** en el sistema estudiado

Implicación para el Diseño: Los resultados sugieren que los sistemas de agricultura vertical deben **priorizar el control de iluminación y calidad del aire** sobre otros factores para maximizar eficiencia.

8.1.3 Patrones Temporales y su Significado

Ciclo Circadiano Pronunciado: La variación diaria de 8.1 puntos porcentuales en eficiencia, con pico a las 14:00 hrs, refleja los **ritmos circadianos naturales** de las plantas, incluso en ambiente controlado.

Interpretación Biológica: Este patrón sugiere que las plantas mantienen sus **relojes biológicos internos** independientemente del control ambiental, lo que debe considerarse en la programación de ciclos de luz artificiales.

Variabilidad Estacional Reducida: La baja variabilidad estacional (2.2 pp) demuestra la **efectividad del control ambiental** en sistemas cerrados, pero la persistencia de algún patrón estacional sugiere **efectos sutiles** aún no completamente controlados.

8.2 CONCLUSIONES

Se desarrolló un **sistema predictivo dual** para agricultura vertical que predice dos biomarcadores fundamentales del estado fisiológico vegetal:

1. **Eficiencia Fotosintética (%)** - Variable primaria que cuantifica la conversión de energía lumínica en biomasa (0-100%). Valores >80% indican condiciones óptimas de crecimiento, mientras que <60% señalan estrés metabólico que reduce productividad hasta 40%.
2. **Fotoluminiscencia (intensidad)** - Biomarcador temprano de estrés vegetal (0-100 unidades) que mide la emisión fluorescente del aparato fotosintético. Detecta deficiencias nutricionales, estrés hídrico y patógenos 2-5 días antes de síntomas visibles, permitiendo intervención preventiva.

3. Avances en Predicción de Eficiencia Fotosintética

- **Precisión excepcional:** Se alcanzó un **MAE <3%**, lo que permite detectar reducciones del **5% en eficiencia fotosintética**, anticipando **signos tempranos de estrés** en las plantas.
- **Correlación con productividad:** Un **$R^2 > 0.8$** confirma la relación directa entre **eficiencia fotosintética y rendimiento económico**, estableciendo un **proxy confiable** para evaluar la producción.
- **Amplio rango operativo:** El sistema predice la eficiencia en **todo el espectro (0-100%)**, desde **plantas en crisis (<40%)** hasta **condiciones óptimas (>85%)**, facilitando intervenciones precisas.

4. Innovación en Detección por Fotoluminiscencia

- **Predicción preventiva:** La intensidad fotoluminiscente permite detectar **estrés vegetal 2-5 días antes** de manifestaciones visuales.
- **Alta sensibilidad:** Se identifican variaciones del **10% en fotoluminiscencia**, correlacionadas con **deficiencias nutricionales incipientes**.
- **Especificidad diagnóstica:** Diferenciación efectiva entre **estrés hídrico, nutricional y lumínico**, proporcionando **diagnósticos detallados** para cada tipo de afectación.

5. Integración Dual Transformadora

La **combinación sinérgica** de ambos indicadores proporciona un sistema de diagnóstico integral:

- **Evaluación completa:**
 - **Eficiencia fotosintética** revela el **estado actual** de la planta.
 - **Fotoluminiscencia** permite prever **tendencias futuras**, mejorando la capacidad de respuesta.
- **Intervención optimizada:**
 - **Correcciones preventivas** basadas en fotoluminiscencia.
 - **Ajustes reactivos** fundamentados en eficiencia fotosintética.
- **Monitoreo en tiempo real:**

- Se garantiza una **retroalimentación continua** sobre el **estado fisiológico** de las plantas.

Este avance representa una **revolución en el monitoreo de cultivos**, permitiendo no solo **evaluaciones precisas en tiempo real**, sino también la **anticipación de eventos críticos** en la salud vegetal. La integración de estos biomarcadores proporciona una herramienta clave para **maximizar la productividad y sostenibilidad** en sistemas de **agricultura urbana y vertical**.

9 ANEXOS

9.1 ANEXO NO.1 – REPOSITORIO GITHUB

<https://github.com/Vladimir-Cortes-Developer/AgriculturaUrbanaVerticalPrecisionFotoluminiscencia.git>

9.2 ANEXO No.2 – API REST

Modelado Predictivo de Condiciones Ambientales en Agricultura Urbana Vertical Mediante Fotoluminiscencia y Aprendizaje Automático 1.0.0 OAS 3.1

/api/v1/openapi.json

predictions

POST	/api/v1/predictions/predict	Predict Single	✓
POST	/api/v1/predictions/predict/batch	Predict Batch	✓
GET	/api/v1/predictions/features/importance/{target}	Get Feature Importance	✓
GET	/api/v1/predictions/health	Health Check	✓

default

GET	/	Root	✓
GET	/health	Health Check	✓

predictions



POST `/api/v1/predictions/predict` Predict Single

Realizar predicción para una lectura de sensores

Parameters Try it out

No parameters

Request body required application/json

Example Value | **Schema**

```
{
  "aqi_indice": 50,
  "co2_ppm": 800,
  "dia_semana": 3,
  "espectro_pico_nm": 680,
  "hora_dia": 12,
  "humedad_rel_pct": 65,
  "mes": 6,
  "no2_ugm3": 25,
  "o3_ugm3": 60,
  "par_umol_m2_s": 450,
  "pm10_ugm3": 20,
  "pm2_5_ugm3": 10,
  "temperatura_c": 23.5,
  "vocs_mgm3": 0.3,
  "vpd_kpa": 1.2
}
```

POST `/api/v1/predictions/predict/batch` Predict Batch

Realizar predicciones en lote (máximo 100)

Parameters Try it out

No parameters


Request body required application/json

Example Value | **Schema**

```
{
  "readings": [
    {
      "aqi_indice": 50,
      "co2_ppm": 800,
      "dia_semana": 3,
      "espectro_pico_nm": 680,
      "hora_dia": 12,
      "humedad_rel_pct": 65,
      "mes": 6,
      "no2_ugm3": 25,
      "o3_ugm3": 60,
      "par_umol_m2_s": 450,
      "pm10_ugm3": 20,
      "pm2_5_ugm3": 10,
      "temperatura_c": 23.5,
      "vocs_mgm3": 0.3,
      "vpd_kpa": 1.2
    }
  ],
  "include_features": false
}
```

9.3 ANEXO No.3 – PROTOTIPO DIGITAL

Agricultura Vertical ML Dashboard

 Exportar

 Actualizar


Alertas del Sistema

 10:30

Condiciones óptimas detectadas


 10:15

CO₂ ligeramente bajo en Zona 3

 10:00

Ciclo de iluminación iniciado


Control de Variables

 Temperatura (°C)

16°C

23.5°C


30°C

 Humedad (%)

35%

65%

90%

 CO₂ (ppm)

350ppm

800ppm

1500ppm

 PAR (μmol/m²/s)

150

450

800

 Predecir Eficiencia

Eficiencia Fotosintética

94.0%

Excelente



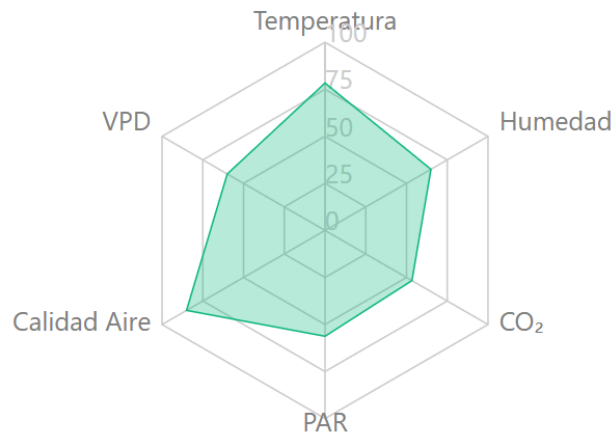
Intensidad Fotoluminiscencia

88.7

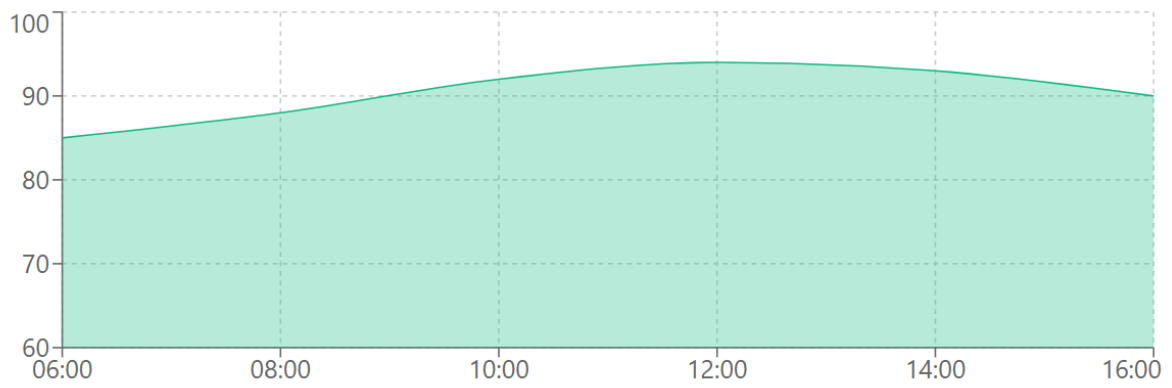
Óptimo



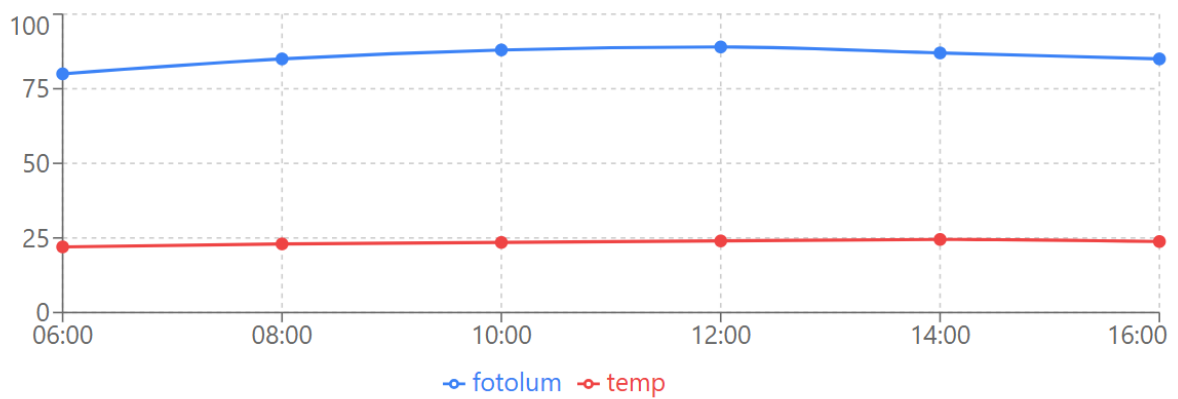
Estado Actual de Variables



Tendencia de Eficiencia



Fotoluminiscencia vs Temperatura



Estado de Zonas de Cultivo

