

Maestría en

INTELIGENCIA ARTIFICIAL APLICADA

**Trabajo previo a la obtención de título de
Magister en Inteligencia Artificial Aplicada**

AUTOR/ES:

Carpio Salvatierra Harry Alberto

Padilla Ramírez Sergio Ignacio

TUTOR/ES:

Karla Estefanía Mora Cajas

Fernanda Paulina Vizcaíno Imacaña

**TEMA: Aplicación De Modelos De Machine Learning Para
La Predicción Del Rendimiento De Rosas Utilizando
Variables Climáticas Y Agronómicas Históricas En Una
Empresa Florícola Del Cantón Mejía, Ecuador.**

Certificación de autoría

Nosotros, **Harry Alberto Carpio Salvatierra y Sergio Ignacio Padilla Ramírez**, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

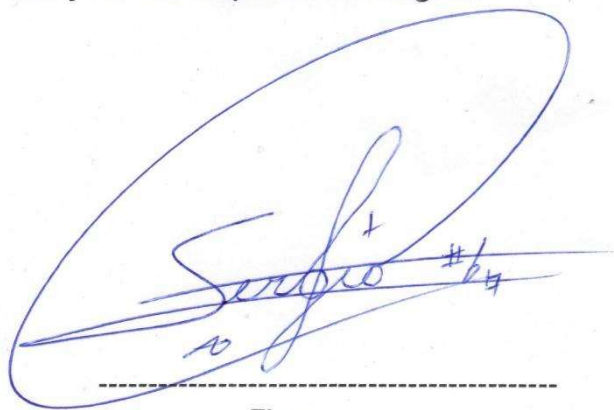
Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.



Firma

Harry Alberto Carpio Salvatierra

CI: 0920106796



Firma

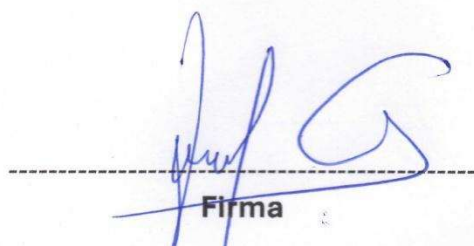
Sergio Ignacio Padilla Ramírez

CI: 1751265065

Autorización de Derechos de Propiedad Intelectual

Nosotros, **Harry Alberto Carpio Salvatierra y Sergio Ignacio Padilla Ramírez**, en calidad de autores del trabajo de investigación titulado “**APLICACIÓN DE MODELOS DE MACHINE LEARNING PARA LA PREDICCIÓN DEL RENDIMIENTO DE ROSAS UTILIZANDO VARIABLES CLIMÁTICAS Y AGRONÓMICAS HISTÓRICAS EN UNA EMPRESA FLORÍCOLA DEL CANTÓN MEJÍA, ECUADOR.**”, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, agosto 2026



Firma

Harry Alberto Carpio Salvatierra

CI: 0920106796



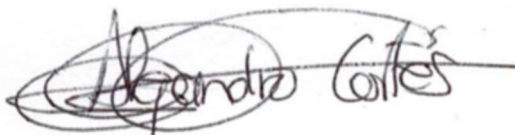
Firma

Sergio Ignacio Padilla Ramírez

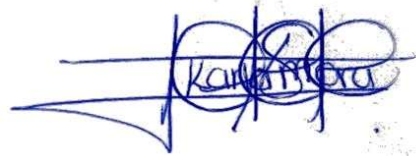
CI: 1751265065

Aprobación de dirección y coordinación del programa

Nosotros, **Alejandro Cortés López** Director EIG y **Karla Estefanía Mora Cajas** Coordinadora UIDE, declaramos que: **Carpio Salvatierra Harry Alberto, Padilla Ramírez Sergio Ignacio** son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés López Director de la
Maestría en Inteligencia Artificial
Aplicada



Karla Estefanía Mora Cajas
Coordinadora de la
Maestría en Inteligencia Artificial Aplicada

DEDICATORIA

A la memoria de mi padre y al ejemplo
de vida y fortaleza de mi madre.

A María Isabel y a nuestra Sol,
luz y calidez de mi vida.

Harry

A Dios mi Padre quien es mi fortaleza.
A mi Madre por su apoyo incondicional
y su ejemplo esfuerzo.

Sergio

AGRADECIMIENTOS

Al apoyo y amor de mi familia, padres, hermanos, esposa e hija. Su compañía siempre ha sido fuente de ánimos, motivación y fortaleza.

A la Universidad Internacional del Ecuador y la Florícola objetivo de este estudio, por propiciar y facilitar el desarrollo de este trabajo.

A Sergio, por la generosidad irrestricta de compartir sus conocimientos, sin los cuales este trabajo no habría sido posible.

- Harry

A mi Madre quien en todo este proceso animó y me a través de su ejemplo me ha enseñado el valor del esfuerzo y la constancia en cada aspecto de mi vida.

A la Universidad Internacional del Ecuador y Los Ingenieros de la Florícola: Esteban Chiriboga, Santiago Serrano, Wilfrido Cazar y en especial a Juan Carlos Coello que fue me mentor personal en todo este proceso sin sus aportes este proyecto no habría sido plausible.

A Harry, por su paciencia para escuchar mis ideas y con su experiencia ayudarme poder plasmar una de mis metas profesionales en la industria.

- Sergio

RESUMEN

En floricultura, la programación de la producción suele apoyarse en la experiencia del equipo técnico y en la interpretación de curvas históricas, sin cuantificar las interacciones entre los factores que condicionan el desarrollo del cultivo. En la unidad productiva estudiada, una empresa florícola ubicada en el cantón Mejía, Ecuador, la proyección semanal del equipo técnico presenta un error absoluto medio de 743,79 tallos y una tasa de acierto del 41,87 % dentro del intervalo operativo de tolerancia del 85 % al 115 %. Estos indicadores se calcularon sobre 27 167 combinaciones de bloque, variedad y semana generadas entre noviembre de 2023 y agosto de 2026. La investigación construye y valida un modelo de aprendizaje automático orientado a pronosticar, con un horizonte de una semana ISO, la producción semanal de tallos de rosa para cada bloque y variedad. Para ello, se integraron cuatro fuentes operativas: cosecha diaria, poda de formación, estación meteorológica externa y registros del microclima de invernadero. La base consolidada reunió 179 804 registros diarios correspondientes a 64 bloques y 139 variedades, a partir de los cuales se conformó un panel semanal de 26 170 observaciones. Mediante correlación cruzada se estimó empíricamente el tiempo de maduración entre el pinch y la cosecha, cuya mediana fue de 98 días. Se analizaron ocho variantes de modelado, estructuradas como una ablación controlada sobre cuatro ejes de diseño, mediante una validación temporal walk-forward con ventana expansiva y embargo; la selección se realizó con base en el error absoluto medio. La mejor variante, una red LSTM con corrección residual por variedad y rezago biológico explícito, redujo el error absoluto medio a 474,0 tallos y elevó la tasa de acierto al 66,08 %: un 36,3 % menos de error y 24,21 puntos porcentuales más de acierto, aunque sobre un horizonte efectivo distinto del que emplea el planificador. La explicabilidad atribuyó entre el 55,5 % y el 80,9 % de la capacidad predictiva a la señal autorregresiva y menos del 8 % al clima, predictor marginal en el rango de variación observado.

Palabras clave: predicción de producción, aprendizaje automático, floricultura, rosa de invernadero, validación temporal walk-forward, redes LSTM.

ABSTRACT

Floricultural production planning commonly relies on the technical team's experience and the interpretation of historical yield curves, without quantitatively representing the interactions among the factors that shape crop development. At the production unit examined, a flower-growing company located in Mejía canton, Ecuador, the weekly forecast prepared by the technical team has a mean absolute error of 743.79 stems and a hit rate of 41.87% within the operational tolerance range of 85% to 115%. These indicators were calculated from 27,167 block, variety, and week combinations generated between November 2023 and August 2026. This study builds and validates a machine learning model for forecasting weekly rose stem production by block and variety one ISO week ahead. Four operational sources were integrated: daily harvest records, formative pruning, an external weather station, and greenhouse microclimate measurements. The consolidated dataset comprised 179,804 daily records from 64 blocks and 139 varieties, from which a weekly panel of 26,170 observations was constructed. Cross-correlation analysis was used to estimate the maturation period between pinching and harvest empirically, yielding a median of 98 days. Eight modeling variants, structured as a controlled ablation study across four design dimensions, were assessed through expanding-window walk-forward temporal validation with an embargo and selected according to mean absolute error. The best variant, an LSTM network with per-variety residual correction and an explicit biological lag, lowered mean absolute error to 474.0 stems and raised the hit rate to 66.08 % — a 36.3 % error reduction and a 24.21 percentage-point gain — though on an effective horizon that differs from the planner's. Explainability attributed between 55.5 % and 80.9 % of predictive capacity to the autoregressive signal and less than 8 % to climate, a marginal predictor within the observed range of variation.

Keywords: production forecasting, machine learning, floriculture, greenhouse rose, walk-forward temporal validation, LSTM networks.

TABLA DE CONTENIDOS

CAPITULO 1	1
1. INTRODUCCIÓN.....	1
1.1. Definición del Proyecto	1
1.2. Justificación e Importancia del Trabajo de Investigación	5
1.3. Alcance	7
1.4. Objetivos	8
CAPITULO 2	10
2. REVISIÓN DE LITERATURA	10
2.1. Estado del Arte	10
2.2. Marco Teórico.....	15
2.3. Terminología.....	23
CAPITULO 3	35
3. DESARROLLO	35
3.1. Análisis y Limpieza de Datos.....	35
3.2. Entrenamiento de Modelos	63
CAPITULO 4	86
4. ANÁLISIS DE RESULTADOS.....	86
4.1. Ventana Amplia por Variante	86
4.2. Ventana Común	88
4.3. Predicho vs. Real	92
4.4. Explicabilidad e Importancia de Variables.....	97

4.5.	Respuestas a las Preguntas de Diseño	103
4.6.	Cuantificación de incertidumbre	110
CAPITULO 5		113
5.	CONCLUSIONES Y RECOMENDACIONES.....	113
5.1.	Conclusiones	113
5.2.	Recomendaciones	117
Referencias.....		122

LISTA DE TABLAS

Tabla 1 Descripción de las fuentes de datos de entrada.....	35
Tabla 2 Resumen de datos de fuente de datos Davis.....	39
Tabla 3.....	42
Tabla 4 Tipos de columnas relevantes de HInputs	43
Tabla 5 Tramos con lecturas largas de temperatura por mes (sensores congelados).....	51
Tabla 6 Resumen de las fuentes de entrada y su rol en la base unificada	59
Tabla 7 Nulos por columna, antes vs. después del relleno	61
Tabla 8 Chequeo rápido de calidad de datos en de base unificada	62
Tabla 9 Ejes de diseño	64
Tabla 10 Definición de variantes	65
Tabla 11 Combinaciones de cada ventana candidata para SEQUENCE_WINDOW	81
Tabla 12 Resultado de la búsqueda de SEQUENCE_WINDOW.....	82
Tabla 13 Métricas relevantes de las variantes secuenciales	84
Tabla 14 Peso por grupo de variables (%), normalizado por cada modelo	98
Tabla 15 Resumen resultados de preguntas de diseño.....	103

LISTA DE FIGURAS

Figura 1 Cantidad de registros en cada archivo ISSI.....	44
Figura 2 Frecuencia de registro de eventos en los archivos ISSI	45
Figura 3 Distribución de valores nulos por columna.....	45
Figura 4 Diagrama de caja de los valores de temperatura en septiembre del 2024.....	46
Figura 5 Diagrama de caja de los valores de humedad en septiembre del 2024.....	46
Figura 6 Histograma de temperatura del Bloque 21 en septiembre del 2025.....	47
Figura 7 Diagrama de cajas de todos los valores de temperatura	48
Figura 8 Diagrama de caja de todos los valores de humedad.....	49
Figura 9 Diagrama de caja de todos los valores del sensor de luzA.....	50
Figura 10 Datos crudos temperatura en diciembre del 2025 de los bloques 1, 4 y 10.....	51
Figura 11 Temperatura de todos los bloques en diciembre del 2025	52
Figura 12 Humedad de todos los bloques en diciembre del 2025.....	52
Figura 13 Valores “congelados” de temperatura removidos. Diciembre del 2025.	53
Figura 15 Tabla salida de extraer_header_hinputs.py	54
Figura 14 Pipeline de procesamiento de archivos de ISSI	55
Figura 16 Tabla de salida de extraer_bloque_metrica.py	56
Figura 17 Tabla de salida de procesar_hinputs.py con datos intra-invernadero	56
Figura 18 Tabla de salida de procesar_hinputs.py con datos extra-invernadero.....	57
Figura 19 Tabla agregada con valores diarios válidos, salida de agregar_promedio_diario.py	57
Figura 20 Tabla de salida del script pivotear_promedio_diario.py	58
Figura 21 Top 20 de bloques y variedades con más filas	62
Figura 22 Cantidad de registros por días en la base unificada.....	63
Figura 23 Cantidad total de tallos cosechados por día en la base unificada	63

Figura 24 Cobertura temporal de las dos fuentes de datos de entrada	67
Figura 25 Cobertura semanal de datos de proyección humana	68
Figura 26 Distribución de target y serie temporal de tallos cosechados	68
Figura 27 Disponibilidad de datos climáticos por invernadero	69
Figura 28 Curvas CCF de las 6 variedades con la correlación más fuerte	70
Figura 29 Distribución del lag de pico de todas las variedades	71
Figura 30 Distribución del ancho de las ventanas biológicas de las variedades con correlación fuerte	72
Figura 31 Panel semanal del bloque 20 y la variedad K0IK	73
Figura 32 Gráfico de walk-forward con embargo	74
Figura 33 Representación de fold 0	75
Figura 34 Ventana expansiva de folds	75
Figura 35 Búsqueda de hiperparámetros por validación cruzada temporal de HistGBT	77
Figura 36 MAE de entrenamiento de HistGBT global	78
Figura 37 Búsqueda de hiperparámetros por validación cruzada temporal de CatBoost	78
Figura 38 RMSE de entrenamiento de CatBoost global	79
Figura 39 MAE por fold (train vs. test), variante LSTM_GLOBAL	83
Figura 40 MAE por fold (train vs. test), variante TCN_GLOBAL	83
Figura 41 MAE por fold (train vs. test), variante LSTM_RESID_BIO	84
Figura 42 MAE de las variantes versus el del baseline, usando sus propias ventanas	86
Figura 43 R^2 de las variantes versus el del baseline, usando sus propias ventanas	87
Figura 44 Tasa de acierto de las variantes versus el del baseline, usando sus propias ventanas	88
Figura 45 MAE de las variantes versus el del baseline, usando la ventana común	89
Figura 46 R^2 de las variantes versus el del baseline, usando la ventana común	89
Figura 47 Tasa de acierto de las variantes versus el del baseline, usando la ventana común	90

Figura 48 Comparativa de MAE de cada variante, ventana amplia vs común	91
Figura 49 Comparativa de R2 de cada variante, ventana amplia vs común.....	91
Figura 50 Tasa de acierto de cada variante, ventana amplia vs común	92
Figura 51 Gráficos de dispersión de variantes HistGBT.....	93
Figura 52 Gráficos de dispersión de variantes secuenciales	93
Figura 53 Frecuencia en que cada variante le gana variable por variable al baseline	94
Figura 54 Mapa de calor de R^2 de la predicción semanal por variedad y variante.....	96
Figura 55 Peso de los grupos de variables por cada variante	99
Figura 56 Importancia individual de variables de HistGBT y CatBoost globales	102
Figura 57 Importancia individual de variables de LSTM_GLOBAL.....	102
Figura 58 Importancia individual de variables de LSTM_RESID_BIO	103
Figura 59 Comparativa de variantes de la pregunta 1	104
Figura 60 Comparativa de las variantes HistGBT de la pregunta 2	105
Figura 61 Comparativa de las variantes CatBoost de la pregunta 2.....	106
Figura 62 Comparativa de las variantes de la pregunta 3	107
Figura 63 Comparativa de las variantes de la pregunta 4	108
Figura 64 Comparativa de las variantes de la pregunta 5	109
Figura 65 Comparativa de las variantes de la pregunta 6	110
Figura 66 Nivel de confianza de las variantes estudiadas	111
Figura 67 Predicción, su banda de confianza del 90% y valor de cosecha real.....	112
Figura 68 Comparativa de reducción de MAE por variante y tipo de ventana	113
Figura 69 Ranking final de las variantes y el baseline	114
Figura 70 Comparación de métricas entre las dos mejores variantes de cada familia de modelos ...	116
Figura 71 Cobertura empírica del intervalo conformal de las dos mejores variantes de cada familia de modelos	117

CAPITULO 1

1. INTRODUCCIÓN

1.1. Definición del Proyecto

La producción de rosas bajo invernadero depende de un conjunto de factores climáticos y agronómicos que determinan tanto el volumen como el momento en que cada tallo alcanza el punto de corte comercial. La planificación de esa producción se fundamenta, en la unidad productiva analizada, en un enfoque empírico-tradicional que se apoya en la consolidación manual de registros en hojas de cálculo, en la lectura de curvas de producción históricas y en el criterio del personal técnico, construido sobre la observación directa del cultivo. El problema radica en que ese procedimiento no modela de forma cuantitativa las interacciones entre las variables que gobiernan el desarrollo del cultivo como la temperatura, la humedad relativa, la radiación acumulada y las labores agronómicas; de modo que la proyección resultante recoge tendencias generales, pero no la magnitud del efecto de cada factor sobre cada combinación de bloque y variedad

El seguimiento de esa proyección se realiza mediante el porcentaje de cumplimiento, definido como el cociente entre la producción obtenida y la proyección emitida:

$$\% \text{ Cumplimiento} = \frac{\sum \text{Producción}}{\sum \text{Proyección}} \times 100$$

Constituye el indicador oficial de gestión; dicho indicador se calcula sobre totales, de manera que las sobreestimaciones y las subestimaciones de las distintas combinaciones bloque-variedad se compensan entre sí al sumar. La compensación produce un ajuste global que no refleja la magnitud del error a nivel del bloque-variedad de cada semana proyectada de modo que el total agregado oculta una dispersión que el nivel de decisión sí experimenta, por este motivo la empresa establece como zona de tolerancia operativa del cumplimiento entre el 85 % y el 115 % a nivel de variedad-bloque.

El presente estudio cuantifica esa diferencia aplicando el mismo criterio de tolerancia, sobre todas las combinaciones de proyecciones disponibles, se definió para ello la tasa de acierto como la proporción de combinaciones de bloque, variedad y semana ISO cuyo cociente entre producción real y valor pronosticado se sitúa dentro de la banda operativa del 85 % al 115 % que emplea la unidad productiva. Sobre las 27 167 proyecciones registradas entre noviembre de 2023 y agosto de 2026, la proyección del equipo técnico alcanza una tasa de acierto del **41,87 %**. Su error absoluto medio asciende a **743,79 tallos** por combinación y semana, con una raíz del error cuadrático medio de **1 455,20 tallos** y un error porcentual absoluto medio del **27,13 %**, es importante mencionar que la dispersión tampoco se corrige ampliando la tolerancia dado que admitiendo desviaciones de hasta un ± 20 % respecto de la producción real, la proporción de combinaciones acertadas alcanza apenas el **53,50 %**, de modo que cerca de la mitad de las proyecciones se desvía más de un quinto de lo efectivamente cosechado.

El contraste entre ambos niveles de medición constituye el hallazgo que estructura el planteamiento: el indicador agregado declara la operación dentro de tolerancia, mientras que a nivel desagregado más de la mitad de las decisiones individuales no cumplen, esa dispersión que indicador oficial no capturada es determinante para el compromiso de pedidos, la asignación de mano de obra y el dimensionamiento de la capacidad de postcosecha y constituye el margen de mejora que este trabajo se propone acotar. El desempeño de la proyección elaborada por el equipo técnico se adopta, en consecuencia, como línea base de contraste.

La complejidad del problema radica en la interacción de múltiples variables que influyen en el rendimiento del cultivo, entre estas se incluyen variables climáticas internas del invernadero, como temperatura, humedad relativa, déficit de presión de vapor (VPD) y grados día de crecimiento (GDD), así como variables externas como radiación, temperatura y humedad ambiental. Adicionalmente, se consideran variables agronómicas como la variedad cultivada, los *pinch* realizados y los registros históricos de cosecha. Si bien existe conocimiento técnico que permite identificar tendencias

generales, como el impacto de semanas frías o cálidas en el desarrollo del cultivo, la relación entre estas variables y la producción no ha podido ser modelada de manera precisa mediante enfoques.

El estudio se desarrolló sobre datos históricos comprendidos entre el 28 de agosto de 2023 y el 5 de junio de 2026, equivalentes a 145 semanas ISO continuas. La base analítica consolidada contiene 179 317 registros diarios a nivel de bloque y variedad, de los cuales 179 061 poseen identificador de bloque y variedad y constituyen el insumo efectivo del modelado, distribuidos sobre 64 bloques productivos y 141 variedades; esta base integra cuatro fuentes operativas independientes: el registro diario de cosecha; el registro semanal de poda de formación (pinchs), con 15.689 observaciones entre la semana 35 de 2023 y la semana 18 de 2026, que consigna año, semana, bloque, variedad y tallos pinchados; los datos intradía de la estación meteorológica externa Davis, con cadencia de quince minutos y 94 443 observaciones de temperatura, humedad, radiación solar y evapotranspiración que cubren la totalidad del período; y los datos de microclima de invernadero por bloque, con registros cada dos minutos procedentes del sistema de sensores instalado en septiembre de 2024 y disponibles únicamente a partir de esa fecha. Las fuentes descritas fueron integradas y se agregaron a la granularidad de semana ISO por bloque y variedad constituyendo así la unidad de análisis del proyecto; una vez consolidada la base se aplicaron criterios de elegibilidad de variedades y de cobertura mínima de la ventana de agregación que dio como resultado el panel modelable que comprende 26 170 observaciones sobre 61 bloques y 118 variedades, entre el 20 de noviembre de 2023 y el 25 de mayo de 2026. El archivo consolidado que alimenta el modelado ocupó aproximadamente 2,1 MB en formato Parquet comprimido.

La concordancia entre las dos fuentes de producción disponibles respalda el uso del archivo de proyecciones como referencia: la suma semanal de tallos cosechados calculada de forma independiente por este estudio coincide con la producción registrada por la empresa en el **99,1 %** de las 24.965 combinaciones en común.

La variable objetivo se definió operativamente como el número de tallos cosechados por bloque y variedad, agregado a resolución semanal, dicha variable objetivo es la magnitud con la que la empresa construye sus proyecciones operativas y calcula su indicador oficial de cumplimiento, lo que asegura coherencia entre la predicción del modelo y los procesos reales de planificación siendo formalmente $y(b, v, s)$ la producción del bloque b y la variedad v en la semana ISO:

$$y(b, v, s) = \sum_{d \in s} \text{Tallos Cosechados}(b, v, d)$$

Donde d recorre los días pertenecientes a la semana s . El horizonte de predicción implementado es de una semana ISO hacia adelante ($s+1$), coincidente con el ciclo semanal de emisión de la proyección interna: el equipo de producción emite cada jueves su proyección para la semana siguiente. La elección permite un contraste directo contra la referencia humana bajo idéntico horizonte, que es la condición mínima para que la comparación resulte equitativa. La extensión a horizontes de dos a cuatro semanas, correspondientes al ciclo de planificación comercial, se identifica como línea de trabajo futuro; su implementación exigiría redefinir las variables autorregresivas, dado que el rezago de una semana deja de estar disponible a partir de $s+2$.

Los resultados obtenidos permitieron superar el desempeño de la proyección del equipo técnico con dos familias de modelos.

Entre las variantes de objetivo semanal nativo, es decir, aquellas que predicen directamente el total de la semana siguiente y resuelven por tanto la misma tarea que el planificador, la de mejor desempeño fue HistGBT global. Evaluada sobre las mismas semanas objetivo, las mismas combinaciones de bloque y variedad y el mismo valor real de producción que la referencia humana, alcanzó un error absoluto medio de 629,79 tallos frente a los 743,79 de la proyección del equipo técnico, lo que representa una reducción del error del 15,3 %, y una tasa de acierto del 52,55 % frente al 41,87 % de la referencia, esto es, 10,68 puntos porcentuales por encima.

Las variantes de aprendizaje profundo alcanzaron reducciones superiores. La mejor de ellas, *LSTM_RESID_BIO*, que combina una red recurrente global con una corrección residual por variedad y el predictor biológico explícito, registró un error absoluto medio de 474,01 tallos, una reducción del 36,3 % respecto de la proyección del equipo técnico, y una tasa de acierto del 66,08 %, esto es, 24,21 puntos porcentuales por encima de la referencia. Debe precisarse que estas variantes operan sobre un grano temporal distinto: cada observación predice la producción del día siguiente y las predicciones diarias se agregan por suma hasta obtener el total semanal, de modo que su horizonte efectivo es de uno a dos días. La comparación bajo horizonte idéntico al del planificador corresponde a las variantes de objetivo semanal nativo, y así se reporta en el capítulo de resultados.

El estudio aporta además tres hallazgos con valor propio. El primero es la estimación empírica de la ventana de maduración entre el pinch y la cosecha mediante análisis de correlación cruzada, con una mediana de 98 días y ventana específica identificable en 35 de las 118 variedades analizadas. El segundo es la constatación de que la capacidad predictiva de los modelos descansa entre un 55 % y un 81 % en la señal autorregresiva y en menos de un 8 % en las variables climáticas, lo que obliga a matizar el papel del clima como predictor marginal en el rango de variación observado. El tercero es la incorporación de intervalos de predicción conformales con cobertura empírica verificada, un elemento del que la proyección manual carece por completo.

1.2. Justificación e Importancia del Trabajo de Investigación

El sector florícola constituye uno de los pilares más importantes de la agroexportación en Ecuador, siendo uno de los principales generadores de divisas no petroleras y empleo en el país, de acuerdo con datos de Expoflores (Asociación de Exportadores de Flores de Ecuador, 2024) y el Banco Central del Ecuador (Banco Central del Ecuador, 2025), la floricultura ecuatoriana representa un porcentaje significativo de las exportaciones agrícolas, con un enfoque predominante en la producción de rosas, las cuales son reconocidas a nivel internacional por su calidad. El sector florícola genera miles de

empleos directos e indirectos, especialmente en zonas de la Sierra como el cantón Mejía, donde se concentra una alta densidad de empresas florícolas.

La precisión en la planificación de la producción no solo tiene implicaciones operativas internas, sino que impacta directamente en la capacidad de cumplimiento de compromisos comerciales internacionales, la estabilidad de ingresos y la competitividad del sector. Los problemas como el incumplimiento de pedidos, la sobreproducción o la desincronización de picos productivos en fechas clave representan riesgos económicos importantes para las empresas.

A nivel técnico, la predicción de la producción agrícola ha sido ampliamente estudiada en distintos cultivos mediante el uso de técnicas de *Machine Learning*. Estudios con modelos avanzados como XGBoost, CatBoost y redes neuronales han demostrado superar con amplio margen a los enfoques estadísticos tradicionales y a las estimaciones manuales empíricas, al capturar adecuadamente las relaciones no lineales entre el microclima del invernadero y la tasa de floración, (Van Ploeg & Heuvelink, 2005) demuestran que la integración de datos climáticos, agronómicos y modelos de aprendizaje automático permite mejorar la capacidad predictiva en sistemas agrícolas complejos. Asimismo, revisiones sistemáticas como evidencian que modelos como Random Forest, redes neuronales y métodos de ensamble son comúnmente utilizados para la predicción de rendimiento en diferentes cultivos.

Sin embargo, a pesar de estos avances, existe un vacío importante en la aplicación de estos enfoques en cultivos florícolas, particularmente en la predicción de la producción de rosas en entornos de invernadero. Los estudios revisados se enfocan principalmente en cultivos extensivos como trigo, arroz o maíz, y en menor medida en hortalizas, como se observa en investigaciones como, donde se aborda la predicción de rendimiento en tomate bajo condiciones controladas; no obstante, no se identifican investigaciones de alto nivel que aborden específicamente la predicción de producción en rosas utilizando técnicas avanzadas de *Machine Learning*.

La ausencia de investigaciones de *Machine Learning* en la predicción de la producción de rosas representa una oportunidad tanto académica como aplicada; desde el punto de vista académico, el desarrollo de modelos predictivos en este contexto contribuye a la generación de conocimiento en un área poco explorada o por otro lado desde el punto de vista empresarial, la implementación de estas herramientas permitiría mejorar la precisión de las proyecciones, optimizar la planificación productiva y reducir los riesgos operativos asociados a la incertidumbre en la producción.

1.3. Alcance

El presente proyecto se enfoca en el desarrollo de un modelo predictivo para la estimación del rendimiento de rosas a nivel de bloque y variedad, utilizando datos históricos climáticos y agronómicos de una empresa florícola ubicada en el cantón Mejía, Ecuador. El estudio incluirá el procesamiento, depuración y consolidación de datos históricos, así como la evaluación comparativa de diferentes modelos de predicción, tanto de *Machine Learning* como de estadística clásica, con el fin de identificar la mejor alternativa en términos de precisión.

El diseño del flujo de trabajo y visualización de proyecciones corresponde estrictamente a un prototipo funcional o Prueba de Concepto (PoC). Se aclara de forma taxativa que una futura puesta en producción real requeriría fases posteriores de validación de software, pruebas A/B de impacto en el negocio y la implementación de ciclos de *MLOps* (Monitoreo y Despliegue Continuo), las cuales quedan fuera del límite y alcance de la presente tesis. El estudio se limita a validar la capacidad predictiva del modelo utilizando la información histórica disponible de la empresa objeto de estudio; no se contempla dentro del alcance del proyecto la implementación en tiempo real del modelo en sistemas productivos, ni el desarrollo de soluciones de automatización a nivel de infraestructura tecnológica. Asimismo, el estudio se limita a la información disponible de una sola empresa florícola, por lo que los resultados no necesariamente serán generalizables a todo el sector.

1.4. Objetivos

1.4.1. Objetivo general

Desarrollar y validar un modelo de aprendizaje automático para la predicción de la producción semanal de tallos de rosa a nivel de bloque y variedad, con un horizonte de una semana, a partir de variables climáticas, agronómicas y autorregresivas históricas, en una empresa florícola del cantón Mejía, Ecuador.

1.4.2. Hipótesis

El entrenamiento de modelos de aprendizaje automático sobre variables biológicas, climáticas y autorregresivas permitirá mejorar la predicción de la producción semanal de tallos cosechados a nivel de bloque y variedad, con un horizonte de una semana ISO, superando a la proyección elaborada por el equipo técnico de la unidad productiva tanto en la tasa de acierto del 41,87 %, como en el error absoluto medio, que en esa misma referencia asciende a 743,79 tallos por combinación de bloque, variedad y semana.

1.4.3. Objetivos Específicos

- Depurar y consolidar la información histórica de cosecha, pinch y clima a nivel de bloque y variedad, verificando su calidad y su consistencia mediante contraste cruzado entre fuentes.
- Estimar empíricamente la ventana de maduración entre el pinch y la cosecha mediante análisis de correlación cruzada, y derivar de ella los predictores biológicos y climáticos del modelo.

- Implementar y comparar ocho variantes de modelo, organizadas como una tabla de ablación controlada sobre cuatro ejes de diseño: familia de algoritmo, grano del objetivo, tipo de especialización y presencia de rezago biológico explícito.
- Evaluar el desempeño de las variantes mediante validación temporal de tipo walk-forward con ventana expansiva y embargo, empleando el error absoluto medio como métrica rectora.
- Contrastar el desempeño del modelo seleccionado frente a la proyección elaborada por el equipo técnico, sobre las mismas semanas objetivo, las mismas combinaciones de bloque y variedad y el mismo valor real de producción.
- Formular una propuesta de despliegue operativo del modelo, con su esquema de monitoreo y su política de reentrenamiento, orientada a su uso como herramienta de apoyo a la planificación productiva.

CAPITULO 2

2. REVISIÓN DE LITERATURA

2.1. Estado del Arte

2.1.1. Predicción de Rendimiento Agrícola con Machine Learning

El aprendizaje automático se ha incorporado de forma extendida a la predicción de rendimiento agrícola durante la última década. Las revisiones de (Van Klompenburg et al., 2020), (Paudel et al., 2021), (Benos et al., 2021) y (Shawon et al., 2025) muestran una expansión sostenida del Machine Learning en agricultura; entre las familias de los modelos probados y entrenados destacan las Random Forest, los métodos de Gradient Boosting y las redes neuronales. Las métricas reportadas con mayor frecuencia incluyen RMSE, R^2 y MAE; no obstante, estas revisiones también evidencian heterogeneidad en cultivos, escalas espaciales, horizontes y protocolos de validación; por ello la selección de un modelo que optimice la predicción de la producción no puede ser generalizada en función de resultados ajenos si no que debe ser validado en función de la realidad específica de la unidad productiva, del cultivo, la calidad de los datos, la escala y la formulación del objetivo.

(Van Klompenburg et al., 2020) revisaron 50 estudios de Machine Learning y un conjunto adicional de 30 investigaciones de Deep Learning y establecieron que la temperatura, la precipitación y el tipo de suelo como las variables más empleadas además identificaron a las redes neuronales artificiales como una familia recurrente en predicción de rendimiento y señalaron el uso creciente de CNN y LSTM.

(Shawon et al., 2025) de manera similar en una revisión de 97 trabajos publicados entre 2017 y 2024, encontraron un uso destacado de regresión lineal, Random Forest, Gradient Boosting, CNN y LSTM y reportaron la temperatura, la composición del suelo y los índices de vegetación

como variables predominantes. La raíz del error cuadrático medio, el coeficiente de determinación y el error absoluto medio fueron las métricas de evaluación más reportadas.

(Paudel et al., 2021) propusieron un flujo modular para predicción de rendimiento a gran escala que combina principios agronómicos, información meteorológica, teledetección, suelo y salidas de modelos de cultivo. Su énfasis en características interpretables, experimentos repetibles y prevención de fuga de información es especialmente relevante para este estudio. Aunque la escala regional y los cultivos analizados difieren del panel Bloque × Variedad de rosas, el antecedente demuestra que el aporte del Machine Learning depende tanto de la arquitectura como de la corrección temporal y la reproducibilidad del proceso.

Estas tres fuentes prueban que el aprendizaje automático es una herramienta útil para la predicción de rendimiento agrícola, pero comparten una limitación su evidencia proviene de cultivos extensivos a campo abierto, medida en unidades de masa por unidad de superficie, y en ventanas de agregación estacionales o anuales, ninguna de ellas informa sobre la predicción de un conteo semanal de tallos en un cultivo ornamental protegido

2.1.2. Clima Externo y Microclima Intra Invernadero

El monitoreo inteligente de invernaderos integra sensores, Internet de las Cosas y sistemas de apoyo a la decisión. (Abou-Mehdi-Hassani, 2025) revisaron 54 trabajos publicados entre 2019 y 2023 y describieron al IoT como una infraestructura esencial para registrar y controlar variables ambientales, la implementación de estos dispositivos genera la disponibilidad de series de temperatura, humedad, radiación y otras covariables dinámicas que permiten a su vez monitorear o predecir el ambiente sin embargo dichas predicciones no equivalen a predecir el rendimiento físico: el microclima es una entrada potencial del modelo de producción, no su variable objetivo.

Las investigaciones sobre la predicción del microclima en invernaderos han empleado tanto modelos físicos como distintos algoritmos de Machine Learning. En un invernadero comercial se compararon redes neuronales, máquinas de vectores de soporte, métodos de bagging y boosting, y dinámica de fluidos computacional para estimar la temperatura y la humedad relativa. Aunque los modelos alcanzaron una concordancia elevada, su desempeño varió según la estación y la composición del conjunto de entrenamiento. Otro estudio integró factores dinámicos, Random Forest y una red neuronal para pronosticar temperatura, humedad relativa, radiación fotosintéticamente activa y concentración de CO₂. Esta combinación superó, en el contexto evaluado, a configuraciones basadas en BPNN y LSTM, lo que evidencia la utilidad de seleccionar variables relevantes e integrar relaciones lineales y no lineales en predicciones ambientales de múltiples horizontes (Sun & Chang, 2023).

De manera complementaria, los pronósticos meteorológicos y los algoritmos de Machine Learning también se han utilizado para optimizar el clima de los invernaderos, particularmente mediante la predicción de la temperatura. En conjunto, estos antecedentes demuestran que el microclima puede modelarse con precisión operativa, pero su relación con la producción depende de la respuesta retardada de la planta, las prácticas de manejo y la estructura jerárquica del cultivo. En consecuencia, este conjunto de resultados respalda dos decisiones conceptuales: incluir variables climáticas internas y externas cuando estén disponibles y evaluar la estabilidad temporal del modelo; sin embargo, no permite concluir que un predictor de microclima estime correctamente el número de tallos cosechados ni que supere una proyección semanal de cosecha (Sedig, 2019).

2.1.3. Predicción Directa de la Producción Física Bajo Invernadero

(Gong et al., 2021) combinaron redes convolucionales temporales y redes recurrentes para predecir rendimiento de cultivos de invernadero. La arquitectura aprovechó la capacidad de

las TCN para extraer patrones temporales a distintas escalas y la memoria de las RNN. Los experimentos se realizaron con datos de producción de tomate expresados en masa por unidad de superficie. Este antecedente fundamenta el uso exploratorio de TCN y LSTM en la tesis, aunque la especie, la unidad de respuesta y el horizonte de las secuencias no son equivalentes al total semanal de tallos de rosa por Bloque × Variedad.

En otra investigación se compararon KNN, Random Forest, LightGBM y un ensamble apilado para estimar el rendimiento de tomate cultivado en invernadero. En el conjunto de datos analizado, el ensamble alcanzó el mayor coeficiente de determinación, con un R^2 de 0,896, y el menor error cuadrático medio, con un MSE de 0,008. Estos resultados evidencian que la combinación de modelos puede mejorar el desempeño de una tarea de regresión en condiciones de cultivo protegido. No obstante, las métricas reportadas no deben trasladarse directamente a esta tesis, puesto que difieren el cultivo, la escala de la variable objetivo y el diseño de evaluación (Mancer et al., 2024).

2.1.4. Producción de Rosas: Antecedentes de Inteligencia Artificial y de Modelado

La literatura sobre la aplicación de Machine Learning a la producción de rosas ornamentales bajo invernadero es reducida, pero no inexistente. (Albarico, 2023) comparan la regresión logística multinomial, las máquinas de vectores de soporte, las redes neuronales artificiales y los bosques aleatorios para determinar el tipo de mejora ambiental que conviene aplicar en un invernadero de rosas con el fin de sostener la producción a lo largo del año, y reportan el bosque aleatorio como el modelo de mejor desempeño, evaluado mediante puntaje F1 ponderado, precisión y exhaustividad ponderadas y coeficiente kappa de Cohen.

Por otro lado (Ahlawat et al., 2015) desarrollaron un sistema experto difuso para predecir el rendimiento de rosa bajo invernadero a partir de siete variables de entrada: temperatura, radiación solar, humedad, concentraciones de nitrógeno, fósforo y potasio, y área foliar,

evaluado mediante coeficiente de determinación, raíz del error cuadrático medio, error absoluto medio y error porcentual absoluto medio, es importante enfatizar que al ser un sistema de lógica difusa codifica reglas formuladas por un experto y no ajusta sus parámetros a partir de un registro histórico, de modo que constituye un antecedente de predicción de rendimiento de rosa mediante inteligencia artificial, pero no de Machine Learning.

La producción de rosa cuenta además con antecedentes de modelado cuantitativo por dos vías complementarias. (Silberbush & Lieth, 2004) formulan y calibran un modelo de absorción de nitrato y potasio organizado explícitamente sobre ciclos sucesivos de corte de flor; su variable de salida es la absorción de nutrientes, pero establece un principio que este estudio recoge: la producción de rosa debe interpretarse como un proceso periódico condicionado por el desarrollo de la planta. Por su parte, (Abdelmigid et al., 2022) combinan redes convolucionales, redes con memoria a corto y largo plazo y perceptrones multicapa para predecir características fenotípicas de la rosa de Taif a partir de información espacial y temporal recogida en diez fincas de Arabia Saudita. Ambos trabajos sustentan la pertinencia de representar los ciclos productivos y las relaciones temporales del cultivo, pero ninguno estima el número de tallos comerciales que se cosechará; de manera complementaria (Demir et al., 2024) construyen modelos de predicción de rendimiento en rosa de aceite en Turquía a partir de imágenes multiespectrales captadas con vehículos aéreos no tripulados, comparando bosques aleatorios, árboles de decisión, el algoritmo CHAID y *splines* de regresión adaptativa multivariante.

En Ecuador, (Herrera et al., 2024) combinan la arquitectura YOLO en su versión 5 con el algoritmo DeepSORT y un filtro de Kalman para contar botones abiertos y cerrados de rosa de la variedad Explorer en vídeo captado por drones y alcanzan una precisión media del 94,1 % y un coeficiente de determinación de 0,998 frente al conteo manual; sobre la misma variedad,

(Herrera et al., 2024) emplean máquinas de vectores de soporte para clasificar etapas del proceso productivo, con una exactitud reportada del 99,21 %.

La actual revisión permite situar con precisión la distancia entre estos antecedentes y el problema analizado en el presente estudio. (Albarico, 2023) y (Herrera et al., 2024) resuelven problemas de clasificación. (Ahlawat et al., 2015) predice rendimiento mediante reglas expertas y sobre información agregada. (Abdelmigid et al., 2022) predice fenotipo. (Herrera et al., 2024) y (Demir et al., 2024) estiman el estado fenológico del cultivo a partir de imágenes lo cual implica generar un inventario in situ de la fenología del cultivo pero es una tarea distinta de pronosticar un volumen futuro. (Silberbush & Lieth, 2004) modelan absorción de nutrientes; sin embargo, ninguno de los estudios constituye un trabajo de pronóstico del volumen de producción de tallos multivarietal con ventanas de tiempo definidas.

2.2. Marco Teórico

La estrategia general del proyecto se basó en un enfoque de modelado impulsado por datos utilizando técnicas de *Machine Learning* y *Deep Learning* para predecir el rendimiento de los cultivos florícolas. En el contexto de invernaderos, el *Machine Learning* resulta superior a los modelos biofísicos explicativos debido a su adaptabilidad, además, nos permite aprovechar los datos históricos de rendimiento y del microclima (temperatura, humedad, radiación, CO₂) para modelar relaciones complejas y no lineales, evitando así los problemas de los enfoques biofísicos: la necesidad de una parametrización exhaustiva, la alta sensibilidad a los parámetros y la restricción a tipos de plantas específicos (Gong et al., 2021).

El éxito de los algoritmos se midió por su capacidad para generalizar estos patrones climáticos y fenológicos sobre un conjunto de datos de prueba, minimizando el error frente a la predicción humana (*baseline*). Para garantizar la validez científica y evitar la fuga de información temporal, la estrategia de validación utilizará validaciones cruzadas basadas en ventanas de tiempo o particiones estrictas donde

los datos de prueba siempre sean cronológicamente posteriores a los de entrenamiento (Paudel et al., 2021).

2.2.1. Familias de Algoritmos de Machine Learning para Regresión Aplicada

El modelado del rendimiento florícola en invernaderos es un problema de alta dimensionalidad, donde variables fuertemente correlacionadas (como la radiación solar y la temperatura) afectan el desarrollo de la flor (Sun & Chang, 2023). Con base en el estado del arte, se entrenarán las siguientes familias de algoritmos:

2.2.1.1. Familia de Ensamblados basados en Árboles

Los métodos de ensamble son altamente efectivos porque combinan múltiples árboles de decisiones para reducir el sesgo y la varianza, mitigando el sobreajuste (Mancer et al., 2024).

Los modelos que se considerarán de esta familia son:

- ***Random Forest (RF)***

Es uno de los algoritmos más robustos y populares para predecir rendimientos. Es idóneo para la floricultura porque maneja de forma excelente la alta dimensionalidad, es estable ante datos ruidosos de los invernaderos y puede modelar interacciones complejas sin requerir un escalado exhaustivo de las características (Shawon et al., 2025).

En este algoritmo se decide cuántos árboles construir. Definamos como M a estos árboles, para cada árbol se repiten los siguientes pasos:

1. Se extrae una muestra *bootstrap* del conjunto de datos de entrenamiento.
2. Se seleccionan f características al azar de entre todas las características.
3. Se elige la mejor división utilizando solo f características y se divide el nodo en dos nodos hijos.

4. Se repiten los pasos 2 y 3 hasta que se cumpla alguno de los criterios de parada definidos.

La principal diferencia con los árboles regulares radica en el muestreo aleatorio de características en cada división, lo que aumenta la aleatoriedad y reduce la correlación entre las salidas de los diferentes árboles (Joseph & Tackes, 2022). Al realizar las predicciones, se utiliza cada uno de estos M árboles para obtener una predicción. Para problemas de regresión, se calcula el promedio, mientras que para problemas de clasificación se toma la mayoría de los votos. La función que se aprende para la predicción de la regresión es la siguiente:

$$\hat{y} = \frac{1}{M} \sum_{t=1}^M T_t(x)$$

En donde $T_t(x)$ es la salida del t^{th} árbol del bosque aleatorio (RF).

- ***Gradient Boosting Trees***

Basado en la construcción de árboles, secuencialmente, para corregir los errores residuales del árbol anterior (RF). Particularmente, algoritmos como *XGBoost* y *LightGBM* son extremadamente eficientes procesando grandes volúmenes de datos tabulares y capturan relaciones no lineales con alta precisión.

El *Gradient Boosting* es una forma más general de optimización que se plantea en términos de minimizar una función de coste. Existen diversos algoritmos de *boosting*, y la idea básica detrás de todos ellos es esencialmente la misma. Para ilustrar su funcionamiento, consideremos Adaboost, el más fácil de entender, que funciona de la siguiente manera:

1. Se observación $w_i = \frac{1}{N}$ para $i = 1, 2, \dots, N$. Se inicializa el modelo de conjunto $\hat{F}_0 = 0$.
2. Utilizando inicializa M , el número máximo de modelos a ajustar, y se establece el contador de iteraciones $m = 1$. Se inicializan los pesos de los pesos de observación $w_1, w_2, \dots, w_N$, se entrena un modelo $\hat{f}_m$ que minimice el error ponderado e_m , definido como la suma de los pesos de las observaciones mal clasificadas.
3. Se agrega el modelo al conjunto: $\hat{F}_m = \hat{F}_{m-1} + \alpha_m \hat{f}_m$ donde $\alpha_m = \frac{\log 1 - e_m}{e_m}$.
4. Actualiza los pesos $w_1, w_2, \dots, w_N$ para que aumenten los pesos de las observaciones que fueron clasificadas erróneamente. La magnitud del aumento depende de α_m ; valores mayores de α_m dan como resultado pesos mayores.
5. Se incrementa el contador del modelo $m = m + 1$. Si $m \leq M$, vuelve al paso 2.

La estimación mejorada (*boosting*) viene dada por:

$$\hat{F} = \alpha_1 \hat{f}_1 + \alpha_2 \hat{f}_2 + \dots + \alpha_M \hat{f}_M$$

Al aumentar la ponderación de las observaciones mal clasificadas, el algoritmo obliga a los modelos a entrenarse más con los datos en los que obtuvieron malos resultados.

El factor α_m garantiza que los modelos con menor error tengan mayor ponderación.

El *gradient boosting* es similar a Adaboost, pero plantea el problema como una optimización de una función de coste. En lugar de ajustar los pesos, ajusta los modelos a un *pseudoresiduo*, lo que implica un entrenamiento más intensivo con los residuos de mayor tamaño. Siguiendo la filosofía del bosque aleatorio, el gradient boosting estocástico añade aleatoriedad al algoritmo mediante el muestreo de observaciones y variables predictoras en cada etapa (Bruce y otros, 2020).

En este estudio también se exploraron los algoritmos *HistGBT* y *CatBoost*, ambos algoritmos son de tipo *gradient boosting* basados en árboles de decisiones. *CatBoost* fue desarrollado por un equipo de investigadores e ingenieros de Yandex, este modelo se emplea en una amplia gama de aplicaciones que abarcan motores de búsqueda, sistemas de recomendación, asistentes personales, vehículos autónomos y predicción meteorológica, entre muchas otras tareas. Además, al tratarse de un proyecto de código abierto, se encuentra a libre disposición de cualquier persona o institución para su uso e implementación (CatBoost, 2026). Su funcionamiento se basa en optimizar el algoritmo de *gradient boosting* tradicional mediante dos pilares: el procesamiento de variables categóricas en tiempo de entrenamiento y la mitigación del sesgo de gradiente para reducir el sobreajuste; utiliza árboles simétricos, donde se aplica el mismo criterio de división en todo un nivel del árbol, lo que facilita una estructura equilibrada y una inferencia más veloz (Dorogush, 2018).

Por su parte, el algoritmo *HistGBT* (abreviatura de *Histogram-based Gradient Boosting*) está diseñado para optimizar la búsqueda del mejor *split* al ejecutar los árboles de decisión en grandes conjuntos de datos (Brownlee, 2021).

- ***Stacked Ensemble Models***

También conocidos como modelos de ensamble apilado, utilizan un metamodelo para combinar las predicciones de varios algoritmos base, por ejemplo, *Ridge*, *Random Forest* y/o *XGBoost*. Estos algoritmos son idóneos para cultivos florícolas porque gestionan eficazmente la multicolinealidad inherente a los datos de invernadero donde humedad, temperatura y déficit de presión de vapor están acoplados. Al combinar modelos, logran un ajuste superior y reducen significativamente el margen de error (Mancer et al., 2024).

2.2.1.2. Familia de Deep Learning

El aprendizaje profundo es fundamental para modelar la naturaleza secuencial del desarrollo floral, donde el estrés ambiental de periodos pasados afecta la cosecha actual.

- **Redes Neuronales Recurrentes (RNN) y Long Short-Term Memory (LSTM)**

Las celdas LSTM incluyen "puertas" de entrada, salida y olvido que regulan la retención de información a largo plazo. El rendimiento florar depende de eventos que ocurren cronológicamente como la acumulación térmica o las horas de luz, es por esto por lo que la combinación de RNN y LSTM resulta de utilidad ya que fue diseñada para analizar datos temporales y espaciales de manera simultánea, por lo tanto, esta combinación analiza secuencias temporales completas de parámetros ambientales combinadas con el historial de rendimiento, esto hace que se adapte a las fases de crecimiento de la flor (Shawon et al., 2025).

- **Temporal Convolutional Networks (TCN)**

Este algoritmo aplica convoluciones causales dilatadas a lo largo del tiempo (Gong et al., 2021). Al fusionar TCN con RNN/LSTM, el modelo extrae características climáticas a diferentes escalas temporales (diarias, semanales), mejorando la comprensión de cómo los picos de clima extremo afectan la longitud del tallo y el tiempo de floración.

2.2.1.3. Familia de Modelos Lineales y de Soporte Vectorial

- **Support Vector Regression (SVR)**

Es sumamente robusto para regresión en espacios de alta dimensionalidad. A través de funciones kernel no lineales, SVR puede proyectar los datos de entrada del invernadero para encontrar un hiperplano que minimice el error de predicción del rendimiento (Sun & Chang, 2023).

2.2.2. Criterios Matemáticos para la Cuantificación del Error Predictivo

Para determinar con rigurosidad qué algoritmo es el modelo ganador, se especificarán las siguientes métricas de error estándar para regresión, altamente recomendadas en la literatura de predictiva agrícola (Mancer et al., 2024).

- Raíz del Error Cuadrático Medio (*RMSE*)

Es la métrica primaria. Mide la raíz de las diferencias cuadráticas medias entre el rendimiento floral predicho y el cosechado. El *RMSE* penaliza de forma severa los errores grandes. En floricultura, subestimar o sobreestimar drásticamente un pico de producción para fechas clave (como San Valentín o el Día de la Madre) tiene consecuencias logísticas y económicas catastróficas. El *RMSE* guía al modelo para evitar estos fallos extremos.

Se define como (Oracle, 2026):

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

En donde:

- $\hat{y}_i$ son los valores pronosticados
- y_i son los valores observados
- n es el número de observaciones

- Error Absoluto Medio (MAE)

Mide la magnitud promedio de los errores absolutos. Es altamente interpretable ya que expresa el margen de error promedio en las mismas unidades de la variable objetivo (por ejemplo, número de tallo por metro cuadrado o kilogramos). Esto facilita la comunicación técnica con los agrónomos, gerentes de finca y otros involucrados en el proceso de toma de decisiones.

Se define como (DataCamp, 2026):

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

En donde:

- $\hat{y}_i$ son los valores pronosticados
- y_i son los valores observados
- n es el número de observaciones

- Coeficiente de Determinación (R^2)

Expresa la proporción de la varianza en el rendimiento florícola real que el modelo logra explicar a partir de las variables de entrada, esto permite determinar la "bondad de ajuste". Un modelo ganador debe demostrar un R^2 alto (cercano a 1.0), indicando que las fluctuaciones predichas siguen de cerca las variaciones reales de la cosecha.

Se define como (Bruce y otros, 2020):

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

- Error Porcentual Absoluto Medio (MAPE) y RMSE Normalizado (nRMSE)

Transforman el error absoluto en un formato porcentual. Son métricas indispensables si el modelo se escalará a diferentes variedades de flores o fincas de distintos tamaños, permitiendo comparar el éxito del modelo de forma equitativa y transversal.

Se define como (Arize, 2023):

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{A_i - F_i}{A_i} \right| \times 100$$

En donde:

- n es el número de puntos de datos.
- A_i es el valor real para el i -ésimo punto de datos.
- F_i es el valor pronosticado para el i -ésimo punto de datos.

2.3. Terminología

2.3.1. Metodología General

- **CRISP-DM**

El acrónimo de *Cross-Industry Standard Process for Data Mining*, es una metodología que organiza el trabajo en seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue, se consideran como partes de un ciclo iterativo y no como una secuencia lineal, de modo que los hallazgos de una fase pueden obligar a revisar las anteriores; su valor metodológico reside en independizar el proceso analítico de la herramienta y del sector de aplicación (Chapman y otros, 2000).

- **Tabla de Ablación**

Diseño experimental en el que se construye un conjunto de variantes de un mismo sistema, cada una de las cuales modifica un único componente respecto de una configuración de referencia, manteniendo constantes todos los demás. Permite atribuir cualquier diferencia de desempeño observada al componente alterado, en lugar de a una combinación no identificable de cambios simultáneos. Se distingue de una comparación de modelos convencional en que su propósito no es seleccionar el mejor candidato, sino cuantificar la contribución marginal de cada decisión de diseño (Meyes y otros, 2019).

- **Contraste Controlado**

Comparación entre dos configuraciones que difieren exclusivamente en el eje de diseño que se desea evaluar, y que se realiza sobre el mismo conjunto de observaciones, el mismo horizonte de predicción y la misma métrica. La exigencia de igualdad en todo lo demás es lo que convierte la diferencia observada en evidencia atribuible. Un contraste realizado sobre poblaciones distintas no mide el efecto del eje evaluado, sino la diferencia entre poblaciones.

- **Hiperparámetro**

Parámetro cuyo valor se fija antes del proceso de entrenamiento y que gobierna la forma en que el algoritmo aprende, a diferencia de los parámetros del modelo, que se estiman a partir de los datos durante ese proceso. Ejemplos habituales son la profundidad máxima de los árboles, la tasa de aprendizaje o el número de iteraciones. Su selección constituye un problema de optimización en sí mismo y debe realizarse sobre datos ajenos al conjunto de evaluación final (Bergstra & Bengio, 2012).

- **RandomizedSearchCV**

Procedimiento de búsqueda de hiperparámetros implementado en la biblioteca scikit-learn que muestrea aleatoriamente un número fijo de configuraciones dentro de distribuciones definidas para cada hiperparámetro, en lugar de evaluar exhaustivamente todas las combinaciones posibles. La justificación teórica de la búsqueda aleatoria frente a la exhaustiva es que, cuando solo un subconjunto de los hiperparámetros resulta relevante, el muestreo aleatorio explora más valores distintos de esos hiperparámetros con el mismo presupuesto de cómputo (Bergstra & Bengio, 2012). En el presente estudio no se empleó esta implementación: la búsqueda se realizó mediante un procedimiento propio sobre

una rejilla acotada, evaluado con el mismo esquema de validación temporal del estudio

2.3.2. Validación y Muestreo Temporal

- **Walk-forward**

Metodología de optimización basada en validación cruzada, diseñada para identificar parámetros robustos y mitigar el sobreajuste. Se estructura mediante un enfoque iterativo que segmenta los datos históricos en ventanas secuenciales, un período *in-sample* (IS), utilizado para la calibración del modelo, y un período *out-of-sample* (OOS), destinado a la validación con datos no observados (Castany, 2023).

- **Fold**

Cada uno de las particiones o segmentos *train/test* dentro de un esquema de validación cruzada. En *walk-forward*, un fold es un par (*semanas_train*, *semanas_test*).

- **Ventana expansiva**

Es una variante de *walk-forward* donde el conjunto de entrenamiento crece en cada *fold* siempre desde el inicio del historial hasta el punto de corte, en vez de deslizarse manteniendo un tamaño fijo.

- **Embargo**

Se define como la brecha temporal deliberada entre el fin del período de entrenamiento y el inicio del período de prueba dentro de un *fold walk-forward*.

Evita el riesgo de fuga de información de una *feature* con ventana móvil larga.

- **Ventana Propia**

En este estudio, se define de esta forma a la ventana temporal en la que cada variante conserva su propio rango de fechas de prueba, es decir, no se recorta a la intersección con las ventanas de las otras variantes.

- **Ventana común**

Al contrario que en la ventana propia, se usa la misma ventana temporal para todas las variedades y bloques, esto resulta en que todas las variantes tengan una predicción simultánea.

2.3.3. Términos de Dominio

- **Tallos Cosechados**

Número de tallos de rosa cortados y entregados a postcosecha durante un día determinado con una combinación de bloque y variedad. En la base cruda de producción esta magnitud corresponde a la columna S1 RECEIVING; en la base analítica consolidada se denomina Tallos Cosechados. Constituye la variable objetivo del estudio: agregada por suma a resolución de semana ISO, define $y(b, v, s)$, la producción del bloque b y la variedad v en la semana s .

- **Tallos Pinchados**

El pinch es una labor agronómica de manejo del cultivo que consiste en la eliminación del ápice de crecimiento de un brote para redirigir el desarrollo de la planta hacia nuevos tallos productivos. Precede biológicamente a la cosecha del tallo resultante, por lo que constituye una señal adelantada de producción y es la base del predictor biológico del modelo. El registro es de cadencia semanal y contiene el número de tallos sometidos a pinch en una combinación de bloque y variedad.

- **Bloque**

Unidad física de producción de la unidad productiva analizada, correspondiente a un invernadero o sección de invernadero identificada por un código numérico interno. Constituye, junto con la variedad y la semana ISO, la unidad de análisis del estudio. La base analítica registra 64 bloques; tras aplicar los criterios de elegibilidad y de cobertura mínima, el panel modelado comprende 61.

- **Variedad**

Cultivar de rosa sembrado en un bloque determinado, identificado por un código interno de cuatro caracteres. Cada combinación de bloque y variedad presenta un patrón productivo propio, derivado de su ciclo de maduración, su densidad de siembra y su edad de plantación. La base analítica registra 141 variedades; tras aplicar los criterios de elegibilidad, al menos 250 días de historia y 8 000 tallos acumulados, el panel modelado comprende 118, y 30 de ellas disponen además de un modelo individual entrenado exclusivamente con sus propios registros.

- **Ventana de Maduración Biológica**

Intervalo temporal, expresado en días, que transcurre entre el pinch de un brote y la cosecha del tallo que de él resulta. En este estudio se estima empíricamente para cada variedad mediante análisis de correlación cruzada entre las series diarias de tallos pinchados y tallos cosechados, explorando desfases entre 40 y 150 días. Se acepta la ventana específica de una variedad cuando la correlación máxima alcanza o supera 0,25, y su amplitud se delimita por el conjunto de desfases cuya correlación conserva al menos el 80 % del valor de pico. El desfase de pico presenta una mediana de 98 días sobre las 118 variedades analizadas, con un rango entre 42 y 147 días. La estimación específica resulta admisible en 37 variedades; las 81 variedades restantes emplean una ventana genérica de 85 a

110 días. Toda variable climática y biológica del modelo se agrega sobre esta ventana, que termina en la fecha de anclaje.

2.3.4. Ingeniería de Variables

- **Rezago**

También definido como *lag*, es el valor de una variable en un momento anterior en el tiempo, en este estudio, es usado como predictor del presente.

- **Función de correlación cruzada**

Es una función que mide qué tan correlacionadas están dos series cuando una se desplaza respecto a la otra en el tiempo, abreviado como CCF (*Cross-Correlation Function*).

- **Ventana CCF**

También referida como ventana biológica, es el rango de días alrededor del pico de la curva CCF donde la correlación se mantiene sobre el 80% de su valor máximo.

- **Fecha ancla**

Es la fecha de referencia fija dentro de cada semana, corresponde a un día específico que es usado para anclar el cálculo de las ventanas de *features* en el panel semanal.

- **One-hot encoding**

Técnica que consiste en convertir las columnas categóricas con n elementos a columnas binarias con valores de cero o uno.

- **Encoding categórico nativo**

Se trata de permitir que cada familia de modelo maneje las variables categóricas (bloque y variedad) con su propio mecanismo interno, en vez de expandirlas a

columnas binarias, evitando de esta forma tanto la alta dimensionalidad del *one-hot encoding* como el riesgo de fuga de un *target encoding* manual.

- **Imputación**

Técnica mediante la que se rellenan valores faltantes propagando el último valor observado hacia adelante (*ffill, forward-fill*) o el próximo valor observado hacia atrás (*bfill, backward-fill*).

- **Fuga de Información**

También referido como *data leakage*, ocurre cuando una feature contiene, directa o indirectamente, información del futuro respecto al momento que se está prediciendo, provocando que se infle artificialmente su desempeño medido y no sea sostenible en producción.

2.3.5. Modelos y Algoritmos

- **Especialización Global**

Estrategia que entrena un solo modelo con todas las combinaciones de bloque y variedad, incluyendo su identidad como información categórica; permite que las unidades con escasa historia se beneficien de los patrones aprendidos en el resto del conjunto, a costa de imponer una estructura funcional compartida (Montero-Manso & Hyndman, 2021).

- **Especialización Individual**

Estrategia que ajusta modelos separados para grupos específicos; captura particularidades que un modelo compartido no puede representar, pero dispone de menos datos por modelo y resulta más sensible al sobreajuste. La implementación actual se entrena un modelo por variedad para las 30 variedades con mayor volumen; sus bloques se mantienen dentro del mismo modelo y se tratan como categoría cuando corresponde. Por tanto, no se ajusta un modelo

independiente por cada combinación de bloque y variedad (Montero-Manso & Hyndman, 2021).

- **Especialización Residual por Variedad**

Estrategia intermedia consistente en emplear un modelo global como predictor base y ajustar sobre sus residuos un modelo auxiliar específico para cada unidad. Busca combinar la capacidad de generalización del modelo compartido con una corrección de los sesgos sistemáticos propios de cada unidad.

- **Corrección Residual**

Ajuste de un segundo modelo sobre los errores del modelo base. La corrección se suma a la predicción inicial. Debe entrenarse dentro de cada pliegue y solo es útil si los residuos conservan una estructura predecible.

2.3.6. Arquitectura y Entrenamiento de Deep Learning

- **Secuencia (ventana temporal, SEQUENCE_WINDOW)**

Conjunto ordenado de observaciones consecutivas que constituye la unidad de entrada de un modelo secuencial. Su longitud determina cuánta historia puede considerar el modelo para producir una predicción: una ventana corta impide capturar dependencias prolongadas, mientras que una excesivamente larga incrementa el coste de cómputo y diluye la señal reciente.

- **Covariable Estática vs. Dinámica**

Una covariable dinámica cambia en cada paso de la secuencia. Una covariable estática permanece constante, como la identidad de bloque o variedad, y se integra por una entrada separada. En los modelos LSTM y TCN del notebook, estas identidades se representan mediante one-hot encoding

- **Walk-forward multi-fold para modelos secuenciales**

Aplicación de la validación temporal a modelos que reciben secuencias. Ninguna secuencia de entrenamiento puede invadir el período de prueba, y la normalización debe ajustarse solo con el tramo de entrenamiento de cada pliegue.

- **Convolución causal**

Operación de convolución en la que la salida correspondiente a un instante depende exclusivamente de observaciones de ese instante y anteriores, nunca posteriores. Se implementa desplazando el relleno de la señal hacia el pasado, y es la condición que hace admisible el uso de arquitecturas convolucionales en predicción de series temporales (van den Oord y otros, 2016).

- **Convolución dilatada**

Convolución que aplica el filtro con espacios entre entradas. Al aumentar la dilatación entre capas, amplía el campo receptivo sin requerir un número equivalente de capas o parámetros (Bai y otros, 2018).

- **Bloque residual**

Estructura arquitectónica en la que la salida de un conjunto de capas se suma a la entrada de ese mismo conjunto mediante una conexión directa. La suma facilita la propagación del gradiente a través de redes profundas y permite que el bloque aprenda una corrección sobre la señal de entrada en lugar de una transformación completa (He y otros, 2016).

2.3.7. Explicabilidad e Importancia de Variables

- **Importancia por permutación (agrupada)**

Medida del deterioro del desempeño cuando se permutan conjuntamente variables relacionadas. La permutación rompe su asociación con el objetivo y

permite estimar cuánto depende el modelo de ese grupo. Debe calcularse fuera del entrenamiento. La agrupación usada en este proyecto es una extensión operativa de la importancia por permutación (Breiman, 2001; Fisher y otros, 2019)

- **Grupos de importancia**

Conjuntos de variables organizados por su función: autorregresivas, biológicas, climáticas, de calendario o de identidad. Permiten responder qué tipo de información sostiene la predicción, sin reducir la explicación a una columna aislada.

- **Importancia individual (variable por variable)**

Estimación de la relevancia de cada variable por separado. Resulta informativa cuando las variables son aproximadamente independientes, y engañosa cuando existen correlaciones fuertes entre ellas, ya que en ese caso el modelo puede sustituir una variable por otra y la permutación de cualquiera de ellas produce un deterioro menor al que correspondería a la señal que ambas representan.

- **Importancia nativa**

Puntuación interna calculada por un algoritmo para resumir el uso de las variables. Su significado depende del modelo: puede basarse en ganancia, cambios en la predicción u otro criterio propio. No es equivalente a la importancia por permutación y no debe compararse directamente entre algoritmos (Prokhorenkova y otros, 2018).

- **Atajo de autocorrelación**

Fenómeno por el cual un modelo aplicado a una serie con fuerte dependencia temporal concentra su capacidad predictiva en los valores recientes de la propia variable objetivo, en detrimento de las variables explicativas externas. El desempeño resultante puede ser satisfactorio, pero el modelo se comporta

esencialmente como un predictor de persistencia sofisticado, y las conclusiones sobre el papel de las variables externas quedan condicionadas por ese comportamiento.

2.3.8. Cuantificación de Incertidumbre

- **Intervalo de predicción**

Rango de valores dentro del cual se espera que se sitúe una observación futura con una probabilidad especificada. Se distingue del intervalo de confianza en que éste acota un parámetro poblacional desconocido pero fijo, mientras que aquél acota una realización aleatoria aún no observada, y es por tanto necesariamente más ancho (Chatfield, 1993).

- **Predicción conformal**

Marco que transforma errores de calibración en conjuntos o intervalos predictivos. Bajo intercambiabilidad entre los datos de calibración y los futuros, ofrece cobertura marginal finita sin imponer una distribución paramétrica del error ni una forma específica del modelo. Su propiedad distintiva es que ofrece garantía de cobertura sin exigir supuestos sobre la distribución de los errores ni sobre la forma del modelo, requiriendo únicamente intercambiabilidad entre las observaciones de calibración y las de predicción (Vovk y otros, 2005; Angelopoulos & Bates, 2023).

- **Recalibración rodante / expansiva**

Actualización temporal del conjunto de calibración mediante una ventana fija que avanza o un histórico que crece. Busca adaptar los cuantiles de error cuando cambia el proceso, pero no convierte automáticamente una serie temporal en datos intercambiables.

- **Cobertura empírica vs. objetivo**

La cobertura objetivo es el nivel nominal usado para construir los intervalos. La cobertura empírica es la proporción de observaciones futuras que realmente queda dentro de ellos. Una cobertura baja indica intervalos demasiado estrechos; una cobertura mayor que la objetivo puede señalar intervalos poco útiles por su amplitud.

CAPITULO 3

3. DESARROLLO

3.1. Análisis y Limpieza de Datos

3.1.1. Descripción de Archivos de Entrada

La construcción del conjunto de datos de modelado partió de fuentes operativas independientes, la base de cosecha, con registros diarios de tallos recibidos por bloque y variedad, y la base de *pinch*, con registros semanales de tallos pinchados por bloque y variedad. Dado que la variable objetivo se define a nivel diario y el *pinch* se registra a nivel semanal, fue necesario, por un lado, depurar cada fuente de manera documentada y, por otro, resolver la asimetría temporal entre ambas antes de integrarlas en una única base analítica. La totalidad del procedimiento se implementó en un *notebook* de Python reproducible y parametrizado, de modo que su re-ejecución sobre versiones actualizadas de las fuentes genere resultados idénticos y trazables. En la **¡Error! No se encuentra el origen de la referencia.** se describe todas las fuentes de datos utilizadas.

Tabla 1
Descripción de las fuentes de datos de entrada

Archivo	Descripción
Metricas_Diarias_Davis.csv	Contiene los datos climáticos de una estación meteorológica ubicada fuera de los invernaderos.
Base_Diaria_Cosecha_Pinch_Corregido.xlsx	Contiene los datos de <i>pinchs</i> y cosechas realizadas a través del tiempo por cada variedad de rosa.
HInputs_<mes>.csv	Archivo con datos de eventos de sensores IoT dentro de cada invernadero (identificados como bloques). Se trata de un archivo ancho que requiere un esfuerzo de preprocesamiento considerable.
ECO_BD_Proyecciones_Corregido.xlsx	Archivo con las proyecciones realizadas por los técnicos sobre la cantidad de cosecha esperada.

3.1.2. Limpieza datos

3.1.2.1. Cosecha

La base de cosecha comprendía 184.155 registros diarios. Se adoptó como variable objetivo la columna *S1 RECEIVING*, que cuantifica los tallos recibidos por bloque y variedad en cada jornada, y que presentó una calidad elevada con la ausencia total de valores nulos, nulos negativos y ceros, dentro del histórico de datos se evidencio una aparente ausencia de datos 156 días de los cuales se pudo constar que 131 de los sin registro correspondía a domingos, reflejo de que la operación de corte se concentra de lunes a sábado; dado que las 148 semanas ISO del período se encuentran completas, esta característica no representa una pérdida de información y refuerza la pertinencia de un grano de análisis semanal.

3.1.2.2. Pinch

La base de *pinch* comprendía 15.885 registros correspondientes al período comprendido entre la semana 35 de 2023 y la semana 24 de 2026; Sobre ella se aplicaron varias operaciones de depuración.

En primer lugar, se identificó y eliminó un único registro atribuible a un error de captura, el correspondiente al bloque 44, variedad 89CB, en la semana 13 de 2024, que consignaba 560.915 tallos pinchados frente a una meta de apenas 789; dicho valor resultaba inverosímil por dos razones convergentes: excedía en aproximadamente veintisiete veces el máximo histórico de su propia variedad (21.000 tallos, con una mediana de 809) y se situaba fuera de la ventana estacional de despunte intensivo. Al no ser posible reconstruir el valor real a partir de la información disponible, el registro se eliminó en lugar de imputarse, decisión adoptada para preservar la defensividad del conjunto y evitar la introducción de datos artificiales. La eliminación quedó asentada en una bitácora de depuración.

En segundo, dos registros de magnitud elevada pero no confirmables como errores mediante el solo análisis de los datos, la variedad SZOP en la semana 51 de 2025 (60.996 tallos) y la variedad 4UBO en la semana 41 de 2024 (60.000 tallos), se conservaron y se marcaron mediante una variable indicadora para su verificación posterior con el personal de la finca, sin alterar sus valores.

En tercer lugar, cinco combinaciones de año, semana, bloque y variedad que aparecían duplicadas se consolidaron sumando sus tallos pinchados, de modo que cada combinación quedara representada por un único registro semanal.

Finalmente, para resolver la asimetría temporal entre los registros semanales de Pinch y la cosecha diaria, se implementó el algoritmo de reparto proporcional del Mayor Resto (Hamilton). La cuota entera base por día q_i y el residuo r_i se obtienen mediante:

$$q_i = \text{floor}\left(\frac{\text{Pinch}_{\text{semanal}}}{7}\right)$$

$$r_i = \text{Pinch}_{\text{semanal}} \bmod 7$$

Los r_i residuos enteros sobrantes se asignan ordenadamente a razón de un tallo adicional en los primeros r_i días del ciclo semanal, garantizando la conservación exacta del total de tallos en números enteros sin introducir fracciones irreales.

3.1.2.3. Davis

- Fuente de datos y evaluación de calidad

La caracterización del ambiente se basa en los registros de una estación meteorológica Davis con cadencia de quince minutos, que constituye la fuente de clima externo del proyecto. El *export* original presenta particularidades de formato (codificación latin-1, cada línea envuelta en comillas con las internas duplicadas, separador decimal de

coma y faltantes representados como «--») que fueron resueltas mediante una rutina de lectura reutilizable, de modo que el proceso es reproducible ante nuevas descargas.

La base cubre el período comprendido entre el 16 de mayo de 2023 y el 5 de junio de 2026, con 107.018 registros. La evaluación de calidad arrojó un resultado sobresaliente en tres dimensiones. En cuanto a la integridad temporal, el 99,99 % de los intervalos corresponde exactamente a quince minutos, sin registros duplicados y con la totalidad de las marcas temporales correctamente interpretadas; en más de tres años solo se detectaron ocho huecos, que suman apenas un día y siete horas y media. En cuanto a la completitud, únicamente las dos de dirección del viento cuyo sensor no reporta y se descartaron; el resto de las variables presentó alrededor de un 0,1 % de faltantes, equivalentes a noventa y una lecturas dispersas en cincuenta y seis fechas, sin evidencia de caídas prolongadas de sensor, por lo que resultan interpolables.

En cuanto a la sanidad física de los valores, todos los rangos son coherentes: la temperatura oscila entre $-0,4$ y $22,9$ °C, la humedad relativa entre 16 y 98 %, la radiación solar entre 0 y 1.382 W/m^2 sin un solo valor negativo y, como prueba de que el piranómetro está bien calibrado, sin ninguna lectura nocturna superior a 5 W/m^2 en las cerca de cuarenta mil lecturas de noche. Estos rangos, además, confirman el clima de altiplano frío característico del sitio. Finalmente, el 99,1 % de los días dispone de las noventa y seis lecturas completas, condición que habilita el cálculo de integrales diarias (energía solar, DLI y GDD) prácticamente sin interpolación. En conjunto, la fuente combina una calidad muy elevada con una riqueza de sensores tales como: radiación, temperatura, humedad, punto de rocío, viento, lluvia y evapotranspiración

los cuales que habilita la construcción del conjunto completo de variables agroclimáticas descrito a continuación.

○ Variables derivadas: definición, método y unidades

El cálculo sigue una lógica de dos niveles: cada variable se computa primero a la resolución nativa de quince minutos y luego se agrega al día, respetando dos principios fisiológicos. El primero es integrar en el tiempo en lugar de promediar puntos aislados, ya que la planta responde a cantidades acumuladas; el segundo es separar el día de la noche, porque la fotosíntesis ocurre bajo luz y la respiración y la elongación ocurren en oscuridad. Las variables de tipo integral se calculan mediante la regla del trapecio sobre las lecturas de quince minutos. La tabla siguiente resume el conjunto.

Tabla 2
Resumen de datos de fuente de datos Davis

Variable	Fuente	Método de cálculo	Unidad
GDD (tiempo térmico)	Temp	Integral por partes de la temperatura, base 8 °C, tope 28 °C	°C·día
Energía solar diaria	Rad. solar	Integral de la irradiancia: $\sum R \cdot \Delta t$	MJ/m ² /día
DLI (integral lumínica)	Rad. solar	Radiación → PAR → flujo fotónico, integrado en el día	mol/m ² /día
VPD medio diurno	Temp, Hum	$e_s(T) - e_a(T) \cdot HR/100$, promedio en intervalos con luz	kPa
Horas de VPD fuera de banda	Temp, Hum	Intervalos con VPD fuera de [0,5–1,0] kPa × 0,25 h	horas/día

T mínima / media / máxima	Temp	Estadísticos diarios	°C
T media nocturna / diurna	Temp + Rad.	Promedio en intervalos oscuros vs. con luz	°C
DIF (día – noche)	Temp	T media diurna – T media nocturna	°C
Amplitud térmica	Temp	T máxima – T mínima del día	°C
Ratio fototérmico	DLI, GDD	DLI ÷ GDD (con GDD > 0)	mol·m ⁻² ·(°C·d) ⁻¹
Horas productivas	Rad., Temp, Hum	Intervalos con luz, T y VPD en rango × 0,25 h	horas/día
Lluvia diaria	Lluvia	Suma del día	mm/día
Horas de humedad alta	Hum	Intervalos con HR > 85 % × 0,25 h	horas/día
Viento medio / máximo	Viento	Media y máximo del día	km/h

Tres variables merecen una descripción más detallada por la particularidad de su cálculo.

- Importancia fisiológica de cada variable

La selección no es exhaustiva sino dirigida por la fisiología del cultivo. El DLI y la energía solar diaria gobiernan la fotosíntesis y, por tanto, el carbono disponible para formar tallos; en un emplazamiento ecuatorial de fotoperiodo constante, la variación de radiación por nubosidad es el principal motor natural de las oscilaciones de producción. El GDD constituye el reloj del desarrollo: predice cuándo un brote alcanza el punto de corte mejor que el tiempo de calendario y permite alinear el clima con la

fenología. La temperatura nocturna, el DIF y la amplitud térmica determinan la calidad comercial, por ejemplo, las noches frías producen tallos más largos, mejor color y apertura más lenta, un aspecto central en rosa premium. El VPD regula la transpiración y la apertura estomática dentro de una banda óptima, de modo que tanto su exceso (estrés hídrico) como su defecto (transpiración pobre, deficiencias de calcio) son perjudiciales. Finalmente, las horas de humedad alta funcionan como proxy de presión de enfermedad, y la lluvia y el viento aportan contexto de nubosidad y demanda de ventilación.

- Resultados del cálculo diario

El cálculo se aplicó a los 1.117 días del período, de los cuales 1.107 resultaron completos. Antes de interpretar los valores se validó la implementación de la integración mediante un control cruzado: la energía solar diaria calculada por integración propia se comparó con la energía que la estación Davis reporta en Langleys, obteniéndose una correlación de 1,0000 y un error medio absoluto de 0,001 MJ/m². Esta coincidencia confirma que el procedimiento de integración es correcto.

Los valores obtenidos describen con nitidez un ambiente de altiplano frío y luminoso, favorable a la rosa de calidad. La radiación es abundante: el DLI externo promedia 27,0 mol·m⁻²·día⁻¹ (rango 9,6–52,3) y la energía solar 13,1 MJ/m²/día. El régimen térmico es fresco: la temperatura media diaria es de 11,3 °C, con noches frías, media nocturna de 8,9 °C y mínimas que descienden ocasionalmente hasta –0,4 °C, lo que señala un riesgo puntual de helada y un DIF positivo medio de 5,0 °C, favorable a la elongación del tallo. Como consecuencia de este clima fresco, el GDD acumulado es bajo (media de 3,45 °C·día con base 8 °C) y el tope de 28 °C nunca llega a activarse, lo que anticipa ciclos de desarrollo largos.

El hallazgo agronómicamente más relevante se refiere al VPD: el valor medio diurno es de solo 0,45 kPa y en el 64,9 % de los días se sitúa por debajo del umbral inferior de la banda óptima, con un promedio de 7,8 horas diarias fuera de banda. En concordancia, se registran en promedio 12,7 horas diarias de humedad relativa superior al 85 %. Es decir, el ambiente externo es predominantemente húmedo y de bajo déficit de vapor, condición que favorece la presión de Botrytis y puede limitar la transpiración y el transporte de calcio. Esta señal, lejos de ser un defecto de los datos, es información valiosa que el modelo podrá aprovechar. En conjunto, las variables calculadas resultan físicamente consistentes, están completas y capturan tanto el potencial productivo (luz elevada, noches frescas) como los factores de riesgo (baja demanda evaporativa, humedad alta) del emplazamiento.

3.1.2.4. ISSI

Para este estudio se seleccionaron los archivos HInputs_*.csv por contener información climática de cada bloque o invernadero. Estos archivos son la salida cruda del registrador de eventos del sistema de sensores dentro de los invernaderos y no están preparados para el análisis. En la *Tabla 3* se resumen las principales características de los archivos de ISSI.

Tabla 3

Descripción de archivos ISSI de entrada

Característica	Descripción
Separador	;
Codificación	latin-1
Granularidad	Una fila por cada <i>timestamp</i>
Columnas de fecha	Separadas en dos, <i>Date</i> con formato YYMMDD y <i>Time</i> con formato hh:mm:ss
Tipos de datos	Todas las columnas usan float64 excepto las de <i>Date</i> y <i>Time</i> .

Adicionalmente, se encontró que la cantidad de columnas no es fija en todos los archivos, esto a causa de que los sensores disponibles entre cada mes no son constantes, es decir, se agregan o se retiran sensores a lo largo del tiempo, de esta forma, desde septiembre del 2024 hasta marzo del 2026, se tienen 299 columnas y a partir de abril del 2026 la cantidad de columnas se incrementa a 307. En la Tabla 4 se muestran las columnas consideradas, si fueron o no utilizadas para el estudio y el tratamiento aplicado.

Tabla 4
Tipos de columnas relevantes de HInputs

Columna	Uso en el estudio	Tratamiento
Humedad	Si	Sin tratamiento
Humedad (IoT)	Si	Renombrar a “Humedad”
Humedad PMQ(IoT)	Si	Renombrar a “Humedad”
Sensor Luz ePAR (umol)	Si	Sin tratamiento
Sensor Presión (Bar) (IoT)	No	
Sensor Presión PMQ (Bar) (IoT)	No	
Temperatura	Si	Sin tratamiento
Temperatura (IoT)	Si	Renombrar a “Temperatura”
Temperatura PMQ(IoT)	Si	Renombrar a “Temperatura”
Tensiómetro 30cm (cbar)	No	

El nombre de las columnas mostrados en la Tabla 4 corresponde a la métrica de interés, la mayoría de las columnas siguen el patrón “{métrica} Bloque {id}”, por ejemplo, “Temperatura Bloque 4” o “Humedad Bloque 20 (IoT)”. Dentro de cada archivo hay otras columnas que no miden condiciones climáticas pero que registran señales de operación general como bombas, filtros o alarma, definimos a estas columnas como “columnas sin bloque”, las mismas no pueden ser categorizadas como ruido al ser señales reales, sin embargo, son descartadas por no contener información de interés relacionada al clima.

En algunas filas, se presenta un error de transmisión severo en el que no solamente la columna Time resulta ilegible, sino que la corrupción de su valor produce una lectura de un separador (";;") adicional que desplaza los demás valores y el registro queda totalmente ilegible. Este error severo solo se presenta en un registro de los 21 archivos procesados por lo que se decide omitirlo. Además, presenta en 721 registros que requieren reparar el formato de la columna Time, esto representa el 0.14% del total de registros disponibles.

Sobre la frecuencia de los datos disponibles en los archivos, podemos ver en la Figura 1 que la cantidad de registros por mes es relativamente cercana a partir de marzo del 2025, esto coincide con una disminución en la frecuencia temporal entre cada registro que pasó de 1 a 2 minutos a partir del mencionado mes (ver Figura 2).

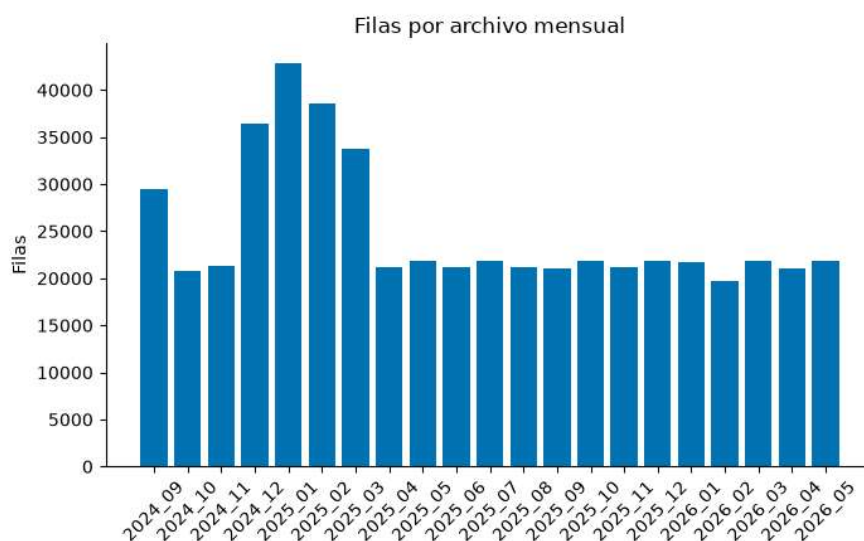


Figura 1 Cantidad de registros en cada archivo ISSI



Figura 2 Frecuencia de registro de eventos en los archivos ISSI

Hasta este punto se disponen de registros con valores de *Date* y *Time* válidos y completos, pero también es relevante identificar si existen columnas de bloques que no contengan información, es decir, invernaderos de los que no se disponen datos climáticos. En la Figura 3 vemos que cerca de 90 columnas de datos climáticos (con bloque) tienen 50% de valores nulos, una de las causas es que en ciertas lecturas, los sensores no estuvieron disponibles. Dado que al final de procesamiento de estos archivos se usarán valores de promedios diarios, estos valores nulos no serán necesariamente relevantes.

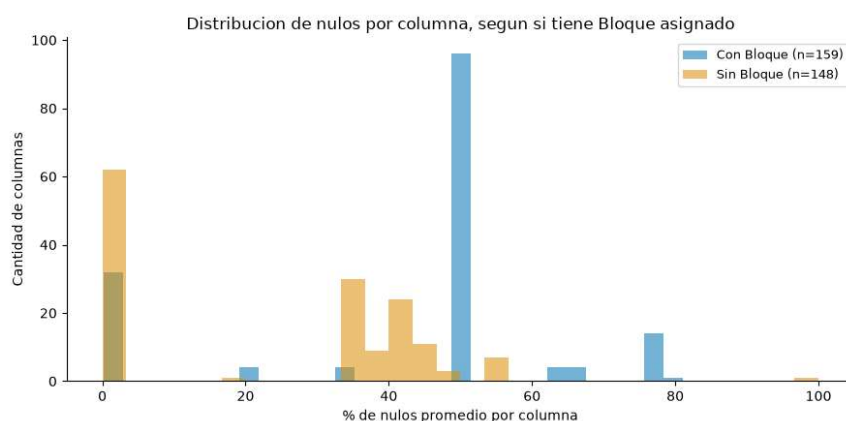


Figura 3 Distribución de valores nulos por columna

Al procesar los archivos ISSI, se enfrentaron dos problemas principales que se describen a continuación, el problema de los valores atípicos y el problema de los sensores congelados.

El problema de los valores atípicos

Se visualizaron los diagramas de caja de los valores de todos los meses disponibles por bloque tanto para Humedad (Figura 5) como para Temperatura (Figura 4) en donde vemos la presencia de valores fuertemente atípicos y valores que a pesar de estar cerca de las cajas, resulta atípicos por la naturaleza de la métrica, por ejemplo, temperaturas de 100° C están cerca del bigote superior pero un invernadero a esa temperatura es físicamente inviable por lo que se deben remover esos valores.

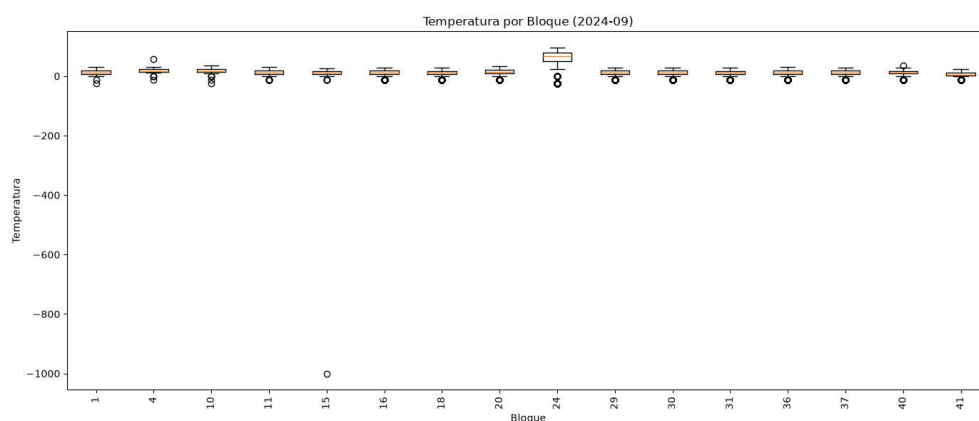


Figura 4 Diagrama de caja de los valores de temperatura en septiembre del 2024

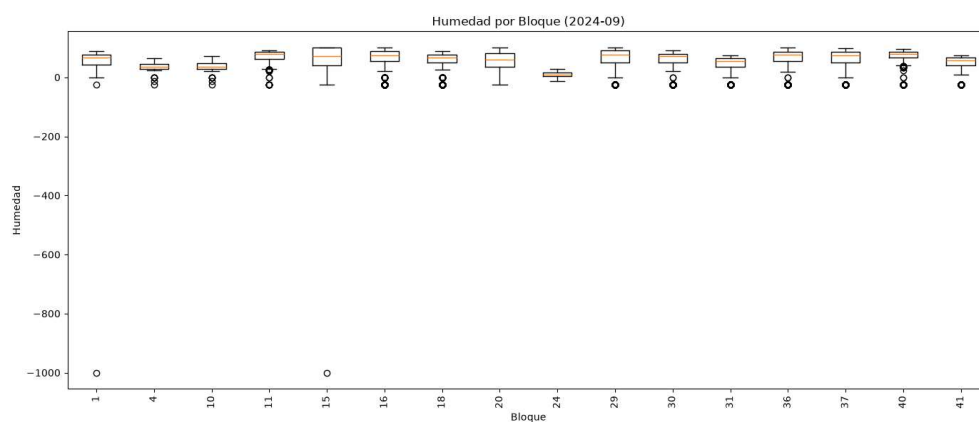


Figura 5 Diagrama de caja de los valores de humedad en septiembre del 2024

Cada Bloque debe presentar una distribución de temperaturas sin valores extremos, por ejemplo, los valores presentados en el Bloque 21 durante septiembre del 2025.

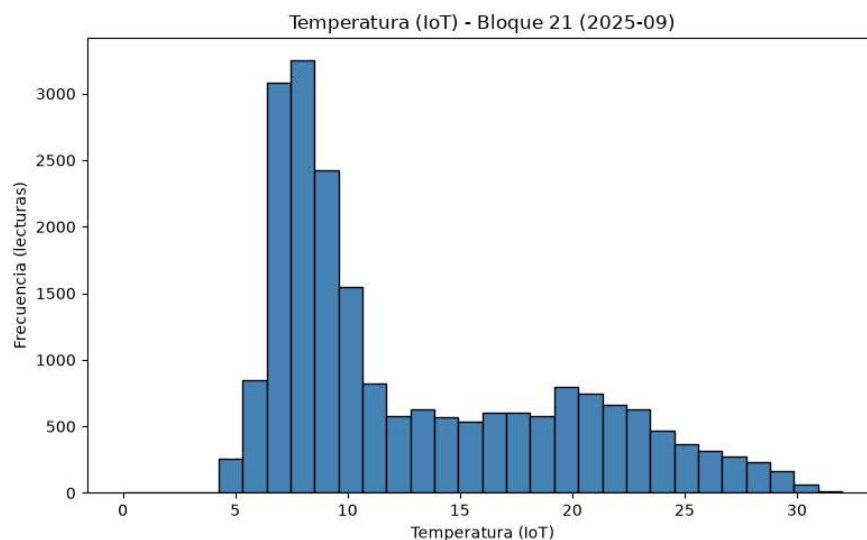


Figura 6 Histograma de temperatura del Bloque 21 en septiembre del 2025

Como primera aproximación, se intentó una limpieza por el método de rango intercuartílico, sin embargo, resultó insuficiente por lo descrito previamente ya que solo usa la posición estadística, no el contexto del dato, su significado ni la causa del valor extremo.

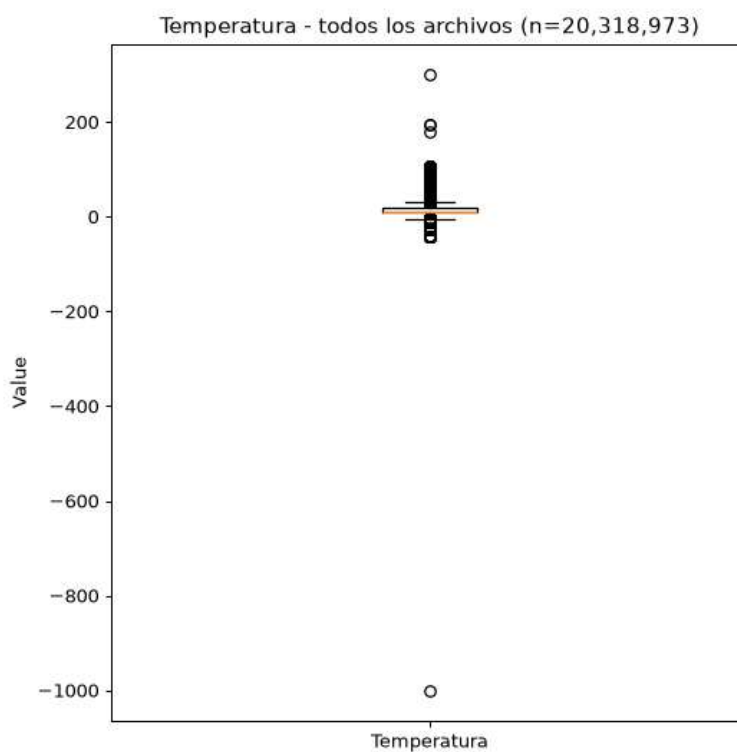


Figura 7 Diagrama de cajas de todos los valores de temperatura

Para el caso de la temperatura (Figura 7), valores por debajo de cero o superiores a 100 son físicamente improbables, esos valores no se registran dentro de un invernadero en condiciones normales. El rango definido para el estudio es de 0 a 40 grados Celsius.

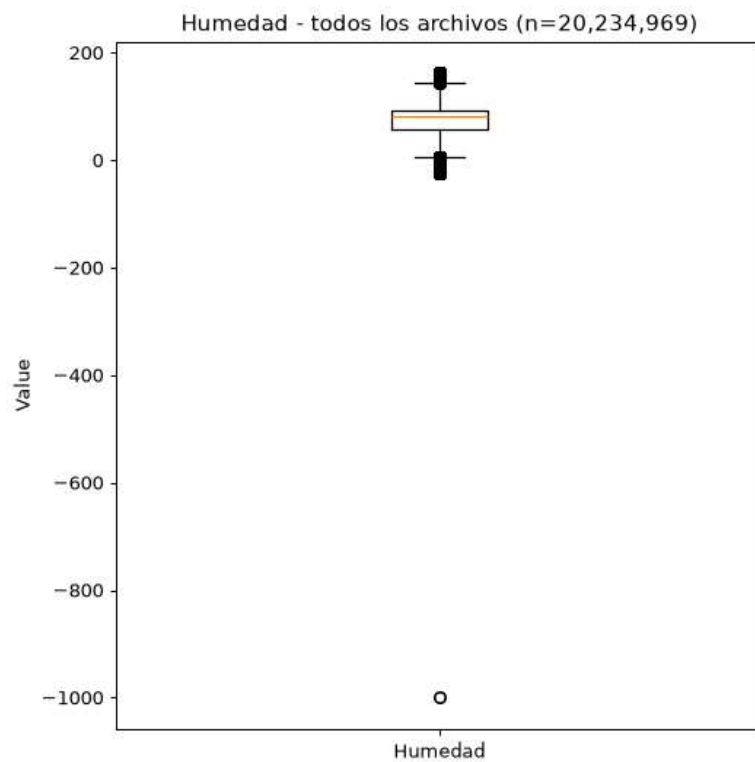


Figura 8 Diagrama de caja de todos los valores de humedad

La humedad (Figura 8) es medida en valores porcentuales en rangos de 0 a 100, por lo tanto, fuera de ese rango son valores que han tenido un problema físico en la lectura y pueden ser ignorados. El rango definido para el estudio es de 25 a 100.

Para finalizar, los valores esperados de luz ePAR (Figura 9) van de 0 a 750, por lo tanto, también es posible establecer un mínimo y máximo válido, sin embargo, para esta métrica, los valores extremos se suavizan convirtiendo la métrica a un valor promediado diario (DLI ePAR (mol/m²/día)).

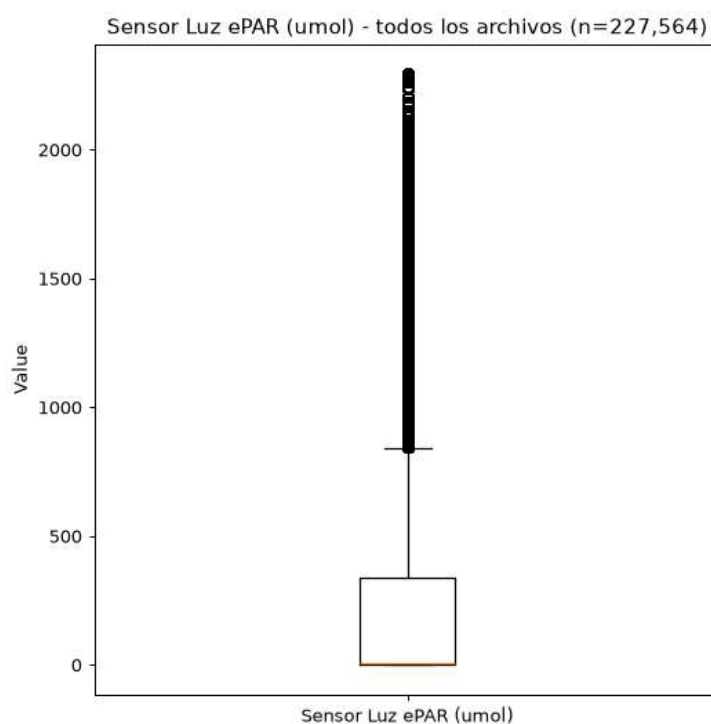


Figura 9 Diagrama de caja de todos los valores del sensor de luz

Como consecuencia, para remover los valores atípicos usamos valores fijos y máximos por cada métrica.

El problema de los valores congelados

Existe un problema conocido en sensores IoT en el ámbito climático, conocido como “*stuck sensor*”, ocurre cuando se observan valores exactamente iguales durante un periodo de tiempo prolongado lo cual es físicamente imposible en variables climáticas que en condiciones normales fluctúan con el transcurso del tiempo. En la Figura 10 vemos cómo existen líneas horizontales en la temperatura de los tres bloques presentados, esto rompe el comportamiento natural de las curvas y puede ser causado por problemas físicos en los sensores.

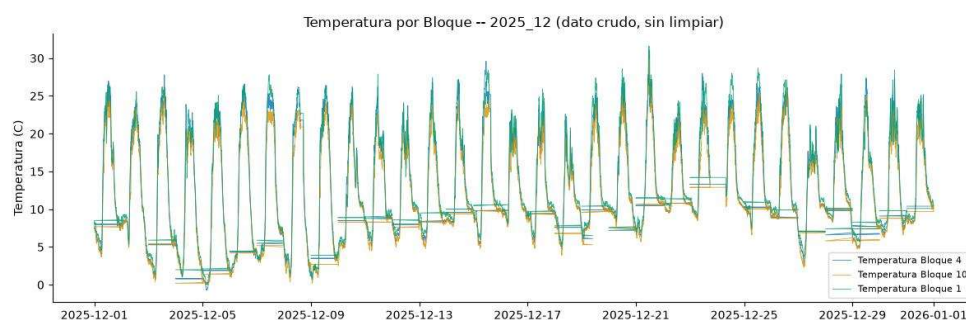


Figura 10 Datos crudos temperatura en diciembre del 2025 de los bloques 1, 4 y 10

Esto fue observado en diferentes meses tanto para los valores de Temperatura como para los de Humedad. La *Tabla 5* resumen todos los casos detectados para la métrica de Temperatura.

Tabla 5

Tramos con lecturas largas de temperatura por mes (sensores congelados)

Mes	Tramos planos largos (≥ 100 lecturas)	Tramo más largo (lecturas)
2024_09	48	598
2024_10	63	10,908
2024_11	51	2,424
2024_12	127	18,573
2025_01	81	42,681 (~59 días)
2025_02	43	24,710 (~34 días)
2025_03	47	9,282
2025_04	51	1,134
2025_05	32	21,877
2025_06	66	21,132
2025_07	60	6,362
2025_08	0	0
2025_09	2	250
2025_10	4	286
2025_11	4	211
2025_12	287	706
2026_01	10	755
2026_02	28	776

2026_03	6	431
2026_04	10	529
2026_05	0	0

Enfocándonos en un mes, de manera gráfica, los valores de sensores congelados se alejan de los que tienen un comportamiento normal como se ve en la Figura 11 para temperatura y la Figura 12 para humedad.

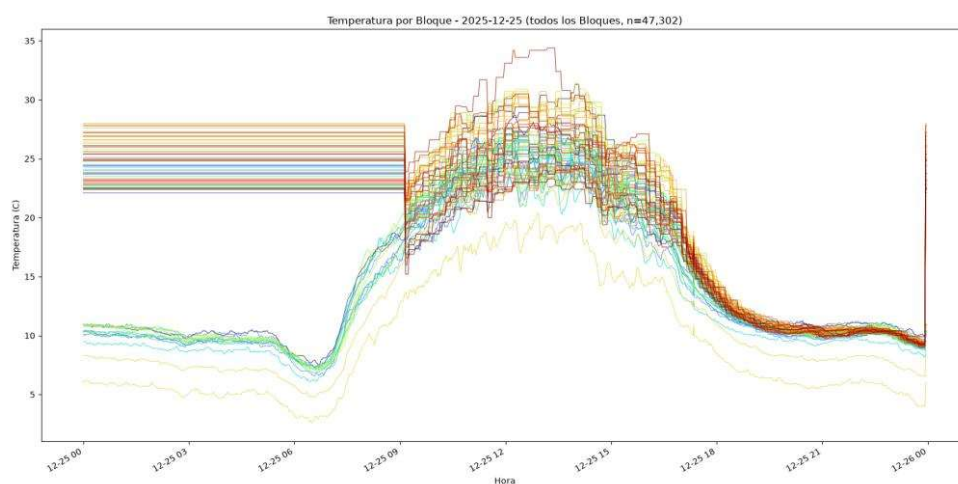


Figura 11 Temperatura de todos los bloques en diciembre del 2025

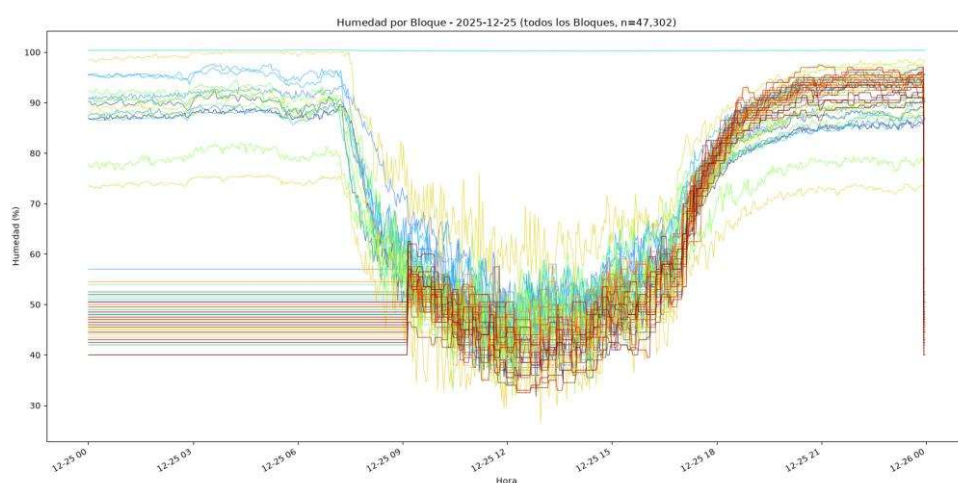


Figura 12 Humedad de todos los bloques en diciembre del 2025

El tratamiento estándar frente a este escenario es omitir los registros afectados en los cálculos posteriores (Integrated Ocean Observing System, 2014), por lo tanto, los removemos de los conjuntos de datos ISSI que serán usados para el entrenamiento de los modelos. En la Figura 13 se muestra un ejemplo de los valores de temperatura que serán usados como entrada para la siguiente fase del estudio; el mismo procedimiento es aplicado para los valores de humedad.

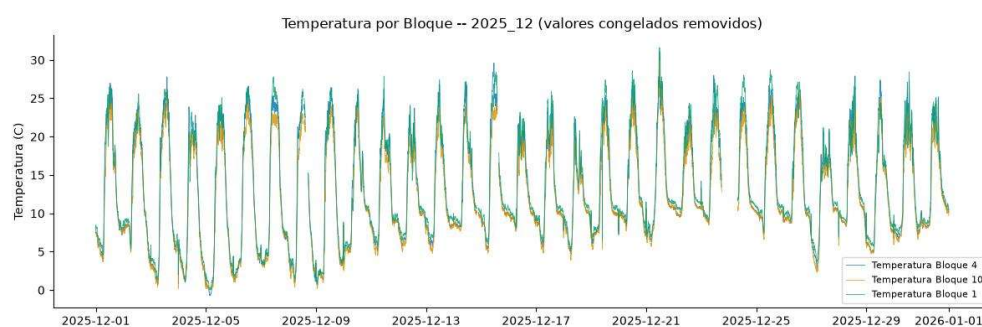


Figura 13 Valores “congelados” de temperatura removidos. Diciembre del 2025.

Dado que los archivos de entrada son anchos, se procedió a transformar los archivos a un formato largo, con las siguientes columnas:

- Date
- Time
- Datetime
- Bloque
- Metric
- Value

Y se establece el formato de salida al tipo de archivo *parquet* para soportar el incremento en el número de filas y mejorar el nivel de compresión.

Dado que se tienen columnas que representan lo mismo, pero con nombres distintos, se procede a estandarizar los nombres tal como se presentó en la Tabla 4.

El pipeline de limpieza de datos ISSI

Teniendo en cuenta las características de los datos previamente descritas, se construyó un pipeline de procesamiento, su ejecución en este estudio se hace manualmente siguiendo la secuencia ilustrada en la Figura 15.

Para facilitar la comprensión, se usa color gris para los archivos de entrada, anaranjado para los de salida y azul para los scripts de Python desarrollados para la limpieza.

El pipeline empieza con los datos de los archivos .csv crudos de entrada, estos son procesados por el script `extraer_headers_hinput.py` para generar una tabla larga con las columnas, el número de meses en que aparecen y el primer y último en donde están disponibles, esto se ilustra en la captura de la tabla generada mostrada en la **¡Error! No se encuentra el origen de la referencia..**

	Original column	N_meses	Primer_mes	Ultimo_mes
0	Date	21	2024_09	2026_05
1	Time	21	2024_09	2026_05
2	Filtro P06 IN	21	2024_09	2026_05
3	Flujo Bomba P05 - Silo 1	21	2024_09	2026_05
4	Flujo Bomba P06 - Silo 2	21	2024_09	2026_05
5	Flujo Bomba P07 - Silo 3	21	2024_09	2026_05
6	EC Monitor Bomba P05 - Silo 1	21	2024_09	2026_05
7	EC Monitor Bomba P06 - Silo 2	21	2024_09	2026_05
8	EC Monitor Bomba P07 - Silo 3	21	2024_09	2026_05
9	Filtro P07 OU	21	2024_09	2026_05

Figura 14 Tabla salida de `extraer_header_hinputs.py`

A partir de esa tabla, el script `extraer_bloque_metrica.py` procesa las columnas con métricas climáticas y extraer la información del bloque al que pertenece, el resultado se muestra en la captura de la tabla mostrada en la Figura 16.

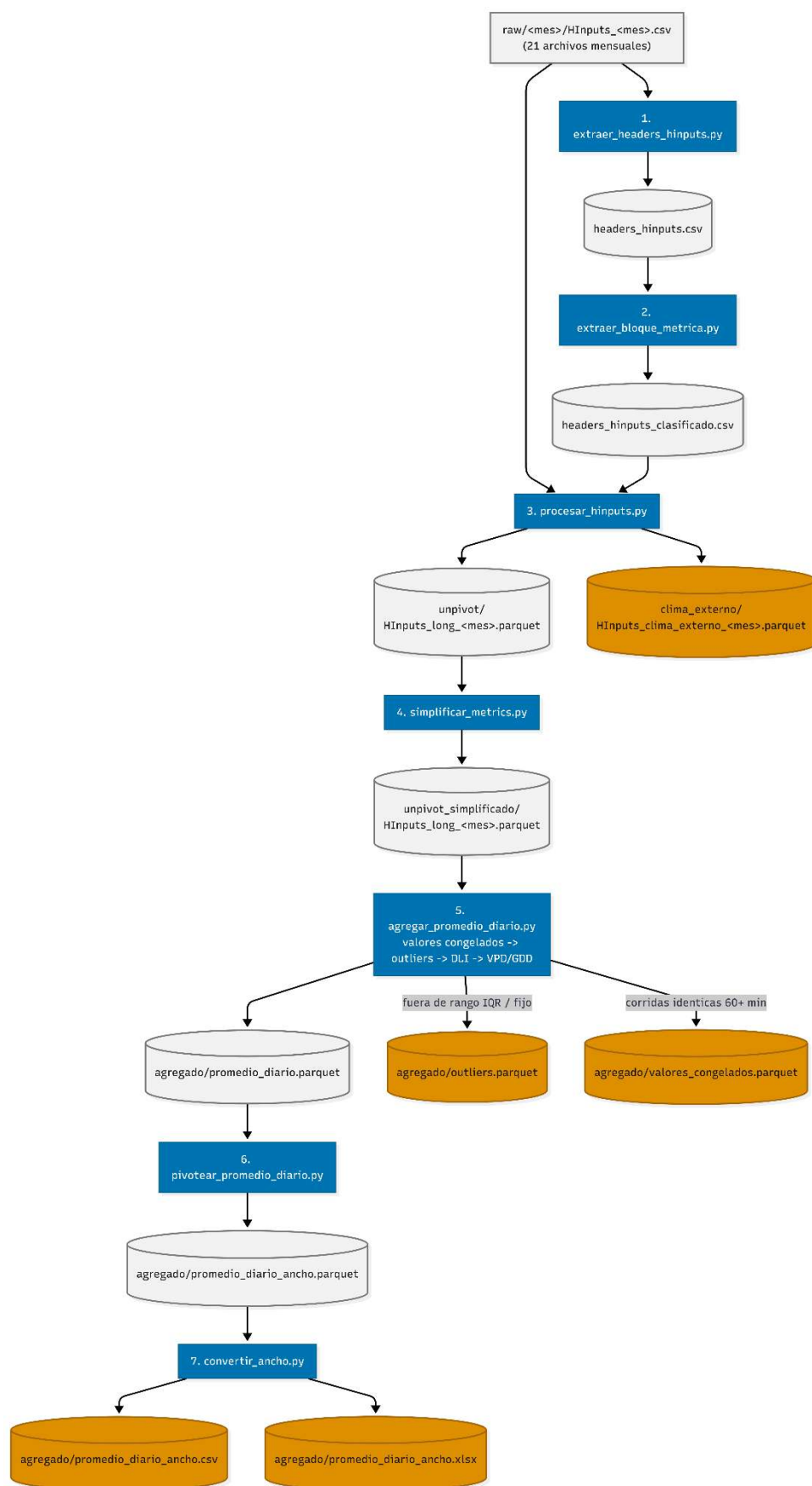


Figura 15 Pipeline de procesamiento de archivos de ISSI

	Original column	N_meses	Primer_mes	Ultimo_mes	Bloque	Métrica
56	Temperatura Bloque 4	21	2024_09	2026_05	4	Temperatura
57	Humedad Bloque 4	21	2024_09	2026_05	4	Humedad
58	Temperatura Bloque 10	21	2024_09	2026_05	10	Temperatura
59	Humedad Bloque 10	21	2024_09	2026_05	10	Humedad
60	Temperatura Bloque 1	21	2024_09	2026_05	1	Temperatura
61	Humedad Bloque 1	21	2024_09	2026_05	1	Humedad
62	Temperatura Bloque 15	21	2024_09	2026_05	15	Temperatura
63	Humedad Bloque 15	21	2024_09	2026_05	15	Humedad
64	Temperatura Bloque 11	21	2024_09	2026_05	11	Temperatura
65	Humedad Bloque 11	21	2024_09	2026_05	11	Humedad

Figura 16 Tabla de salida de `extraer_bloque_metrica.py`

Los dos scripts descritos hasta este punto permiten generar el archivo de entrada al script `procesar_hinputs.py` del que obtiene la información de mapeo de cada columna a su bloque y métrica correspondiente para generar dos salidas, un archivo con los datos “despivotados”, es decir una tabla larga con los datos de día, bloque, métrica climática y valor, esta primera tabla corresponde a los datos intra-invernadero (ver Figura 17). También genera una tabla con los datos correspondientes al clima externo (ver Figura 18).

	Date	Time	Datetime	Bloque	Metric	Value
0	260101	00:00:00	2026-01-01 00:02:00	4	Temperatura	10.1
1	260101	00:00:00	2026-01-01 00:02:00	4	Humedad	87.5
2	260101	00:00:00	2026-01-01 00:02:00	10	Temperatura	9.8
3	260101	00:00:00	2026-01-01 00:02:00	10	Humedad	84.9
4	260101	00:00:00	2026-01-01 00:02:00	1	Temperatura	10.4
5	260101	00:00:00	2026-01-01 00:02:00	1	Humedad	83.6
6	260101	00:00:00	2026-01-01 00:02:00	15	Temperatura	10.1
7	260101	00:00:00	2026-01-01 00:02:00	15	Humedad	94.2
8	260101	00:00:00	2026-01-01 00:02:00	11	Temperatura	10.4
9	260101	00:00:00	2026-01-01 00:02:00	11	Humedad	89.5

Figura 17 Tabla de salida de `procesar_hinputs.py` con datos intra-invernadero

	Date	Time	Datetime	Temperatura Externa (°C)	Radiacion externa (W.m2)	Humedad Relativa Externa (%)	
0	260101	00:00:00	00:02:00	2026-01-01 00:02:00	9.2	0.0	99.8
1	260101	00:00:00	00:04:00	2026-01-01 00:04:00	9.1	0.0	99.8
2	260101	00:00:00	00:06:02	2026-01-01 00:06:02	9.0	0.0	99.8
3	260101	00:00:00	00:08:03	2026-01-01 00:08:03	9.0	-1.1	99.8
4	260101	00:00:00	00:10:04	2026-01-01 00:10:04	9.0	-1.1	99.8
5	260101	00:00:00	00:12:03	2026-01-01 00:12:03	9.0	-1.1	99.8
6	260101	00:00:00	00:14:06	2026-01-01 00:14:06	8.9	0.0	99.8
7	260101	00:00:00	00:16:07	2026-01-01 00:16:07	8.9	0.0	99.8
8	260101	00:00:00	00:18:08	2026-01-01 00:18:08	8.9	-1.1	99.8
9	260101	00:00:00	00:20:08	2026-01-01 00:20:08	8.9	-1.1	99.8

Figura 18 Tabla de salida de procesar_hinputs.py con datos extra-invernadero

El script simplificar_metrics.py toma unifica los valores de la columna *Metrics* de la Figura 17 según lo definido en la Tabla 4.

El script agregar_promedio_diario.py filtra los atípicos y los congelados según lo especificado previamente y genera una nueva tabla con valores diarios promediados según se muestra en la Figura 19, también se incluyen nuevas métricas complejas calculadas a partir de la temperatura, humedad y tiempo: *DLI ePAR (mol/m2/dia)*, *GDD (C-dia)* y *VPD (KPa)*.

	Bloque	Metric	Fecha	Value
0	4	Temperatura	2024-09-06	19.386984
1	10	Temperatura	2024-09-06	20.211746
2	1	Temperatura	2024-09-06	13.692540
3	4	Humedad	2024-09-06	39.430511
4	10	Humedad	2024-09-06	41.358307
5	15	Temperatura	2024-09-06	13.538548
6	11	Temperatura	2024-09-06	13.827967
7	15	Humedad	2024-09-06	77.697697
8	11	Humedad	2024-09-06	67.608972
9	1	Humedad	2024-09-06	54.350526

Figura 19 Tabla agregada con valores diarios válidos, salida de agregar_promedio_diario.py

La tabla de la Figura 19 ya permite volver a transponer la información para que tenga el formato de una tabla de características para el entrenamiento de modelos, esto se logra pivoteándola con referencia al bloque y a la fecha. Para realizarlo se usa el script pivotear_promedio_diario.py, la tabla resultante se muestra en la Figura 20.

	Fecha	Bloque	Temperatura	Humedad	DLI ePAR (mol/m2/día)	VPD (KPa)	GDD (C-día)
0	2024-09-06	4	19.386984	39.430511	NaN	1.363424	13.386984
1	2024-09-06	10	20.211746	41.358307	NaN	1.389284	14.211746
2	2024-09-06	1	13.692540	54.350526	NaN	0.715321	7.692540
3	2024-09-06	15	13.538548	77.697697	NaN	0.345989	7.538548
4	2024-09-06	11	13.827967	67.608972	NaN	0.512052	7.827967
5	2024-09-06	24	0.000000	NaN	NaN	NaN	NaN
6	2024-09-06	16	5.519149	48.754545	NaN	0.463480	0.000000
7	2024-09-06	20	4.826667	41.600000	NaN	0.503299	0.000000
8	2024-09-06	18	4.411905	43.007692	NaN	0.477107	0.000000
9	2024-09-06	29	8.462198	80.170889	NaN	0.219510	2.462198

Figura 20 Tabla de salida del script `pivotear_promedio_diario.py`

Finalmente, el script `convertir_ancho.py` genera dos tablas idénticas a la mostrada en la Figura 20 que se encuentra en formato parquet, pero esta vez en formatos .csv y .xlsx.

3.1.3. Alineación Temporal e Integración de las Fuentes

La integración de las fuentes exigió resolver la diferencia de granularidad temporal, esto es semanal en el pinch y diaria en la cosecha mediante la desagregación del pinch a nivel diario. El procedimiento se sustentó en tres decisiones.

En primer lugar, se estableció la correspondencia temporal entre fuentes. Se verificó que la semana registrada en el pinch coincidiera con la semana ISO estándar (de lunes a domingo). El primer día con cosecha, el 28 de agosto de 2023, es un lunes y corresponde a la semana ISO 35 de 2023, exactamente el punto de inicio de la base de pinch. Adicionalmente, la totalidad de las 139 semanas con registros de pinch tiene cosecha en la misma semana ISO, y el año de pinch se comporta como año ISO, sin discrepancias en los cambios de año. Con ello se confirmó que las claves de unión son el año ISO, la semana ISO, el bloque y la variedad; el bloque se normalizó de valor numérico a texto y la variedad se trató como texto para asegurar la coincidencia exacta.

En segundo lugar, se definió el método de desagregación. Los tallos pinchados de cada combinación de bloque y variedad en una semana determinada se repartieron de manera

uniforme entre los días en que esa misma combinación estuvo en cosecha durante dicha semana. Para preservar la naturaleza entera del recuento de tallos, la asignación se realizó mediante el método del mayor resto, es decir, a cada día se le asignó la parte entera del cociente y el residuo se distribuyó de a una unidad entre los primeros días de la semana en orden cronológico. De este modo se garantiza que la suma de los valores diarios reconstruya de forma exacta el total semanal original, que todos los valores sean enteros y que no se introduzcan fracciones artificiales de tallos.

Finalmente, en tercer lugar, se resolvió el tratamiento de los casos sin correspondencia. Ciento dos combinaciones de bloque, variedad y semana que equivalen al 0,6 % de las combinaciones de pinch y a 85.970 tallos, presentaban registro de pinch pero que al mismo tiempo no tenían ningún día de cosecha en la semana ISO respectiva. Al carecer de días sobre los cuales efectuar el reparto, estas combinaciones se descartaron. Las combinaciones de cosecha sin pinch asociado, en cambio, se conservaron con un valor de tallos pinchados igual a cero.

3.1.4. Base resultante

La base curada, compuesta de los archivos de entrada previamente descritos, se construye haciendo una operación *left join* como se resume en la Tabla 6. Al ser la tabla de clima Davis el ancla de la operación, se procede a rellenar las filas faltantes cuando las condiciones son las adecuadas.

Tabla 6
Resumen de las fuentes de entrada y su rol en la base unificada

Orden	Archivo	Granularidad	Rol en la unión
1	Metricas_Diarias_Davis.csv	Una fila por fecha	Ancla
2	Base_Diaria_Cosecha_Pinch_Corregida.csv	Una fila por cada combinación de fecha, bloque y variedad	<i>Left join</i> por fecha
3	promedio_diario_ancho.csv (ISSI)	Una fila por cada fecha y bloque	<i>Left join</i> por fecha y bloque sobre el resultado anterior

Al ser la fuente de datos de clima Davis el ancla, se logra que no se pierda ninguna fecha del historial meteorológico, el efecto es que los días de cosecha o pinch queda con valores nulos y las combinaciones de fecha y bloque que no tienen correspondencia con alguna lectura de los datos de clima ISSI quedan con valores de clima nulos.

Para los valores ISSI de clima nulos, se aplican las siguientes consideraciones:

1. Calcular el promedio de la métrica entre los demás bloques que sí tienen datos ese día. De esta forma se toma como referencia el valor de la métrica en los otros invernaderos.
2. Si existe el promedio planteado en el numeral anterior, se rellena el valor faltante con él, caso contrario, el valor de la columna quedará como estaba, en nulo.
3. Para mantener la trazabilidad de los valores rellenos, se marca cada fila rellanada con el valor *True* en la columna *Clima_ISSI_Rellenado*.
4. Se recalculan los valores para las métricas complejas *VPD (KPa)* y *GDD (C-día)* a partir de los valores de temperatura y humedad rellenos, el cálculo se realiza de la misma forma que antes, usando la fórmula de Magnus/Tetens y el mismo *GDD_TEMP_BASE* de 6 grados que se usa para el cálculo de los registros no rellenos. Se evita asignar directamente el promedio de *VPD (KPa)* y *GDD (C-día)* para evitar la mezcla de dos bases de datos distintas. Las filas que ya tenían valores de temperatura y humedad desde el principio conservan los valores de *VPD* y *GDD*.

El resumen de las filas nulas antes y después de aplicar el algoritmo de relleno se detalla en la *Tabla 7*.

Tabla 7
Nulos por columna, antes vs. después del relleno

Columna	Nulos antes	Nulos después	Filas rellenadas
Temperatura	101140	61599	39541
Humedad	101348	61599	39749
DLI ePAR ($\text{mol}/\text{m}^2/\text{dia}$)	177951	115827	62124
VPD (KPa)	101688	61599	40089
GDD (C-día)	1016 88	61599	40089

Las columnas Temperatura, Humedad, VPD y GDD terminan con el mismo número de nulos porque, en la práctica, cuando falta una de las dos primeras casi siempre falta la otra también ese mismo bloque y día, y el recálculo de VPD/GDD depende de tener ambas disponibles. Por su parte, los valores de DLI ePAR sigue teniendo muchos más nulos que las demás, es una métrica que muy pocos bloques reportan en general, así que en la mayoría de los días ningún bloque tiene lectura de la que promediar, y el algoritmo de relleno no establece a los valores de ese escenario como rellenables.

Es importante resaltar que el algoritmo de relleno no inventa datos de los que no se disponible ninguna fuente y que los promedios se calculan sobre todos los bloques con datos disponibles en el día en cuestión.

Para verificar que no existan valores evidentemente corruptos remanentes de la unión de las fuentes de datos, se realiza un chequeo rápido con reglas basadas en la plausibilidad física, el resumen de las validaciones se detalla en la *Tabla 8*. Adicionalmente observamos algunas distribuciones y conteos generales para verificar la coherencia de los datos de la base unificada, de esta forma en la Figura 21 se muestran los veinte bloques y veinte variedades con más registros, la diferencia presentada está dentro de los valores esperados.

Tabla 8
Chequeo rápido de calidad de datos en de base unificada

Regla	Filas que incumplen
Tallos Cosechados < 0	0
Tallos Pinchados < 0	0
Humedad fuera de [0, 100]	0
Temperatura fuera de [-5, 45] C	0
VPD (KPa) < 0	0
GDD (C-día) < 0	0

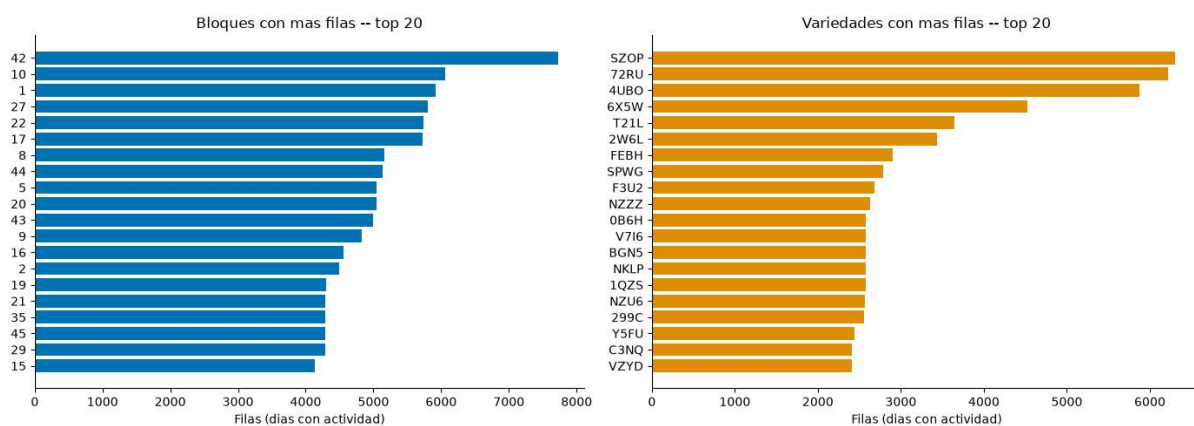


Figura 21 Top 20 de bloques y variedades con más filas

También se verifica que se tenga información de todos los días, las caídas periódicas corresponden a las fechas en que no hubo cosechas, típicamente la florícola no realiza cosechas los domingos por lo que el comportamiento observado se adapta al esperado (ver Figura 22).

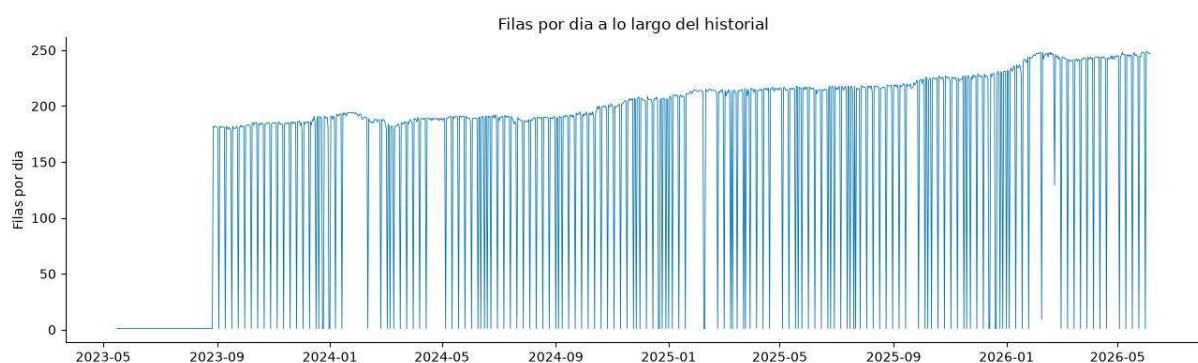


Figura 22 Cantidad de registros por días en la base unificada

Adicionalmente se visualiza en la Figura 23 la serie temporal de la variable objetivo, los tallos cosechados, no observan valores anómalos, la única observación es una cosecha alta los primeros meses del año.

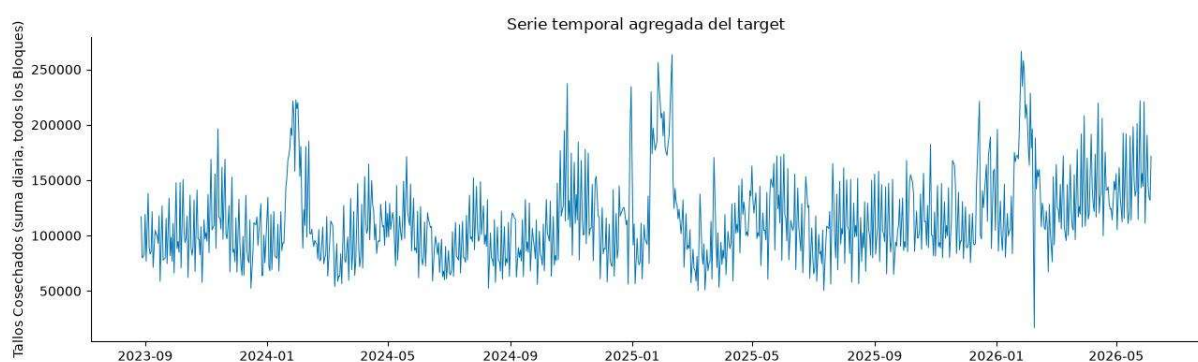


Figura 23 Cantidad total de tallos cosechados por día en la base unificada

3.2. Entrenamiento de Modelos

3.2.1. Diseño y Criterios de Éxito

Para cumplir con los objetivos propuestos, este estudio compara ocho variantes de modelo para pronosticar la cantidad de tallos de rosa cosechados por cada combinación de bloque, variedad de rosa y semana, y las enfrenta contra la proyección humana que ya usa la empresa.

Para estructurar la experimentación, las ocho variantes de los modelos forman una tabla de ablación controlada de tal forma de que cada una de las variantes cambia un eje de diseño a

la vez respecto a las demás, para poder atribuir cualquier diferencia de resultado a ese eje específico y no a una mezcla de cambios simultáneos.

3.2.1.1. Los Ejes de Diseño

La Tabla 9 resume las definiciones de los ejes de diseño experimental planteados para este estudio.

Tabla 9
Ejes de diseño

Eje	Valores posibles	Descripción
Modelo	1. HistGBT 2. CatBoost 3. LSTM 4. TCN	Valores 1 y 2 de la familia de árboles de gradient boosting y valores 3 y 4 de la familia de redes neuronales
Granularidad del <i>target</i>	1. Semanal nativo 2. Diario	Para el valor 1, el modelo predice directamente el total de la semana. En el caso del valor 2, el modelo predice día a día y se suma después, en otras palabras, un agregado semanal.
Rezago biológico explícito	1. Sí 2. No	El valor 1 corresponde al escenario en que el modelo recibe, como predictor explícito, la ventana de días entre el <i>pinch</i> y la cosecha que resulta de él, estimada por variedad. El valor 2 es asignado al caso negado del 1.
Especialización	1. Global 2. Individual 3. Residual por variedad	Global se refiere a un solo modelo para todas las variedades, individual un modelo por cada variedad. La especialización residual por variedad es un modelo global con una corrección liviana por cada variedad pero sin reentrenar la red.

3.2.1.2. Las Ocho Variantes

En la Tabla 10 se detallan las ocho variantes de los modelos considerados y estudiados durante el desarrollo de este trabajo.

Tabla 10
Definición de variantes

Variante ID	Modelo	Granularidad del Target	Rezago biológico	Especialización
HISTGBT_GLOBAL	HistGBT	Semanal nativo	Sí	Global
HISTGBT_INDIV	HistGBT	Semanal nativo	Sí	Individual
CATBOOST_GLOBAL	CatBoost	Semanal nativo	Sí	Global
CATBOOST_INDIV	CatBoost	Semanal nativo	Sí	Individual
LSTM_GLOBAL	LSTM	Diario	No	Global
TCN_GLOBAL	TCN	Diario	No	Global
LSTM_RESID	LSTM	Diario	No	Residual por variedad
LSTM_RESID_BIO	LSTM	Diario	Sí	Residual por variedad

3.2.1.3. Preguntas que el Diseño Permite Responder

Para analizar los resultados de la aplicación de la tabla de ablación, definimos las siguientes preguntas:

1. ¿Qué familia de árboles de *gradient boosting* predice mejor, a igualdad de todo lo demás?
2. ¿Especializar un modelo por variedad ayuda, o el modelo global ya generaliza igual de bien?
3. ¿Cuál aprovecha mejor la secuencia diaria de historia calculada como hiperparámetro?
4. ¿Ayuda agregar una corrección residual por variedad sobre un *LSTM* global, sin pagar el costo de reentrenar la red por variedad?

5. ¿Aporta algo agregar el predictor biológico explícito a un modelo que ya ve toda la historia diaria?
6. ¿Puede algún modelo superar a la proyección humana que la empresa ya usa para planificar?

3.2.1.4. Criterios de Éxito de la Experimentación con los Modelos

Para determinar el éxito de la aplicación del diseño experimental, se definen los siguientes cuatro criterios:

1. Los resultados permiten responder las preguntas planteadas en la sección anterior a esta con evidencia en los datos disponibles.
2. Determinar si alguna de las ocho variantes de modelo supera a la proyección humana en la comparación semanal directa, y en qué medida.
3. Entender la razón del por qué gana el mejor modelo más allá de solamente reportar su métrica, sino que explicar qué variables usa para poder defender la recomendación de despliegue.
4. Producir la recomendación de despliegue en producción con evidencia robusta e identificando sus límites explícitos.

Estos criterios permiten alcanzar dos de los objetivos específicos definidos en el capítulo uno.

3.2.2. Comprensión de los Datos

3.2.2.1. Carga de Datos

Identificamos dos tipos de fuentes de datos de entrada:

1. Base unificada de cosecha, *pinchs* y clima. Esta fuente de datos es una serie diaria de la combinación de bloque y variedad, incluye la información de tallos

cosechados, pinchados y datos climáticos de dos fuentes distintas, ISSI y Davis, es decir, intra y extra-invernadero respectivamente.

2. Proyección humana, es un archivo Excel que detalla la cantidad de tallos que el equipo de producción proyecta cosechar cada semana por bloque y variedad, además contiene la producción real correspondiente. Es la fuente de la variable a predecir y de si fuente de verdad de referencia contra la que se validan los modelos estudiados y sus variantes. Este archivo fecha sus filas usando los lunes como referencia de cada semana para la que se realiza la proyección. Se omiten 3353 filas que no poseen un valor de proyección.

En la Figura 24 se muestra la distribución de la cantidad de registros por fecha para cada fuente de datos ya descrita, la base de unificada de cosecha y clima, al ser diaria, tiene un orden de magnitud superior que la de la proyección humana que es semanal.

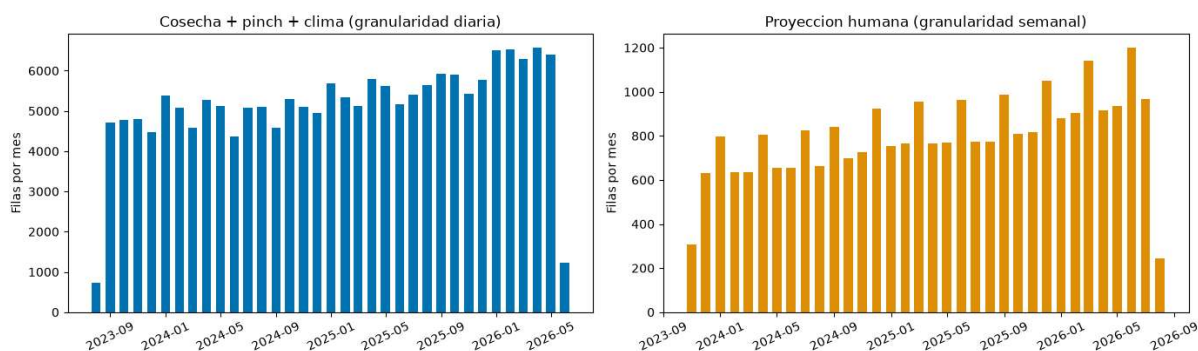


Figura 24 Cobertura temporal de las dos fuentes de datos de entrada

Otro aspecto importante de observar es que la cantidad de datos de las proyecciones realizadas por los trabajadores ha venido incrementándose con el tiempo, esto se puede ver en la Figura 25.

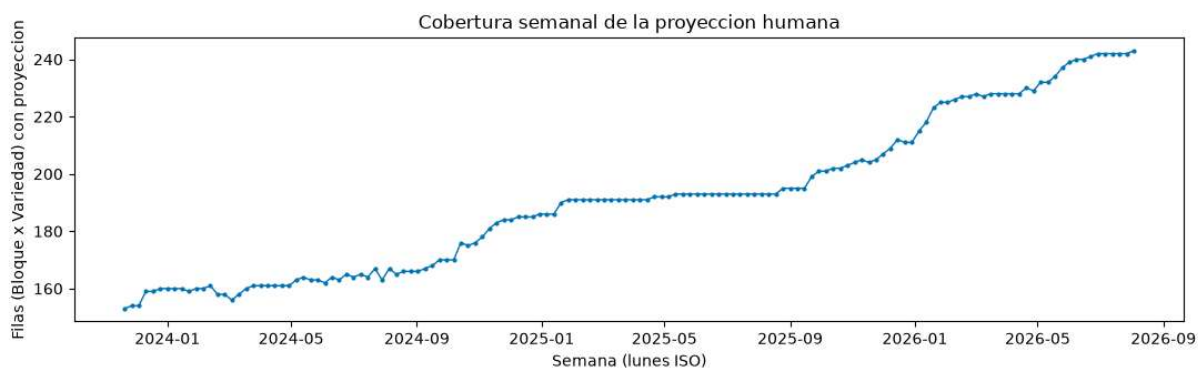


Figura 25 Cobertura semanal de datos de proyección humana

3.2.2.2. Distribución del Target y Disponibilidad de Clima

La distribución del *target* tiene una cola larga a la derecha (ver Figura 26), es decir, los días de cosecha alta son poco frecuentes, eso es típico de datos en contextos de conteo agrícola.

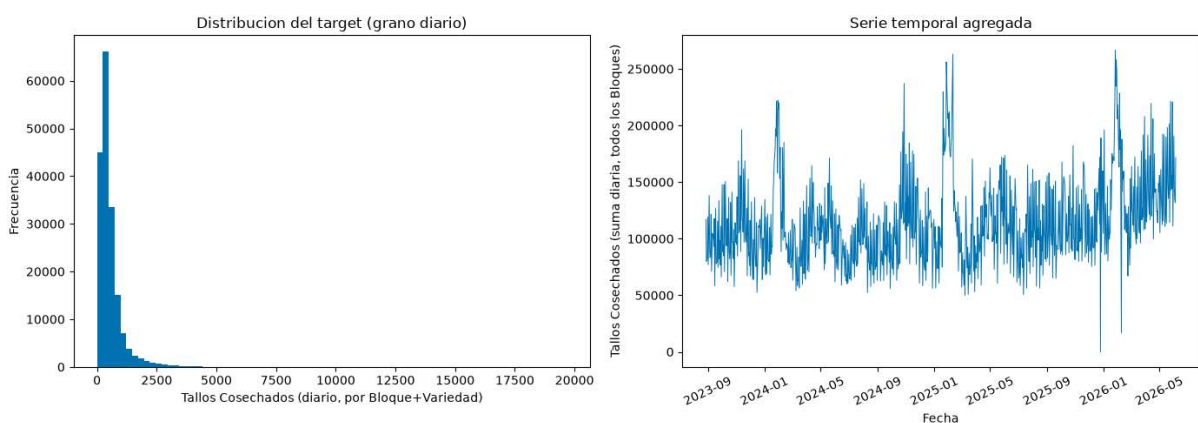


Figura 26 Distribución de target y serie temporal de tallos cosechados

Respecto a los datos de clima dentro de los invernaderos solo se tiene cobertura consistente a partir de septiembre del 2024 (ver Figura 27), recién cerca a la mitad de la información histórica. Para los datos faltantes, se completa con técnicas de relleno explicadas más adelante.

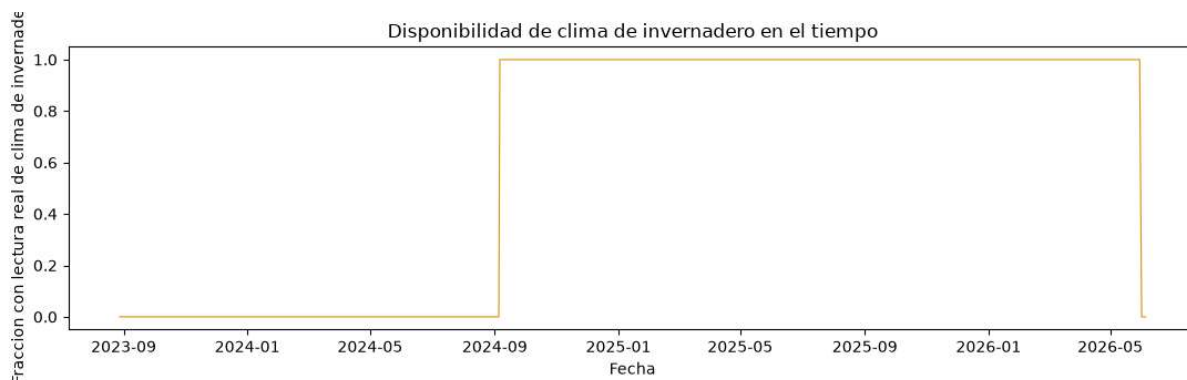


Figura 27 Disponibilidad de datos climáticos por invernadero

3.2.3. Preparación de Datos

3.2.3.1. Rezago y Tallos Pinchados

En series temporales, una forma sencilla de predecir una variable a futuro, por ejemplo, en nuestro caso “cuánto se va a cosechar esta semana”, es mirar el comportamiento pasado de la variable a predecir, en términos técnicos, haciendo una autorregresión pura de la propia variable a predecir o *target*. Sin embargo, a pesar de que un modelo puede lograr una buena predicción apoyándose casi exclusivamente en la inercia del *target*, sin aprender ninguna otra relación agronómica real, esto es un atajo frágil ya que el modelo puede perder efectividad drásticamente si no se tiene un historial reciente de cosecha o cambia el patrón de negocio.

Por lo expuesto previamente, se usará un rezago basado en los tallos pinchados, esta variable cuantifica una práctica que precede biológicamente a la cosecha, es decir, pinchar hoy determina cuánto se cosechará semanas después. Bajo este concepto construimos un predictor biológico explícito sobre los tallos pinchados y no sobre los tallos cosechados. En secciones posteriores se profundiza sobre la configuración del predictor biológico.

3.2.3.2. Análisis de Correlación Cruzada (CCF)

Una vez identificada la relevancia biológica del *pinch*, surge la interrogante sobre cuánto tiempo transcurre entre que se pincha una planta y que se da la cosecha. El conocimiento

agronómico señala que esto difiere dependiendo de la variedad de la planta. Para estimar el valor basados en evidencia y no asumir un único valor para todas las variedades, realizamos el cálculo de la función de correlación cruzada o CCF entre las series diarias de tallos pinchados y cosechados de cada combinación de bloque y variedad.

Por lo tanto, la función de correlación aplicada se expresa a continuación:

$$CCF(k) = \text{correlación}(\text{Pinchados}_t, \text{Cosechados}_{\{t+k\}})$$

Con k representando qué tan parecida es la serie de pinch de hoy a la cosecha k días después y k estando entre 0 y 150 días. El pico de esa curva, es decir, el k con mayor correlación, estima el tiempo de maduración biológico típico de esa variedad. Para determinar la ventana biológica, se promedia la curva entre todos los bloques de la misma variedad, y si el pico tiene una correlación fuerte (mayor a 0.25) se toma como ventana $[a, b]$ al rango de días alrededor del pico donde la CCF se mantiene sobre el 80% de su valor máximo. Para graficarlo, ver la Figura 28 en donde las curvas de correlación alcanzan su pico entre los 85 y 110 días, esta misma ventaja es usada como una ventana genérica para las variedades sin un pico confiable.

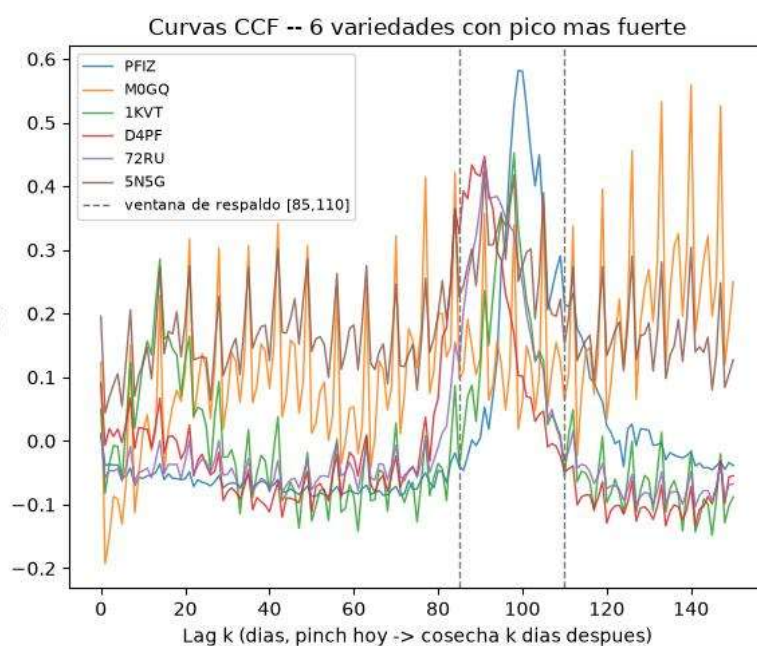


Figura 28 Curvas CCF de las 6 variedades con la correlación más fuerte

Como se muestra en la Figura 29, la mediana del lag o rezago del pico de todas las variedades es de 98 días, esto representa el valor que queda justo en el centro cuando organizamos los datos de menor a mayor, si bien este valor no es usado de forma explícita, sirve como valor referencial. Otro dato referencia es el que se puede inferir al analizar la Figura 30 en la que vemos que, entre las 37 variedades con correlación fuerte, el ancho de la ventana biológica más recurrente es de entre 5 y 10 días.

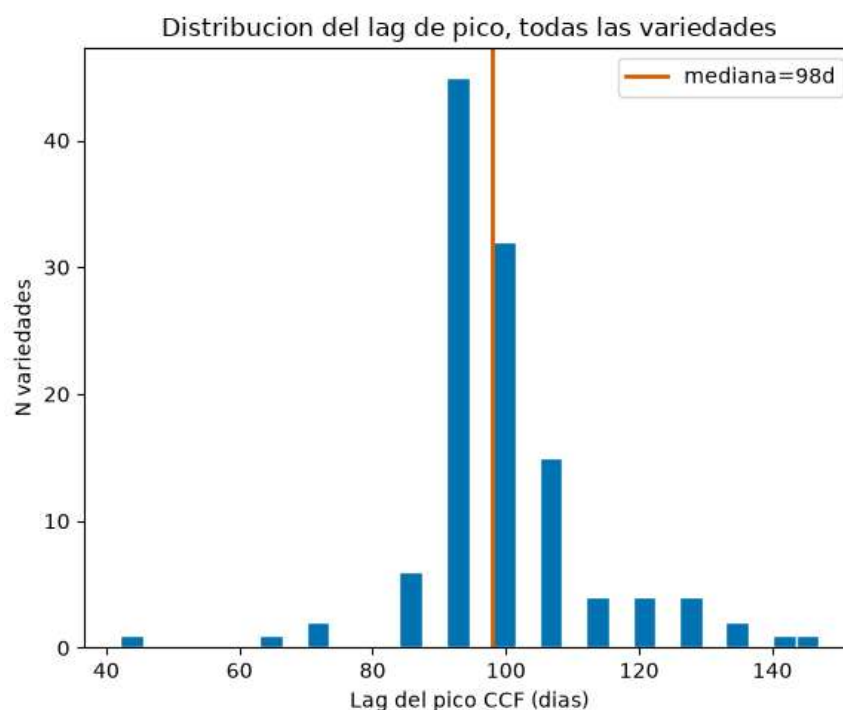


Figura 29 Distribución del lag de pico de todas las variedades

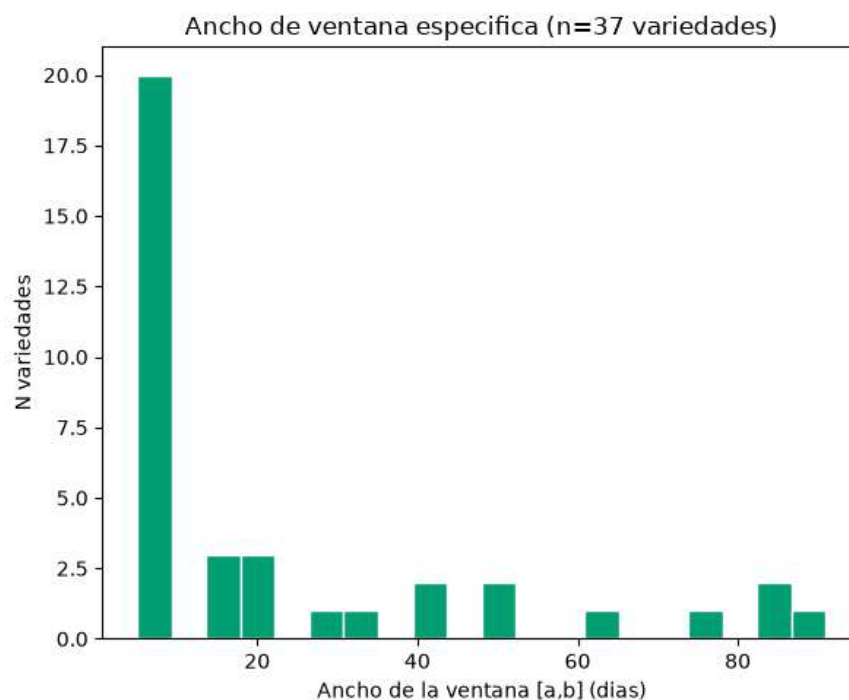


Figura 30 Distribución del ancho de las ventanas biológicas de las variedades con correlación fuerte

3.2.3.3. Imputación de Clima y Series Diarias por Bloque y Variedad

Dado que los datos disponibles de clima intra-invernadero tienen vacíos, en especial antes de los últimos meses del 2024, se estableció una estrategia para completarlos. En primer lugar, se completa con el valor más cercano dentro del mismo invernadero, ya sea *ffill* o *bbill*, si esto falla se utiliza la mediana global de esa variable.

Adicionalmente se construye un nuevo *dataframe* sin datos vacíos por cada combinación de bloque y variedad, este nuevo *dataframe* es el usado para agregar los datos dentro de la ventana biológica definida por el CCF y para calcular el predictor biológico en cualquier fecha.

3.2.3.4. El Predictor Biológico

Definimos al predictor biológico como *FEAT_BIO*, corresponde a la suma de tallos pinchados y tallos cosechados dentro de la ventana biológica (CCF) de la variedad correspondiente y que termina antes del ancho, esto hace que la ventana se produzca por completo en el pasado a la fecha a predecir asegurando que se evite la fuga de información.

Para robustecer la capacidad de que los modelos aprendan un patrón real, solo se modelan las variedades que tengan historia suficiente, esto es, al menos 250 días con datos y al menos 8000 tallos cosechados en total. La cantidad de variedades que cumple con esta definición son 118 de 141, es decir el 84%.

Para ilustrarlo, podemos analizar la Figura 31 en donde:

- La línea vertical punteada es el ancla, es decir, el punto en el tiempo desde el que se mira hacia el pasado para predecir el valor objetivo.
- La franja verde corresponde a la ventana CCF, es decir, el rango $[a, b]$ de días antes del ancla que se suma para obtener el $FEAT_BIO$
- La franja naranja o semana objetivo de predicción cuya suma de tallos cosechados es el valor de la predicción del modelo.

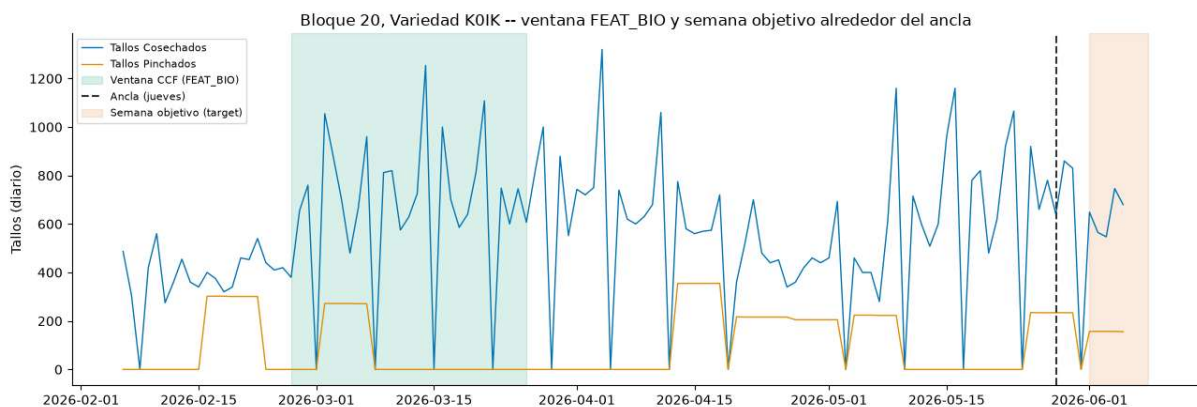


Figura 31 Panel semanal del bloque 20 y la variedad K0IK

3.2.3.5. Validación Temporal

Dado que no podemos validar un modelo de series de tiempo mezclando fechas al azar entre los conjuntos de datos de entrenamiento y prueba, adoptamos el método de avance paso a paso o *walk-forward*. El método consiste en que cada *fold* de validación entrena con todo lo

anterior y prueba con el bloque de semanas siguiente (ver Figura 34). Además, se deja un embargo de una semana entre el fin del subconjunto de entrenamiento y el inicio del subconjunto prueba con el objetivo de evitar que una *feature* con ventana móvil cercana a los 110 días, como *FEAT_BIO*, toque accidentalmente información que cruza el borde del *fold*. Las métricas finales de cada variante se calculan sobre la concatenación de todos los *folds* de prueba, no sobre un único recorte, de esta forma se logra una muestra mucho más amplia y una estimación más confiable del error real que validar una sola vez.

En la Figura 32 se ilustra el *walk-forward* con embargo, se muestran 11 *folds*, el segmento claro de cada *fold* es el subconjunto de entrenamiento seguido de una sección sin color que corresponde al embargo y finalmente un segmento sólido que corresponde al subconjunto de prueba.

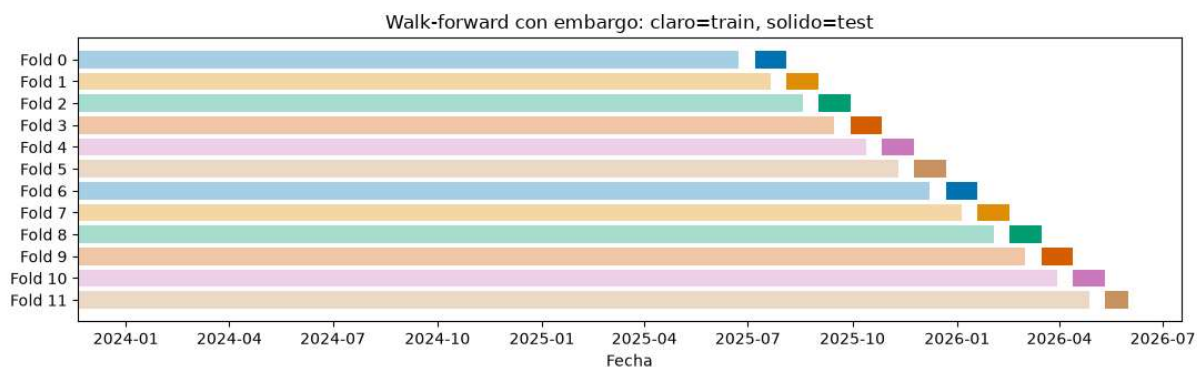


Figura 32 Gráfico de walk-forward con embargo

En la Figura 33 se presenta un acercamiento sobre el primer fold, el fold 0, aquí se ven las semanas de entrenamiento en azul más cercanas al corte, el embargo que queda excluido de los conjuntos entrenamiento y pruebas, y las primeras semanas del conjunto de pruebas para poder visualizar el agujero del embargo explícitamente en vez de inferirlo del grafico general de arriba.



Figura 33 Representación de fold 0

De la Figura 34 es relevante notar las semanas de pruebas se mantienen de tamaño constante y que el de las semanas de entrenamiento va creciendo conforme avanza el tiempo.

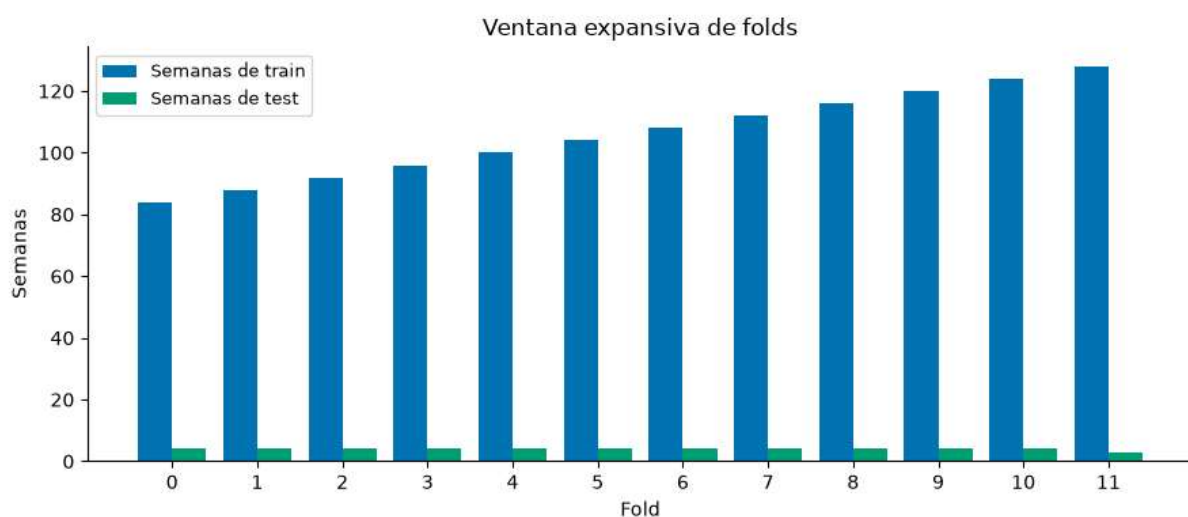


Figura 34 Ventana expansiva de folds

3.2.4. Modelado

En esta sección se entrenan las ocho variantes de modelo definidas en la sección 3.2.1. Los hiperparámetros de cada familia (HistGBT, CatBoost, LSTM, TCN) quedan fijos entre variantes de la misma familia y se buscan una sola vez por familia usando validación cruzada temporal (walk-forward) y esa misma configuración ganadora se aplica por igual a todas las variantes de esa familia (global e individual).

3.2.4.1. Variantes Tabulares Semanales

Las variantes de esta sección pertenecen a la familia *gradient boosting trees*, es decir, HistGBT y CatBoost en sus especializaciones global e individual.

Respecto la codificación de las variables categóricas, en lugar de expandir Bloque/Variedad en columnas binarias como se haría con *one-hot encoding*, que para un aproximado de 250 categorías combinadas dispararía la dimensionalidad, cada modelo usa su mecanismo nativo. HistGBT recibe un encoding ordinal que consiste en asignar un entero arbitrario por cada categoría junto con su parámetro *categorical_features* para que trate esas columnas como categorías, no como una escala numérica real. Por su lado, CatBoost usa su propio encoding interno basado en estadísticas del *target* calculadas de forma que evita fuga de información.

En lugar fijar los hiperparámetros a mano, se realiza una búsqueda con validación cruzada *walk-forward*, esto es, para cada combinación candidata se entrena y evalúa sobre varios *folds* temporales y se promedia el MAE de esos folds. La combinación con menor MAE promedio gana y se busca una sola vez por cada familia de modelo, en este caso HistGBT y CatBoost, esa misma configuración se usa después tanto para la variante global como para la individual de esa familia. De esta forma el eje de especialización se mantiene aislado del eje de hiperparámetros.

En la Figura 35 se muestran los resultados del MAE promedio por cada combinación hiperparámetros probada para el modelo HistGBT, dando como los mejores:

- `max_iter = 150`
- `learning_rate = 0.08`
- `max_depth = 8`
- `min_samples_leaf = 20`

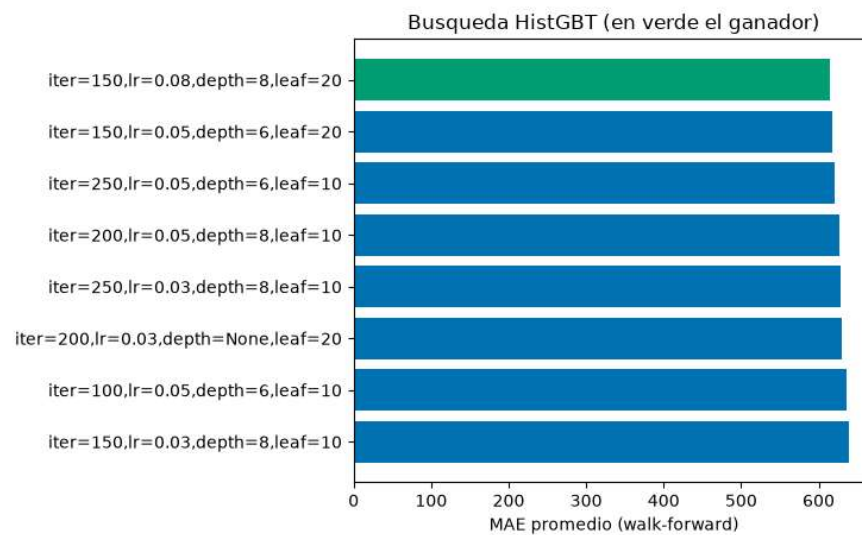


Figura 35 Búsqueda de hiperparámetros por validación cruzada temporal de HistGBT

Para verificar uno de los hiperparámetros ajustados del modelo HistGBT, se grafica en la Figura 36 el MAE por cada iteración en donde se observa la existencia de una asíntota y que después de las 150 iteraciones el MAE no continúa bajando.

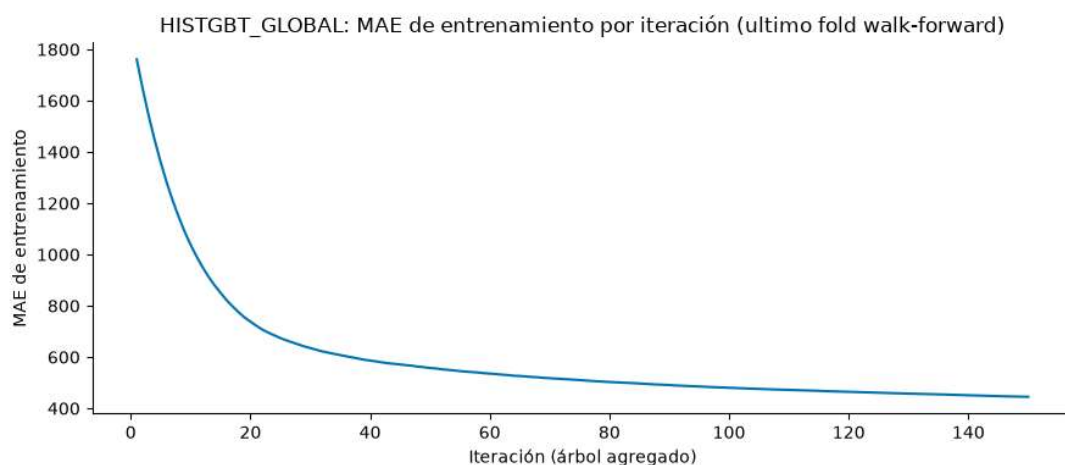


Figura 36 MAE de entrenamiento de HistGBT global

Por su parte, la búsqueda de hiperparámetros para el modelo CatBoost (ilustrado en la Figura 37) dio los siguientes como los mejores, considerando al MAE como la métrica de relevancia:

- iterations = 150
- learning_rate = 0.05
- depth = 6

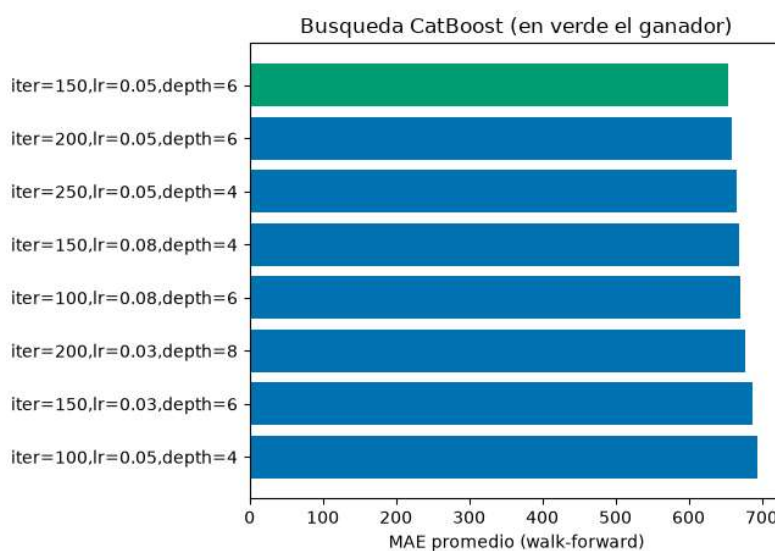


Figura 37 Búsqueda de hiperparámetros por validación cruzada temporal de CatBoost

De manera similar a como se verificó el hiperparámetro de iteraciones de para HistGBT, en la Figura 38 se muestra la curva de RMSE de entrenamiento por iteración del modelo CatBoost global en donde se hace evidente la misma asíntota ya identificada, el RMSE no bajará de manera significativa en iteraciones superiores a 150.

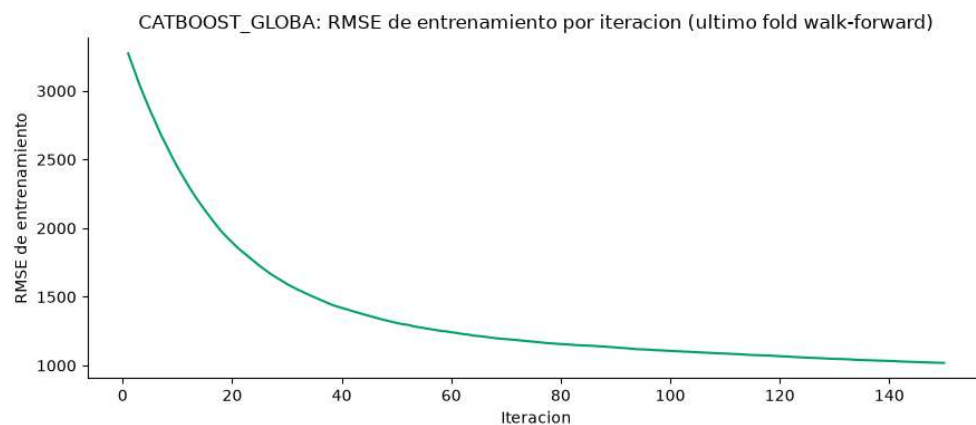


Figura 38 RMSE de entrenamiento de CatBoost global

En la especialización individual se entrena un modelo completo por cada variedad de flor, entrenado y evaluado solo con los datos de esa variedad teniendo como consideración que solo se toman en cuenta las variedades de flores que tienen datos en más de un invernadero o bloque. Adicionalmente no se definen ni pesos ni hiperparámetros compartidos entre las diferentes variedades. La validación *walk-forward* de cada variedad usa un umbral más relajado en comparación con el usado en la especialización global, esto es, 15 semanas para la variante individual del modelo y 20 semanas para la variante global del modelo, porque cada variedad individual de flor tiene menos semanas propias de historia.

Por presupuesto computacional, la especialización individual se acota a las 30 variedades de mayor volumen de cosecha, no las más de 250 elegibles, debido a que entrenar un modelo *walk-forward* completo por variedad es mucho más caro que un solo modelo global, y las

variedades de bajo volumen son además las que menos se benefician de un modelo dedicado ya que disponen de menos datos para poder realizar un buen ajuste.

3.2.4.2. Variantes Secuenciales Diarias

Estas cuatro variantes están basadas en redes neuronales y realizan la predicción con granularidad diaria en lugar de semana como se hace con la familia de *gradient boosting trees*, es decir, cada ejemplo de entrenamiento es una ventana de historia diaria consecutiva que combina los valores de cosecha diaria, clima y calendario por cada combinación de bloque y variedad, estas variantes son usadas para predecir el día siguiente basadas únicamente en los datos del pasado. Las predicciones diarias del período de prueba (subconjunto de prueba) se agregan semanalmente mediante una operación de adición, para poder compararse contra las variantes de *target* semanal nativo y contra la proyección humana. El corte de los conjuntos de entrenamiento y pruebas de cada *fold* cae en el mismo día, definido con la variable *ANCHOR_DOW*, que ya usan las variantes tabulares semanales para anclar sus *features* de ventana biológica y climática, así que las 8 variantes comparten exactamente el mismo punto de corte temporal, no dos convenciones distintas, esto resulta sumamente relevante al momento de realizar el análisis de resultados para asegurar la coherencia semántica de los mismos.

La covariable estática, la variable que no cambia con el paso del tiempo, es el *one-hot* de la combinación de bloque y variedad, es leída antes de reiniciar el índice del grupo, para evitar errores de lectura en que apunten a la fila equivocada del *one-hot*. La variante *LSTM_RESID_BIO* es la única de las cuatro que además agrega *FEAT_BIO* al vector estático, el resto de la arquitectura y el protocolo de entrenamiento es idéntico al de *LSTM_RESID*.

Al igual que con las variantes tabulares, cada familia basada en redes neuronales y basada en datos secuenciales diarios entrena una red nueva por cada *fold* de validación temporal, no un

único recorte cronológico, y las métricas finales son la concatenación de las predicciones en el conjunto de pruebas de todos los *fold*s. Reentrenar una red neuronal por *fold* es más caro que hacerlo en un modelo tabular, así que se usa un subconjunto de *fold*s, 3 en lugar de los 12 de las variantes semanales, elegidos con buen espaciado temporal; suficientes para que las métricas finales no dependan de un único período de pruebas, sin que el costo de cómputo se dispare. La longitud de esa ventana está definida en la variable *SEQUENCE_WINDOW* que tampoco se fija a mano pero que se busca entre varias candidatas, igual que se busca la arquitectura de cada red.

Tabla 11

Combinaciones de cada ventana candidata para SEQUENCE_WINDOW

SEQUENCE_WINDOW	# de secuencias	# de combinaciones de bloque/variedad con secuencia	# de combinaciones totales	% de combinaciones excluidas
80	158261	265	282	6.0
100	152971	262	282	7.1
125	146463	257	282	8.9

Para seleccionar la arquitectura el *walk-forward* para los modelos secuenciales, aplicamos el mismo principio que la búsqueda de hiperparámetros para las variantes tabulares, esto es, se prueban unas pocas arquitecturas candidatas y gana la de menor error de validación, pero con la diferencia de que para estas variantes la validación es un recorte cronológico dentro del período de entrenamiento del primer *fold* que usa el 85% del conjunto para búsqueda y el 15% del conjunto para validación, para no tocar en ningún momento los folds de prueba reales. Una vez elegida la arquitectura, se reentrena una red nueva por cada fold de la validación *walk-forward*, ninguna red se cachea entre las diferentes ejecuciones del notebook, exactamente

igual que las variantes tabulares, porque la validación *walk-forward* ya es parte del propio protocolo de evaluación.

De esta forma, los *folds* de las variantes secuenciales son un subconjunto de los mismos *folds* semanales usados por las variantes tabulares, para que ambas familias de modelo se validen sobre períodos de tiempo comparables.

Tabla 12
Resultado de la búsqueda de SEQUENCE_WINDOW

<i>SEQUENCE_WINDOW</i>	MAE LSTM	MAE TCN	promedio
80	155.493179	146.463852	150.978516
100	160.491776	151.896149	156.193962
125	160.580704	152.101135	156.340919

Empíricamente se estima que mientras más se acerca la ventana al periodo de retraso en que se alcanza el pico de la correlación entre el *pinch* y la cosecha, mejor deben de rendir los modelos secuenciales, sin embargo, como se ve en la Tabla 12, el menor MAE promedio para los modelos secuenciales se logra en una ventana de 80 días.

Curva de entrenamiento de variantes secuenciales

La función que recorre las épocas solo imprime la pérdida cada 5 épocas y no la devuelve para poder asignarla a una variable, esto provoca que no se disponga de un historial por época accesible sin modificar la función. Lo que si devuelve la función (*train_sequential_walk_forward*) es la predicción del conjunto de entrenamiento y de pruebas de cada *fold*, ese valor permite obtener un sustituto legítimo para el MAE del conjunto de entrenamiento versus el de prueba por cada *fold* para las tres redes base (*LSTM_GLOBAL*, *TCN_GLOBAL*, y la red base de *LSTM_RESID_BIO*).

LSTM_RESID no tiene su propia curva, en su lugar, reutiliza estas mismas redes de *LSTM_GLOBAL* y solo le agrega una corrección *Ridge* por variedad.

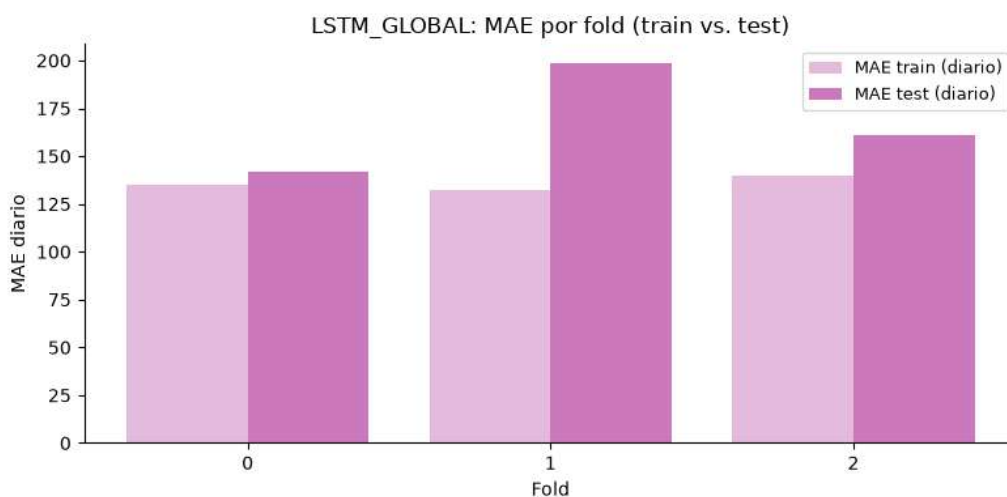


Figura 39 MAE por fold (train vs. test), variante LSTM_GLOBAL

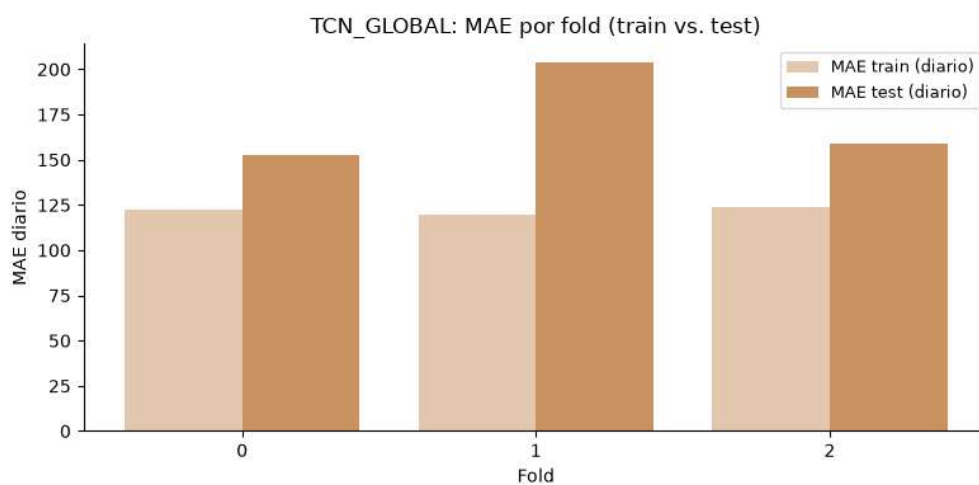


Figura 40 MAE por fold (train vs. test), variante TCN_GLOBAL

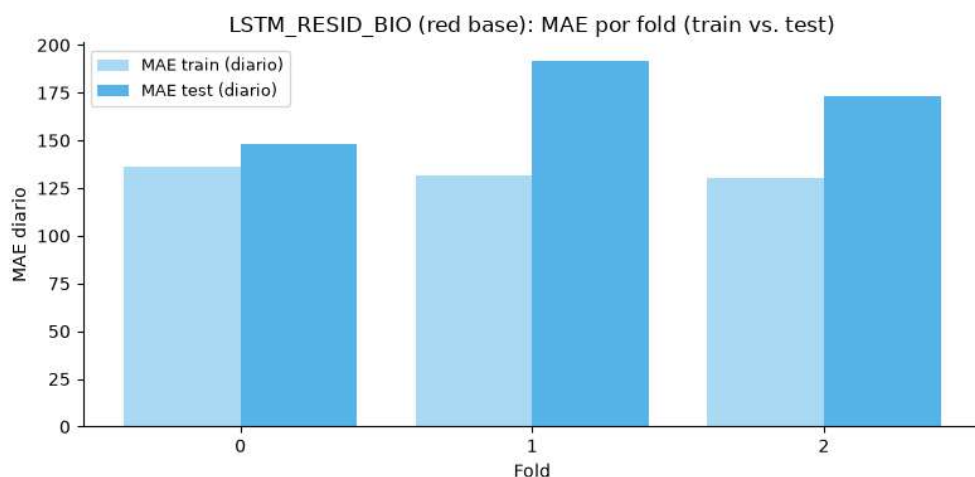


Figura 41 MAE por fold (train vs. test), variante LSTM_RESID_BIO

Resultados de las variantes secuenciales

Las variantes *LSTM_GLOBAL* y *TCN_GLOBAL* parten del mismo diseño, una red globalizada, una ventana de 80 días y la no consideración del *FEAT_BIO*, y quedan cerca en desempeño con un MAE de 524.72 y RMSE de 891.59 para la variante LSTM y con un MAE de 552.9 y un RMSE de 838.64 para la variante TCN. El RMSE más bajo de la TCN sugiere menos errores grandes puntuales, pero su MAPE (24.13 % vs. 20.25 %) y su tasa de cumplimiento (45.56 % vs. 56.38 %) son peores, es decir, falla con más frecuencia en las variedades y semanas de menor volumen, aunque sin errores tan extremos como los que evitaría un RMSE bajo. En ambos casos el *fold* 1 concentra el grueso del error diario y muestra el mismo patrón estacional que ya se observaba en las variantes tabulares.

Tabla 13
Métricas relevantes de las variantes secuenciales

Variante	MAE	RMSE	R2	MAPE	Tasa de cumplimiento
<i>LSTM_GLOBAL</i>	524.72	891.59	0.91	20.25	56.38
<i>TCN_GLOBAL</i>	552.90	838.64	0.92	24.13	45.56
<i>LSTM_RESID</i>	494.23	883.69	0.91	18.31	61.44
<i>LSTM_RESID_BIO</i>	461.11	866.88	0.91	17.70	64.71

La variante *LSTM_RESID* reutiliza los *folds* ya entrenados de *LSTM_GLOBAL* (sin reentrenar) y añade una corrección Ridge por variedad sobre la combinación de predicción y día del año. El ajuste se aplica a un promedio de 118 variedades por *fold* y mejora la LSTM base en las cuatro métricas relevantes. El RMSE apenas cambia unas cuantas unidades siendo esto una señal de que la corrección ajusta el sesgo sistemático por la combinación de variedad y época del año más que los errores grandes aislados.

Se definió que la variante *LSTM_RESID_BIO* repita el mismo esquema de red y corrección residual pero que entrene la red base con *FEAT_BIO* incluido, esto resulta en que sea la mejor variante secuencial con base en las cinco métricas consideradas (*MAE*, *RMSE*, *MAPE* y tasa de cumplimiento). La mejora sobre *LSTM_RESID* aísla el aporte de las variables biológicas, con todo lo demás igual, incluirlas baja el MAE (Figura 41) y sube la tasa de cumplimiento.

En conjunto, la Tabla 13 muestra que ambas mejoras, corrección residual por variedad y *FEAT_BIO*, son acumulativas sin sacrificar el R^2 .

CAPITULO 4

4. ANÁLISIS DE RESULTADOS

4.1. Ventana Amplia por Variante

Para comparar los resultados de la predicción de los modelos con el *baseline*, se cruza la predicción de cada variante con la producción real y la proyección humana por cada combinación de semana, bloque y variedad. Cada variante conserva su propia ventana de comparación con la amplitud que sus propios datos lo permitan, en este aspecto es importante apuntar que no se consideran las intersecciones con otras variantes para evitar el descarte subóptimo de datos.

En la Figura 42 vemos que en general, el *MAE* de los modelos logra una mejora con respecto al del *baseline*, siendo la variante *LSTM_RESID_BIO* el que logró un mejor valor con una reducción aproximada del 40% respecto al *baseline*. Otros modelos como el *HistGBT individual* y el *CatBoost individual* tuvieron un *MAE* peor que el de la proyección humana.

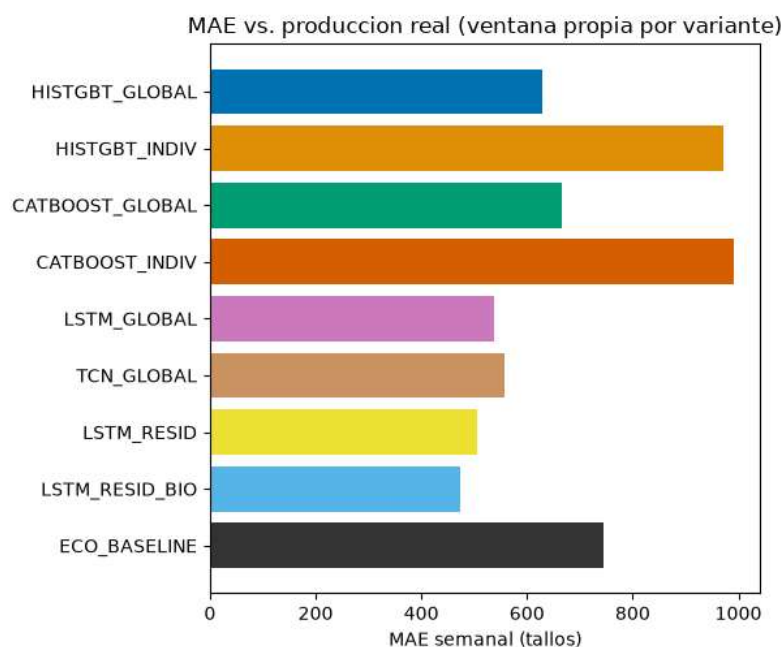


Figura 42 MAE de las variantes versus el del *baseline*, usando sus propias ventanas

Si analizamos el coeficiente de determinación (R^2), podemos observar en la Figura 43 que todas las variantes de los modelos logran una mejora sin que ninguno destaque de forma notable, a pesar de aquello, se puede notar que el presenta un mejor desempeño es el *TCN global*.

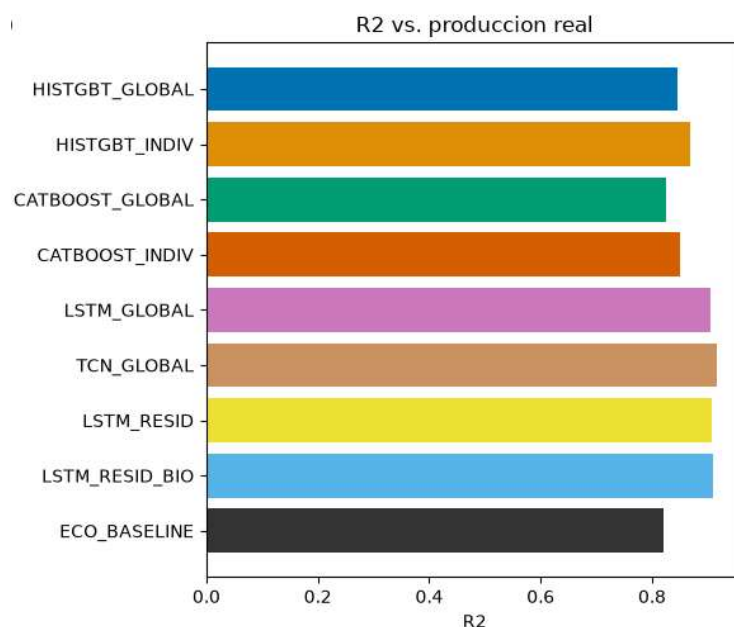


Figura 43 R^2 de las variantes versus el del baseline, usando sus propias ventanas

La tasa de acierto definida por el negocio también tiene una mejora en todas las variantes de los modelos estudiados. Al igual que con el MAE, los modelos que destacan por sobre el resto son el *LSTM_RESID* y el *LSTM_RESID_BIO*, las otras variantes de los modelos tienen un desempeño similar entre ellas. Esto lo podemos ver en la Figura 44.

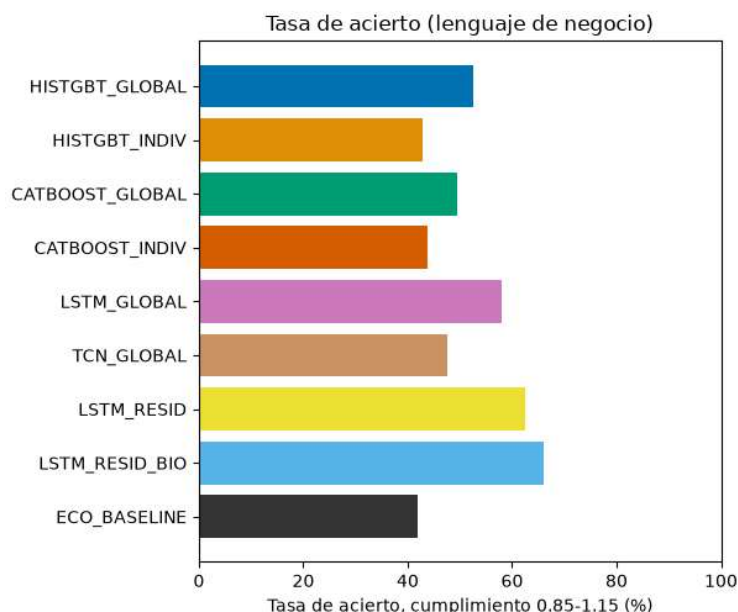


Figura 44 Tasa de acierto de las variantes versus el del baseline, usando sus propias ventanas

4.2. Ventana Común

A diferencia de la ventana amplia por variante, la ventana común usa la misma intersección en la combinación de semana, bloque y variedad, es decir, todas las variantes de los modelos estudiados realizan la predicción de forma simultánea usando la misma ventana.

Al analizar los resultados del MAE ilustrados en la Figura 45, se puede establecer que las variantes basadas en redes neuronales tienen un mejor desempeño bajo estas condiciones, siendo la variante *LSTM_RESID* la que mejor se desempeña.

El coeficiente de determinación para las diferentes variantes mostrado en la Figura 46, al igual que con el MAE, se comporta mejor en los modelos basados en redes neuronales, siendo el menos favorecido el de la variante *LSTM_RESID_BIO*. Si bien las variantes basadas en *gradient boosting trees* tienen buenos valores, estos siguen estando por debajo de los obtenidos por las de redes neuronales.

La última métrica usada para esta comparación, la tasa de asertividad basada en las reglas del negocio muestra como claro ganador a la variante *LSTM_RESID* (ver Figura 47).

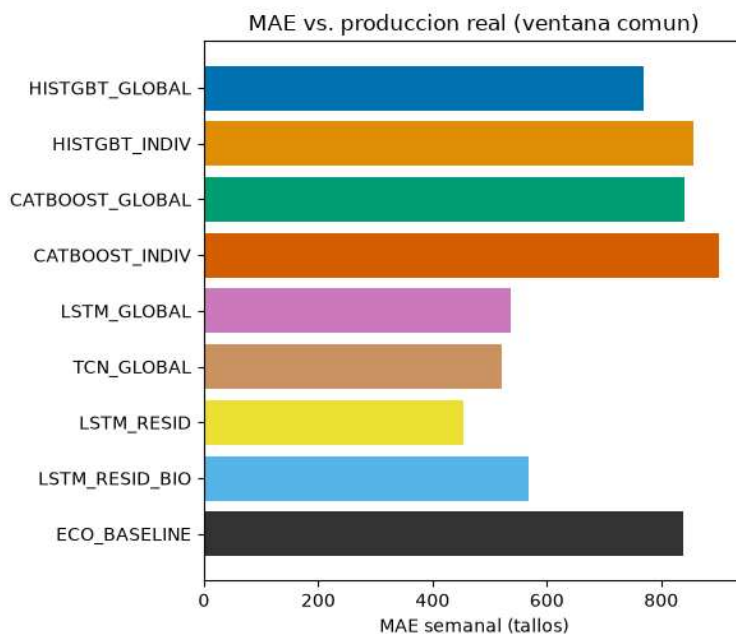


Figura 45 MAE de las variantes versus el del baseline, usando la ventana común

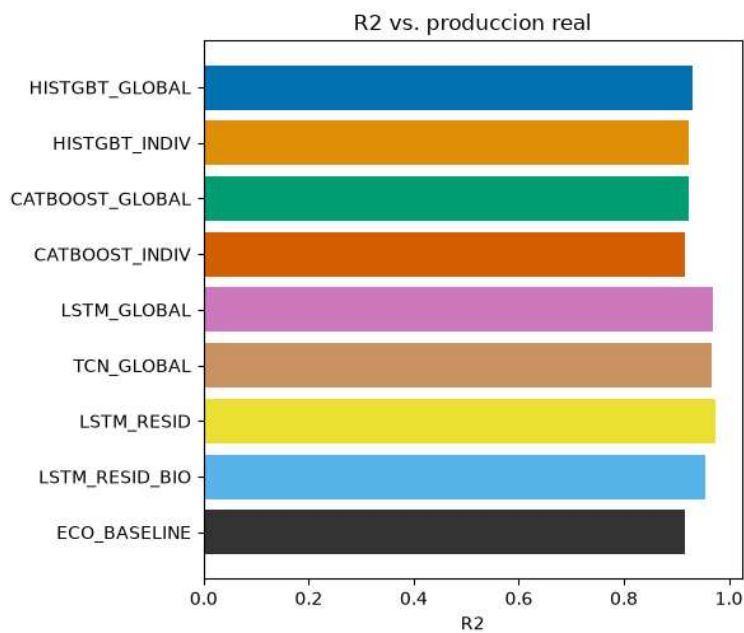


Figura 46 R^2 de las variantes versus el del baseline, usando la ventana común

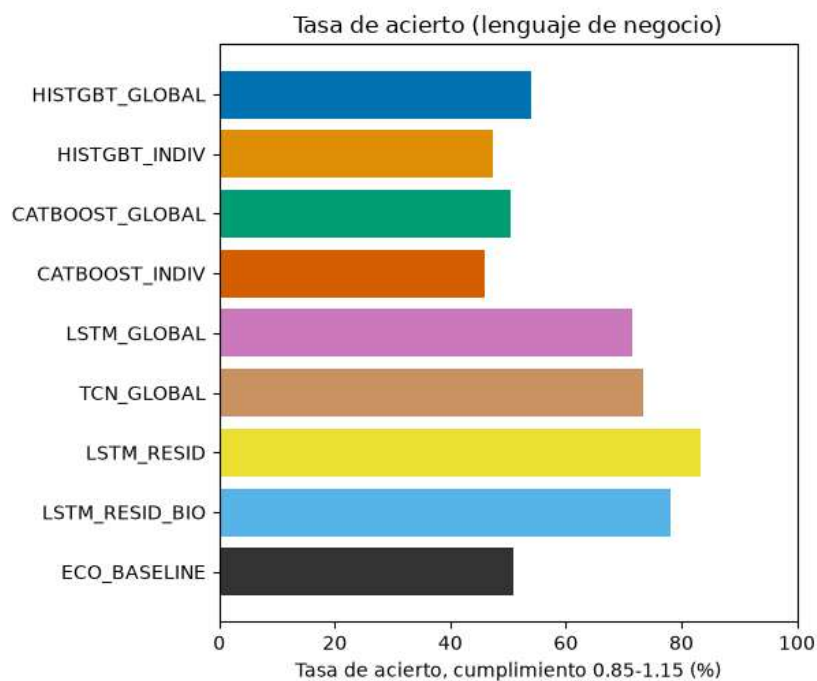


Figura 47 Tasa de acierto de las variantes versus el del baseline, usando la ventana común

Si analizamos las magnitudes obtenidas, vemos que en general la tasa de cumplimiento mejora en todas las variantes usando una ventana común, incluso la del baseline (ver Figura 50). A su vez, el coeficiente de determinación mejora de forma consistente en todas las variantes (Figura 49) pero en el caso del *MAE*, los modelos basados en *gradient boosting* individuales resultaron tener un peor desempeño bajo estas condiciones.

De forma general podemos establecer que es posible preferir el uso de una ventana común por sobre la ventana amplia por variante.

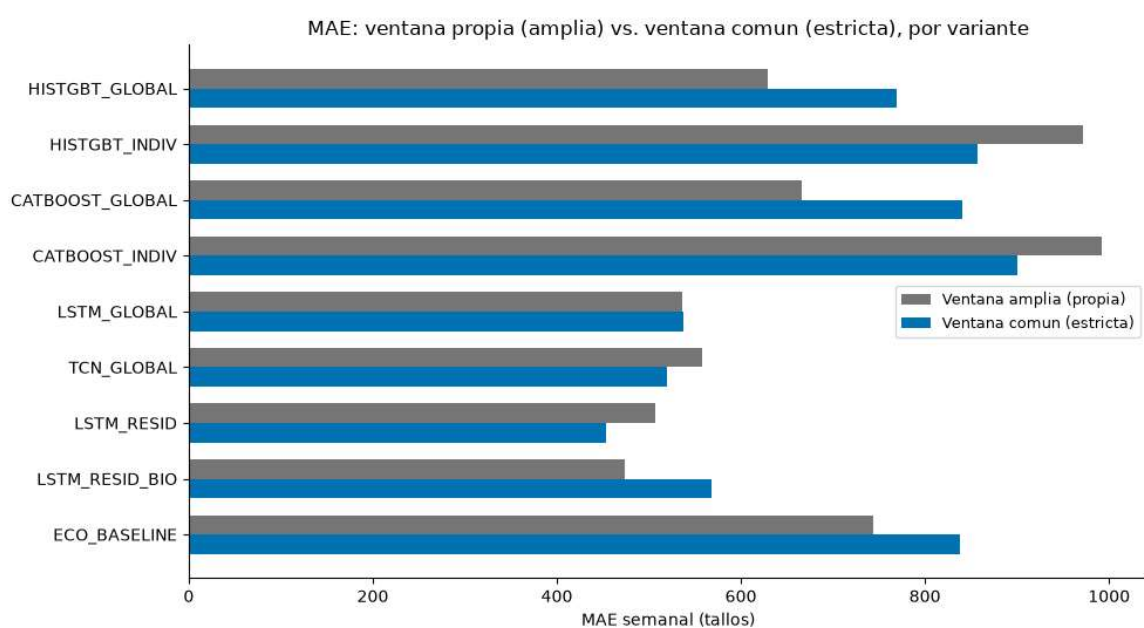


Figura 48 Comparativa de MAE de cada variante, ventana amplia vs común

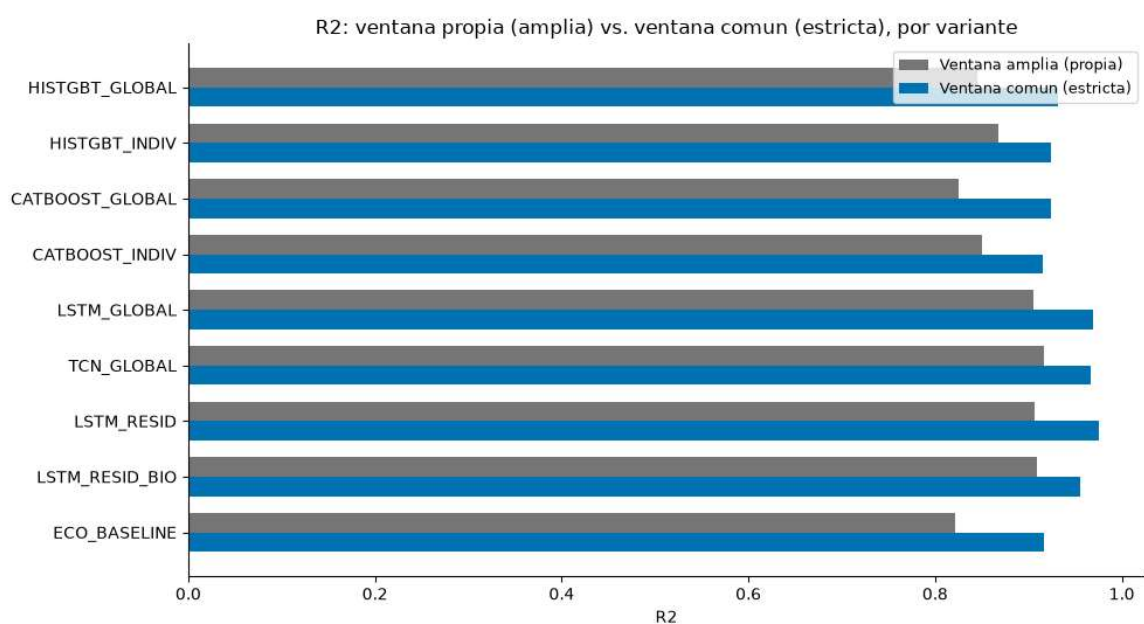


Figura 49 Comparativa de R2 de cada variante, ventana amplia vs común

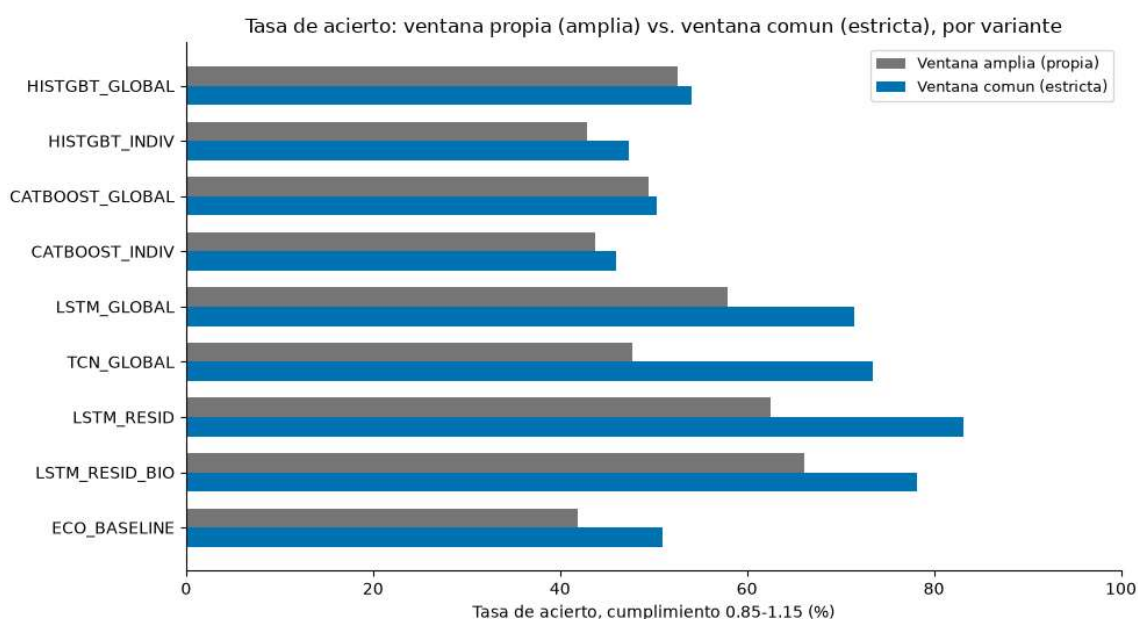


Figura 50 Tasa de acierto de cada variante, ventana amplia vs común

4.3. Predicho vs. Real

En los diagramas de dispersión de la Figura 51 y Figura 52, cada punto es una combinación de bloque, variedad de flor y semana del conjunto de pruebas, la línea punteada representa la diagonal perfecta en la que la predicción iguala a la producción real. Para cada una de las cuatro familias de especialización, esto es, árboles global e individual, y secuencial global y secuencial más residual, solo la variante ganadora de esa familia según MAE. Por eso muestra HistGBT global e individual y LSTM global y LSTM residual con FEAT_BIO.

Los dos paneles de árboles (Figura 51) muestran un puñado de semanas de volumen extremo, esto es visible en que el eje de HistGBT global llega al rango de 60.000 a 80.000 tallos, donde el modelo subestima con fuerza, esto se hace visible al observar que el punto más extremo cae a menos de la mitad de la producción real. El resto de la nube de puntos, por debajo de 20.000 tallos, se ajusta razonablemente bien a la diagonal.

Ambos paneles secuenciales (Figura 52) tienen ejes mucho más acotados de hasta alrededor de 30.000 a 35.000 tallos, al mismo tiempo su nube está más concentrada cerca del origen, y el panel de LSTM global muestra algunas predicciones negativas para semanas de producción real baja. El panel de la

derecha (LSTM + residual + FEAT_BIO, $R^2=0.91$, el más alto de los 4) es visualmente el más ajustado a la diagonal, aunque con una ligera tendencia a sobreestimar en el extremo superior.

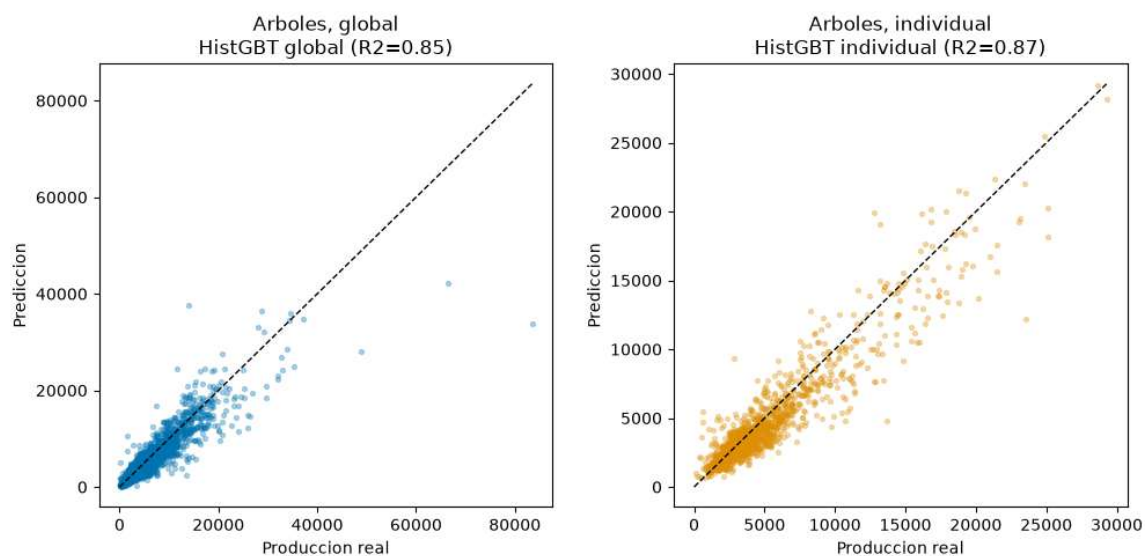


Figura 51 Gráficos de dispersión de variantes HistGBT

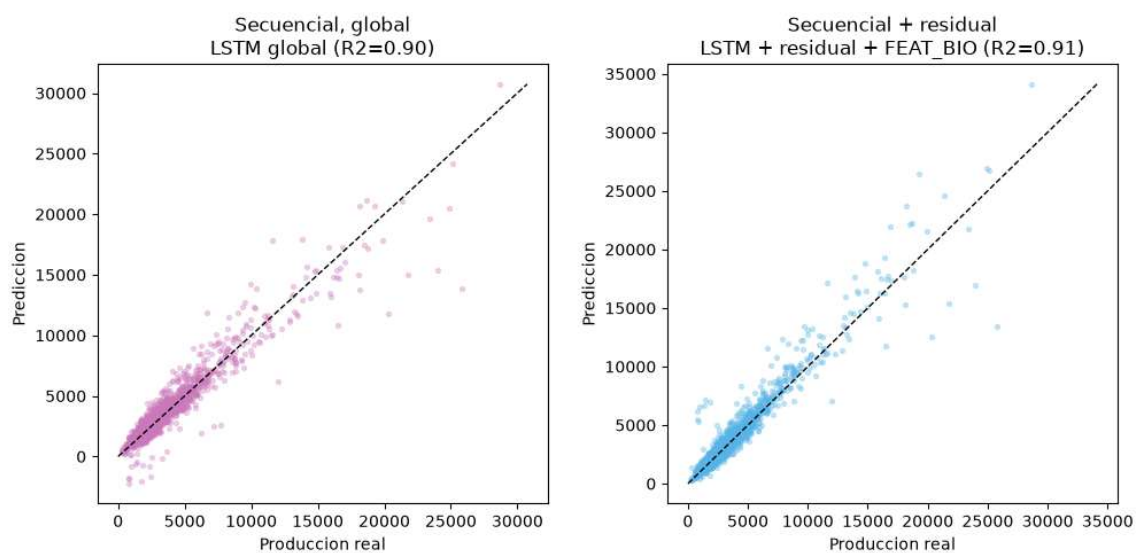


Figura 52 Gráficos de dispersión de variantes secuenciales

En la Figura 53, la línea punteada en 50% es el punto de referencia, en esa línea la probabilidad de que las variantes predigan mejor que la proyección humana, variedad por variedad, se equipara a lanzar una moneda y apostar por cada o cruz. Las 6 variantes globales, es decir, ambas familias de árboles y

las 4 secuenciales, superan holgadamente ese 50% en un rango que va del 67% al 86% de las variedades, con *LSTM_RESID_BIO* a la cabeza con un 85.9% de las veces. Las 2 variantes individuales (*HISTGBT_INDIV* y *CATBOOST_INDIV*) son la excepción marcada, le ganan a la proyección humana en apenas de un 13% al 15% de sus variedades, muy por debajo de las globales y del propio 50% de referencia. Esta interpretación refuerza la idea de que la especialización individual no aporta a realizar mejores predicciones en las variantes de *gradient boosting trees*, vista ahora variedad por variedad en vez de en el agregado.

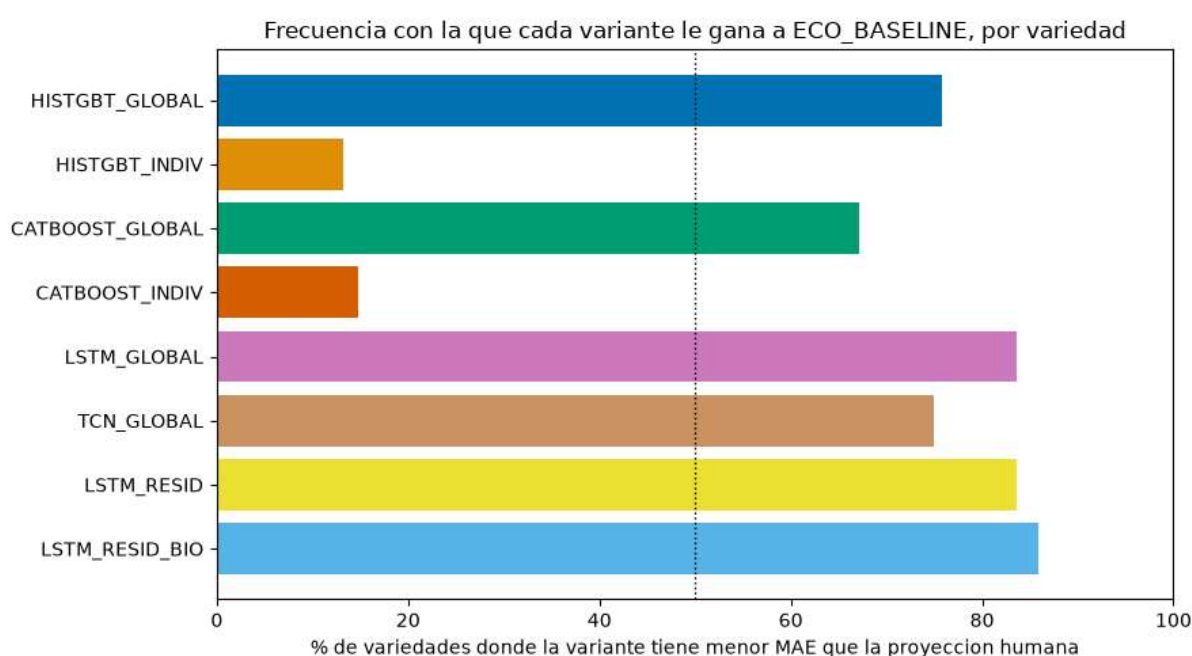


Figura 53 Frecuencia en que cada variante le gana variable por variable al baseline

El análisis sobre cuán buena resulta una variante de los modelos estudiados para cada variedad de flor, también puede abordarse desde un mapa de calor como el mostrado en la Figura 54, en donde cada fila es una variedad con al menos 15 semanas de datos, el color va de rojo ($R^2 = -1$, peor que predecir el promedio) a verde ($R^2 = 1$, ajuste perfecto). Las columnas de *HISTGBT_GLOBAL* y *CATBOOST_GLOBAL* son casi uniformemente verdes, son consistentes variedad por variedad, no solo en el promedio. Las columnas con el sufijo *_INDIV*, correspondientes a las variantes con segmentación

individual, tienen huecos en blanco para las aproximadamente 92 variedades fuera de su cobertura y solo cubren las 30 de mayor volumen, y donde sí tienen dato, son visiblemente menos verdes que su contraparte global.

Las 4 columnas de las variantes secuenciales tienen mucha más variación fila a fila, conviven variedades en verde oscuro con bloques de variedades en rojo intenso con un R^2 muy negativo. La misma variante que gana en el agregado puede fallar de forma severa en variedades puntuales donde los modelos de árboles globales siguen siendo confiables. El *baseline* muestra un patrón parecido al de las secuenciales, bueno en general, pero con sus propios puntos rojos.

La variabilidad de variedad por variedad de la variante *LSTM_RESID_BIO* es un riesgo para monitorear en producción que los modelos de árboles globales no presentan en la misma magnitud.

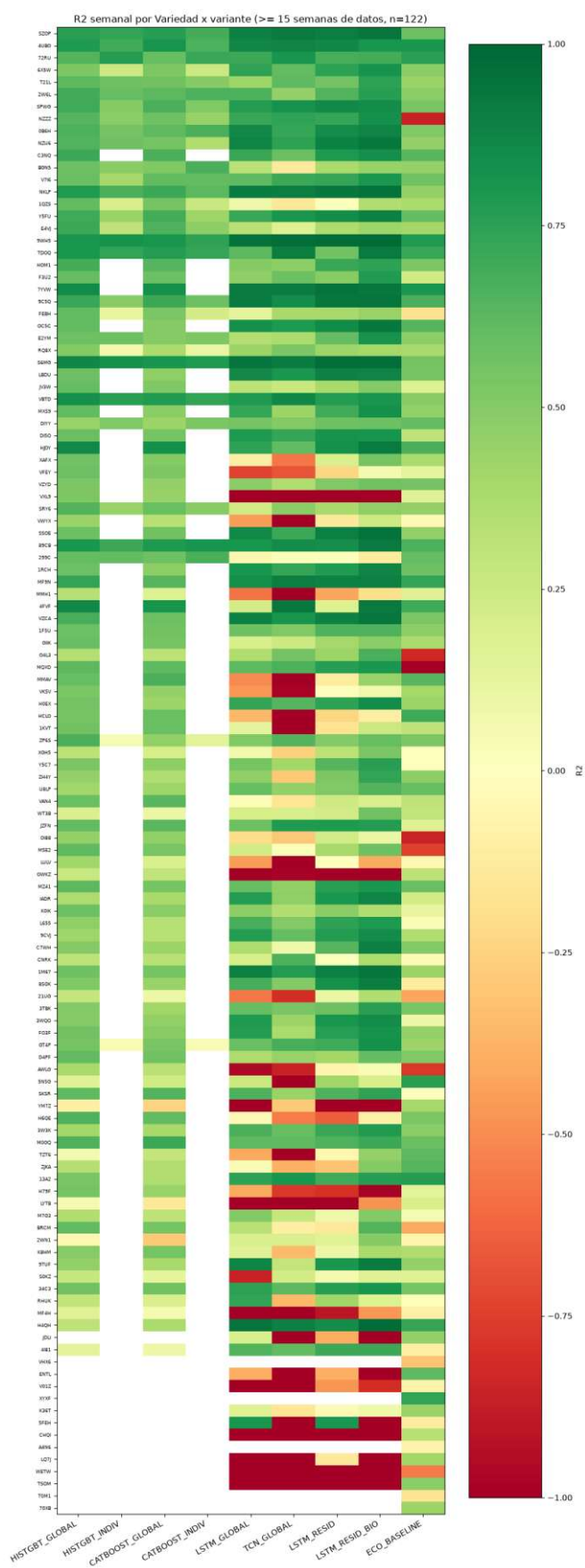


Figura 54 Mapa de calor de R^2 de la predicción semanal por variedad y variante

4.4. Explicabilidad e Importancia de Variables

La capacidad de entender qué usa cada modelo para predecir es tan importante como saber cuál tiene menor error, un modelo que dependa casi exclusivamente de una sola señal frágil (por ejemplo, su propia autocorrelación) es más riesgoso de desplegar que uno que combine varias fuentes de información, aunque su error puntual sea parecido. En esta sección se responde a esa pregunta con un método unificado, aplicado por igual a las 4 familias de modelo, para que los resultados sean directamente comparables entre sí.

4.4.1. Importancia por Permutación Agrupada

El método de la permutación agrupada mide cuánto empeora el error del modelo (aumento de MAE) cuando se mezcla al azar, entre observaciones, un grupo entero de variables relacionadas. Un aumento grande de error indica que el modelo depende fuertemente de esa información, un aumento pequeño o nulo indica que esa información aporta poco. Los grupos usados son los mismos para los cuatro modelos, para que la comparación entre familias sea directa:

- Autorregresivo: la propia cosecha reciente de esa combinación de bloque y variedad.
- Biológico (*FEAT_BIO*): la ventana de *pinch* a cosecha estimada por CCF, presente solo en las variantes que usan *FEAT_BIO*.
- Clima: todas las variables de clima. Incluyen los datos intra y extra-invernadero.
- Calendario: estacionalidad, semana del año o día del año, según el grano del modelo.
- Cobertura: fracción de días con dato dentro de la ventana agregada, solo existe en las variantes de *target* semanal nativo.

- Identidad: de qué bloque y qué variedad es cada observación.

Los porcentajes de cada modelo están normalizados para sumar 100%, así se puede comparar el peso relativo de cada grupo entre modelos con escalas de error muy distintas, por ejemplo, semanal versus diario o árboles versus redes neuronales.

4.4.2. Explicabilidad de los Modelos Secuenciales

Para las redes secuenciales, "permutar una variable" significa mezclar, entre observaciones, la serie temporal completa de un canal de entrada. El mismo principio que en los modelos tabulares, adaptado a que la entrada es una secuencia diaria, de longitud *SEQUENCE_WINDOW*, buscada como hiperparámetro en esta versión y no una fila de columnas.

4.4.3. Importancia Unificada de las Variables

Para lograr una visibilidad unificada, se generó una gráfica de los pesos normalizados por variante de modelo de cada grupo de variables definido, los resultados se presentan en la Tabla 14 en donde solo se han considerado las cinco principales variantes.

Tabla 14

Peso por grupo de variables (%), normalizado por cada modelo

Variables	HISTGBT GLOBAL	CATBOOST GLOBAL	LSTM GLOBAL	TCN GLOBAL	LSTM_RESID_BIO
Autorregresivo	55.5	78.4	74.1	73.8	80.9
Biológico	8.0	5.6	0.0	0.0	4.6
Clima	7.6	5.4	3.7	2.5	1.7
Calendario	7.3	6.4	0.0	0.1	0.0
Cobertura	0.0	0.0	0.0	0.0	0.0
Identidad	21.6	4.1	22.2	23.7	12.9

En la Figura 55 se ve como el grupo autorregresivo domina en las cinco variantes, obteniendo

entre el 55.5% y el 80.9% del peso total de la importancia de las variables. Esto confirma que ningún modelo escapa del todo al atajo de autocorrelación descrito previamente, incluso los modelos secuenciales, que ven toda la ventana de historia diaria completa, todos terminan apoyándose mayormente en el nivel reciente de cosecha.

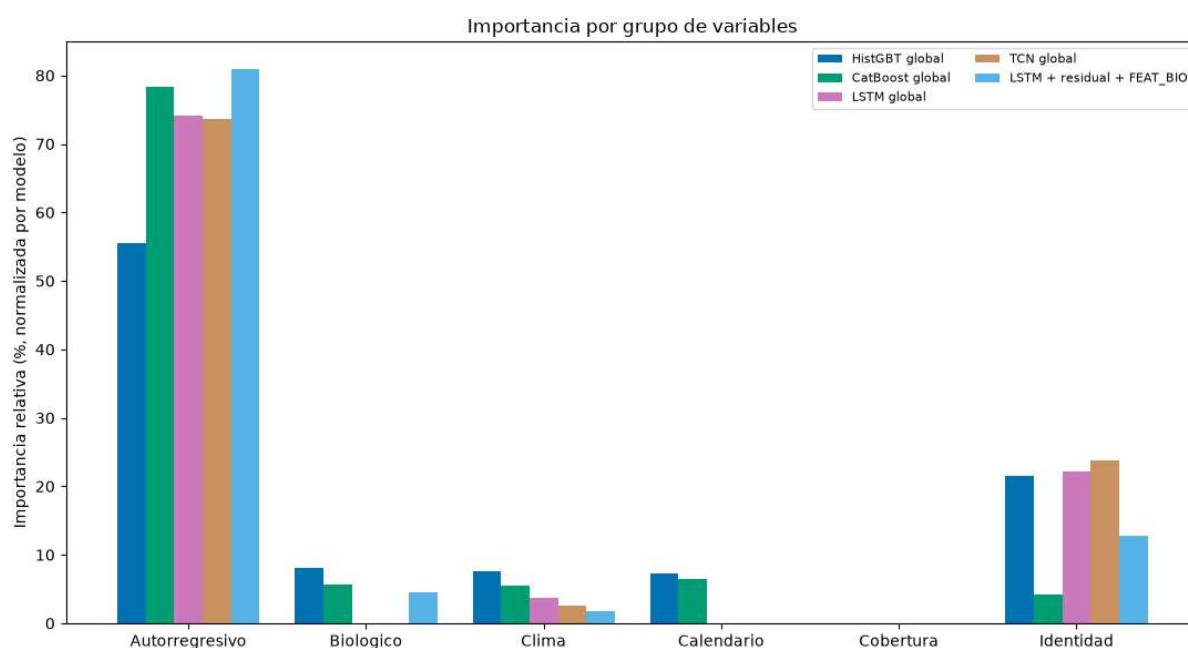


Figura 55 Peso de los grupos de variables por cada variante

De la Figura 55 también se observa un contraste que no era obvio de antemano, CatBoost casi no usa la identidad, mientras que TCN sobresale como la que más depende de ella. Una lectura razonable es que el encoding categórico nativo de CatBoost, es decir, las estadísticas del target ordenadas ya comprimen la información de identidad de forma tan eficiente que el modelo no necesita apoyarse mucho más en saber literalmente qué Bloque o Variedad es, mientras que el TCN, sin un mecanismo de encoding especializado, parece compensarlo construyendo su propia noción de identidad a partir de esa entrada. El grupo biológico pesa poco en términos relativos con una importancia que va del 4.6% al 8.0% en las variantes que lo tienen. Un peso relativamente pequeño en la importancia por permutación no impide que la señal aporte lo suficiente como para mejorar el resultado final; solo indica que, comparada con el

resto de las variables, no es lo que más mueve la aguja del error en la mayoría de las observaciones.

4.4.4. El peso de las Características Autoregresivas

¿Por qué los grupos Clima y Biológico pesan tan poco frente al Autorregresivo, si el sentido común agronómico dice que el calor o el frío extremos matan flores? La expectativa de que el clima debería pesar más es razonable y no es que el clima no importe biológicamente, es que en este conjunto de datos específico casi no varía lo suficiente como para que un modelo estadístico pueda usarlo. Es posible considerar tres verificaciones directas sobre el *weekly_panel* para demostrarlo:

- El clima se mide dentro de invernaderos con manejo activo, no a la intemperie. La temperatura máxima promediada en la ventana biológica tiene un desvío estándar de apenas un aproximado 1°C y un rango completo de 14°C a 22°C en las 26170 filas, no se acerca a un umbral que dañe la planta. Eso es el resultado de mantener las plantas dentro de un invernadero que absorbe los extremos climáticos antes de que lleguen al cultivo. El escenario de calor o frío extremo que uno esperaría que dominara la importancia casi no ocurre en los datos, precisamente porque el manejo agronómico ya lo previene y no porque no sería relevante si ocurriera.
- La ventana de agregación promedia cualquier pico corto. Aun si hubiera un par de días de calor puntualmente alto, lo resumen como una media o una suma sobre varios meses. Un evento de dos o tres días queda diluido entre los 85 a 110 días de clima normal, y su efecto sobre el promedio de la ventana es marginal aunque su efecto biológico puntual no lo haya sido.
- Las variables autorregresivas son, por lejos, las más correlacionadas con el *target*. La correlación entre el autorregresivo y las variables de clima es baja con un valor de

correlación de solo 0.06, así que tampoco es que el autorregresivo esté ocultado el efecto del clima por redundancia, sino que simplemente el clima parte con mucho menos poder predictivo crudo en el rango de valores observado.

Por lo expuesto, la lectura correcta no es que el clima no importe, sino que el clima ya está siendo manejado para no ser el factor limitante. La importancia por permutación mide cuánta información marginal aporta una variable dado el rango que realmente toma en los datos, no cuán crítica sería biológicamente si se le permitiera variar libremente. Un evento climático fuera de lo que el invernadero puede compensar, por ejemplo, una falla del sistema de control seguiría siendo una amenaza real para la cosecha, lo que ocurre es que no hay ejemplos de eso en el historial de entrenamiento para que el modelo lo aprenda.

4.4.5. Importancia individual de las variables

Además de la vista de importancia agrupada, analizamos el detalle variable por variable para los dos modelos tabulares globales y para *LSTM_GLOBAL* y *LSMT_RESID_BIO*.

Para los modelos basados en *gradient boosting trees* analizados a partir de la Figura 56, se tiene la coincidencia de que la variable que más explica las predicciones es *ar_harvest_w1* del grupo de variables autoregresivas, representa la cosecha semanal, es decir, los tallos cosechados de lunes a domingo de la semana inmediatamente anterior al ancla.

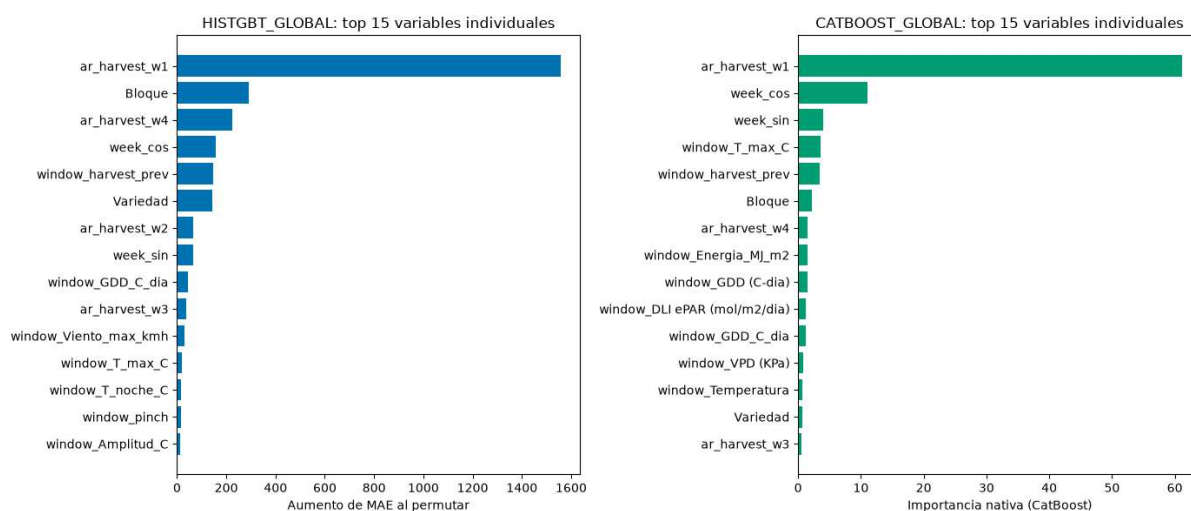


Figura 56 Importancia individual de variables de HistGBT y CatBoost globales

Para los modelos *LSTM* mostrados en la Figura 57 y en la Figura 58, se tiene que la variable más importante, por mucho, es la de tallos cosechados de las semanas previas, es decir, también es una variable autoregresiva.

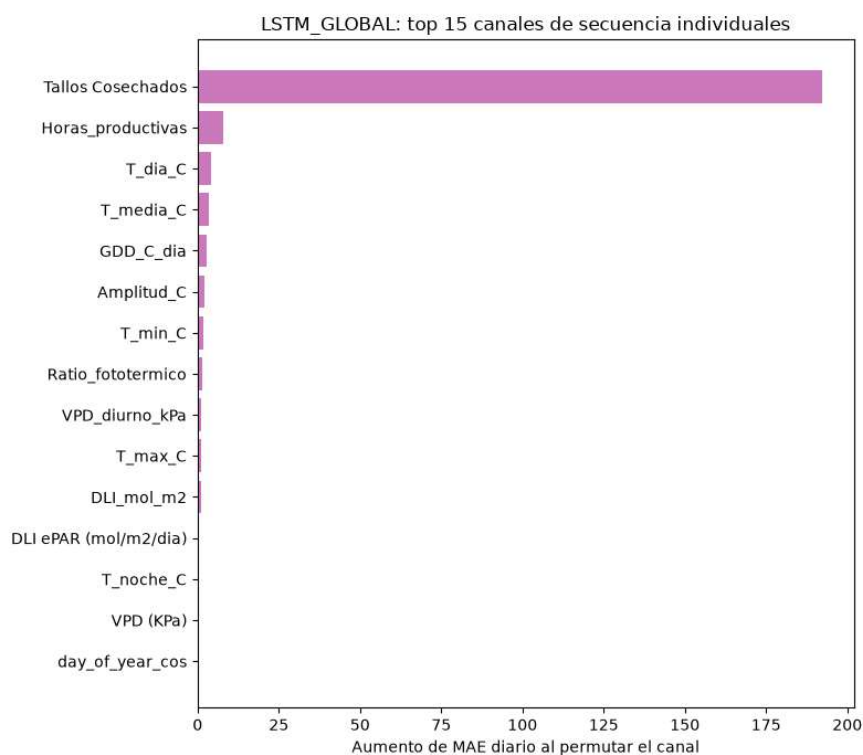


Figura 57 Importancia individual de variables de LSTM_GLOBAL

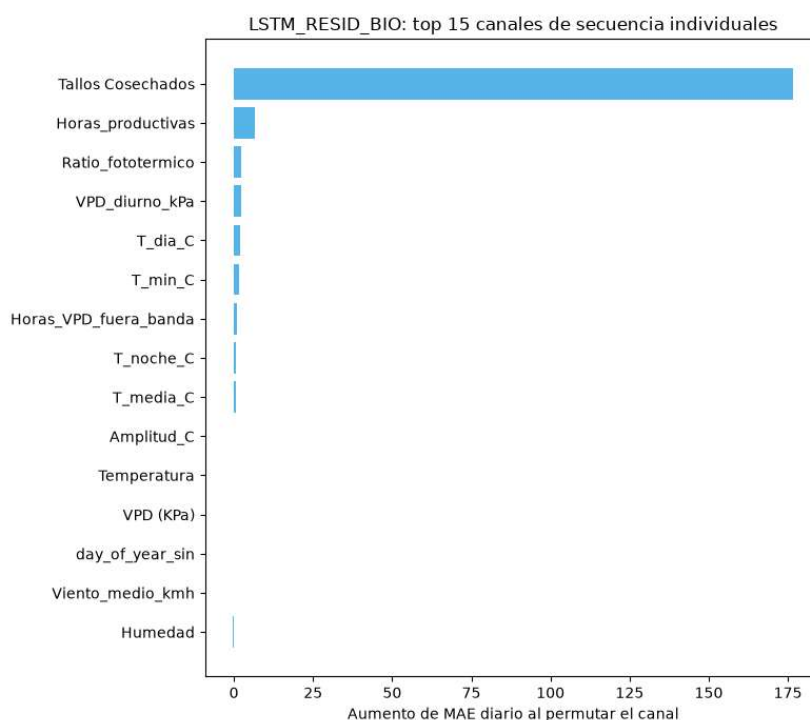


Figura 58 Importancia individual de variables de LSTM_RESID_BIO

4.5. Respuestas a las Preguntas de Diseño

A continuación, se responden cada una de las preguntas de diseño planteadas al inicio del desarrollo. En la Tabla 15 se muestra el resumen de los resultados obtenidos, la columna P (Bootstrap) indica si la diferencia de MAE entre el modelo ganador y el perdedor es distinguible de ruido de muestreo dado el tamaño de cada lado, por ejemplo, una fila que gana con MAE ganador, pero con valor P (Bootstrap) menor a 95% no debería tratarse con la misma confianza que una donde sí lo es.

Tabla 15
Resumen resultados de preguntas de diseño

Pregunta	Ganador	MAE Ganador	MAE Perdedor	Diferencia	P (Bootstrap)
1	HISTGBT_GLOBAL	629.8	666.1	+5.4%	97.9%
2	HISTGBT_GLOBAL	841.0	971.5	+13.4%	100.0%
2	CATBOOST_GLOBAL	879.9	992.1	+11.3%	99.6%
3	LSTM_GLOBAL	536.7	556.0	+3.5%	78.4%
4	LSTM_RESID	506.5	536.7	+5.6%	89.1%
5	LSTM_RESID_BIO	474.0	506.5	+6.4%	89.8%
6	LSTM_RESID_BIO	474.0	743.8	+36.3%	100.0%

4.5.1. Pregunta 1: ¿qué familia de árboles de gradient boosting predice mejor, a igualdad de todo lo demás?

La variante que mejor MAE logra es la de HistGBT global, siendo 5.4% mejor que la correspondiente CatBoost (ver Figura 59). Además, el valor de P (Bootstrap) supera el 95% por lo que el resultado puede ser tratado con confianza.

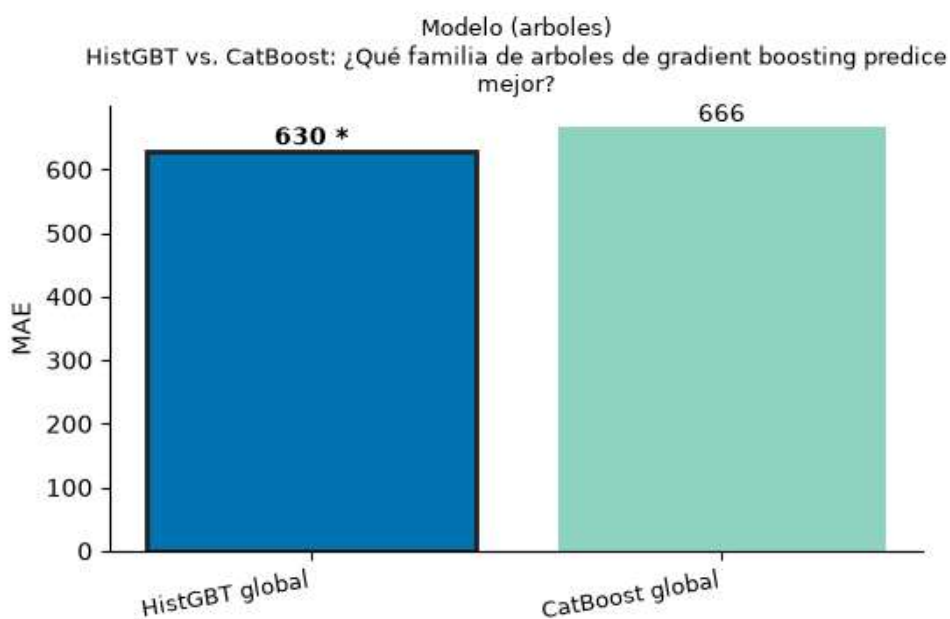


Figura 59 Comparativa de variantes de la pregunta 1

4.5.2. Pregunta 2: ¿especializar un modelo por variedad ayuda, o el modelo global ya generaliza igual de bien?

Según los resultados obtenidos, la estrategia de especialización empeoró el MAE del mejor modelo de *gradient boosting trees*. Para hacer esta comparación se restringen la cantidad de variedades de flores a 30, eso provoca que el MAE del modelo varíe con referencia a la pregunta anterior. El modelo HistGBT con especialización individual por variedad presenta un MAE mayor por lo que podemos decir que esta especialización en este modelo basado en *gradient boosting trees* empeora la predicción (ver Figura 60). Esta afirmación es sustentada por un valor de P (Bootstrap) del 100%.

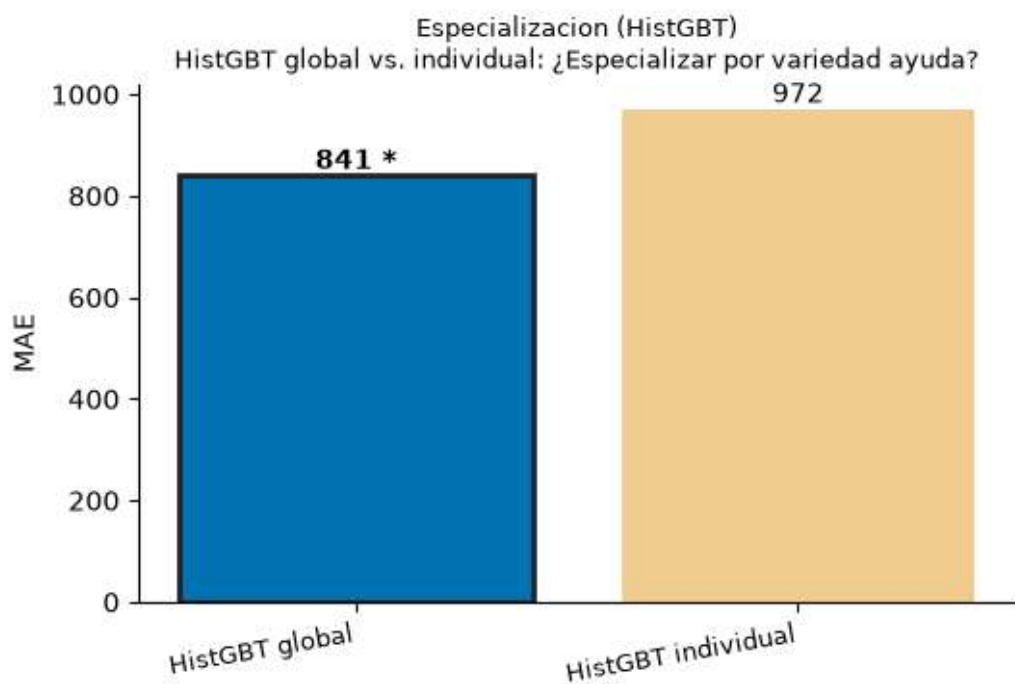


Figura 60 Comparativa de las variantes HistGBT de la pregunta 2

De manera similar ocurre con el modelo CatBoost, la especialización global tiene un MAE menor que el de la especialización individual (ver Figura 61). En este caso con un valor de P (Bootstrap) cercano al 100%, considerando esto, la respuesta a la pregunta planteada es que la especialización global produce mejores predicciones.

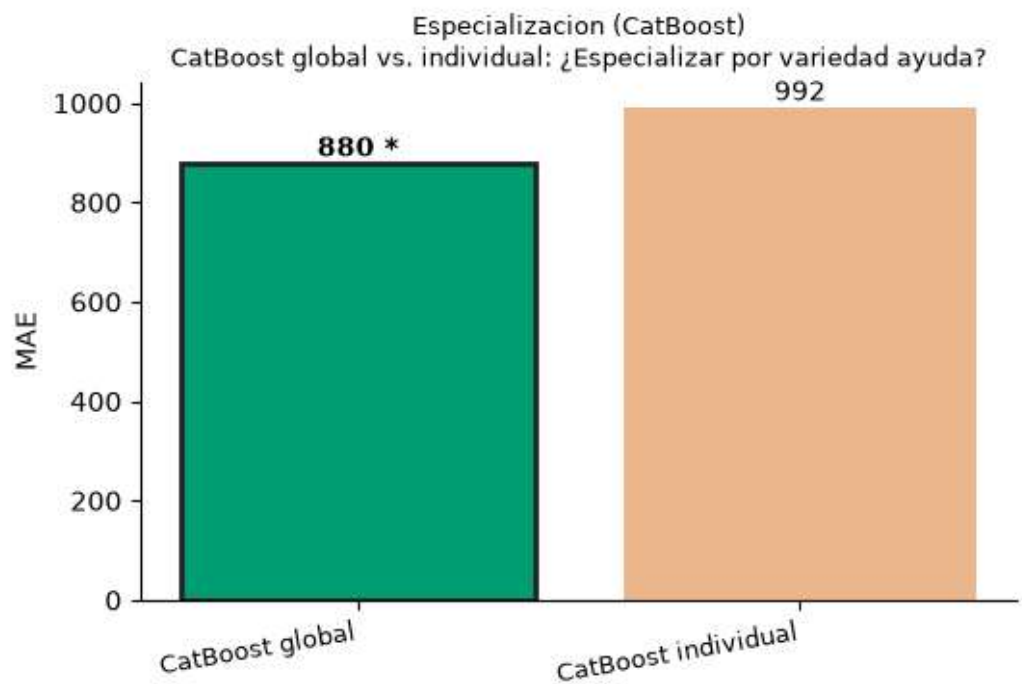


Figura 61 Comparativa de las variantes CatBoost de la pregunta 2

4.5.3. Pregunta 3: LSTM (recurrente) vs. TCN (convolucional causal) ¿cuál aprovecha mejor la secuencia diaria de 80 días de historia?

En el escenario planteado, la variante LSTM global presenta una leve ventaja por sobre la TCN global, sin embargo, para este caso el valor de P (Bootstrap) quedó por debajo del 95% con un valor de 78.4%, es decir, el resultado podría deberse a la existencia de ruido de muestreo. En la Figura 62 se muestran los valores de MAE obtenidos.

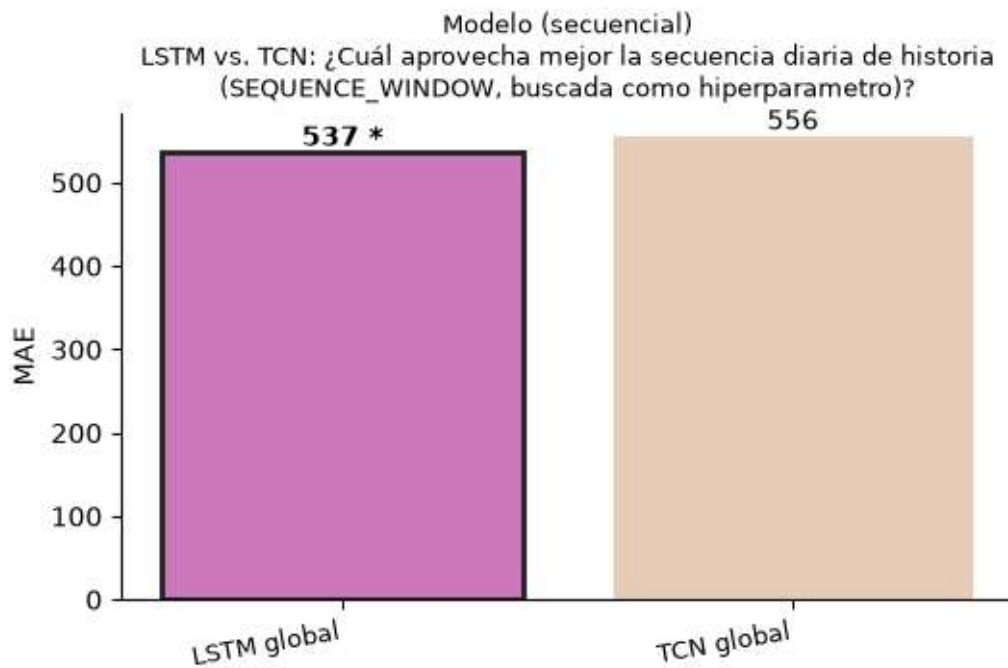


Figura 62 Comparativa de las variantes de la pregunta 3

4.5.4. Pregunta 4: ¿ayuda agregar una corrección residual por variedad sobre un LSTM global, sin pagar el costo de reentrenar la red por variedad?

Al agregar una corrección residual se logra una mejora de 5.6% (ver Figura 63) pero con un P (Bootstrap) de 89% por lo que, al igual que en la pregunta anterior, podría tratarse de ruido de muestreo.

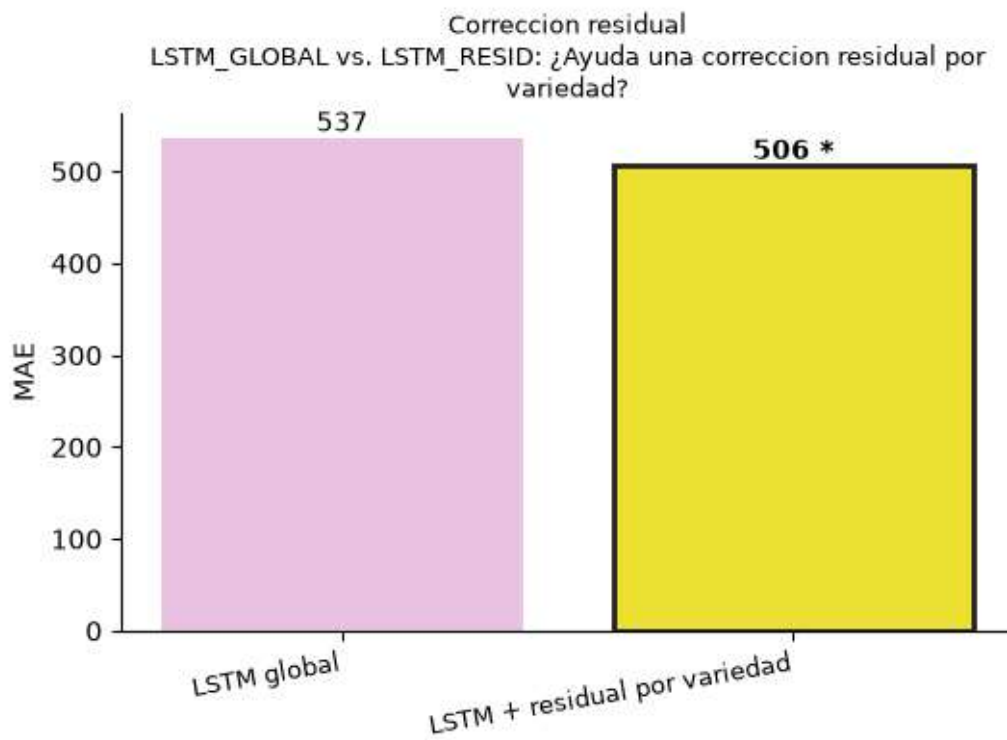


Figura 63 Comparativa de las variantes de la pregunta 4

4.5.5. Pregunta 5: ¿aporta algo agregar el predictor biológico explícito (FEAT_BIO) a un modelo que ya ve toda la historia diaria?

En este caso, agregar el predictor biológico explícito a la corrección residual del LSTM, provoca una mejora leve en el MAE, pero con un P (Bootstrap) por debajo del 95%. Los valores MAE obtenidos para ambos modelos se grafican en la Figura 64.

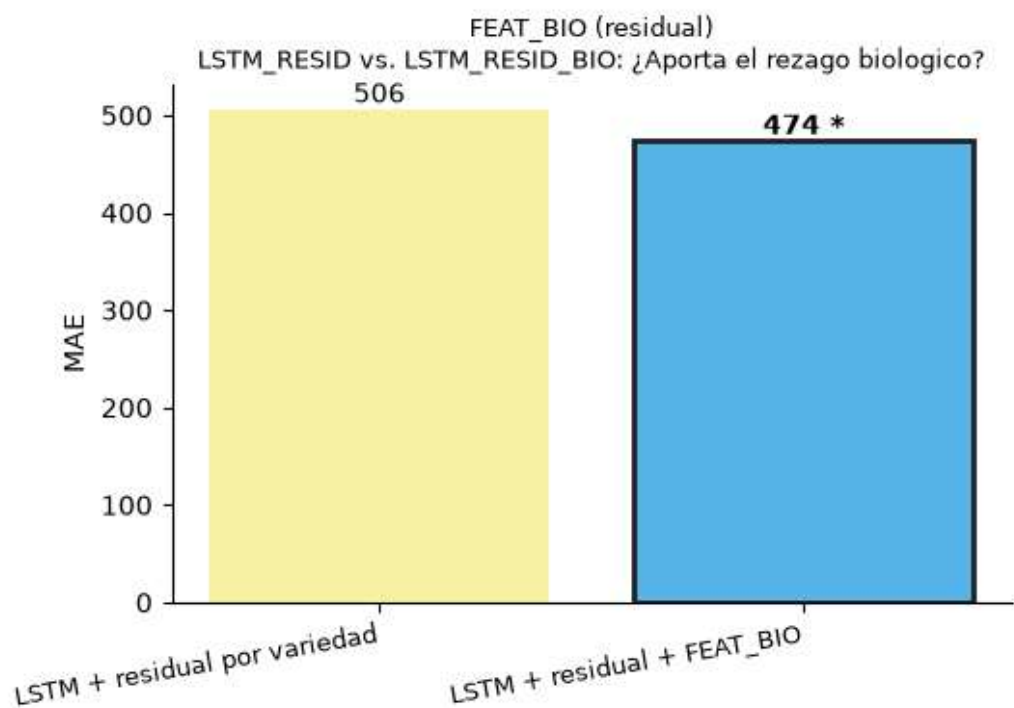


Figura 64 Comparativa de las variantes de la pregunta 5

4.5.6. Pregunta 6: ¿puede algún modelo superar a la proyección humana que la empresa ya usa para planificar?

Cuando se compara el MAE de la variante LSTM_RESID_BIO con el baseline humano, se obtiene una mejora que supera el 35% y genera un valor P (Bootstrap) de 100%.

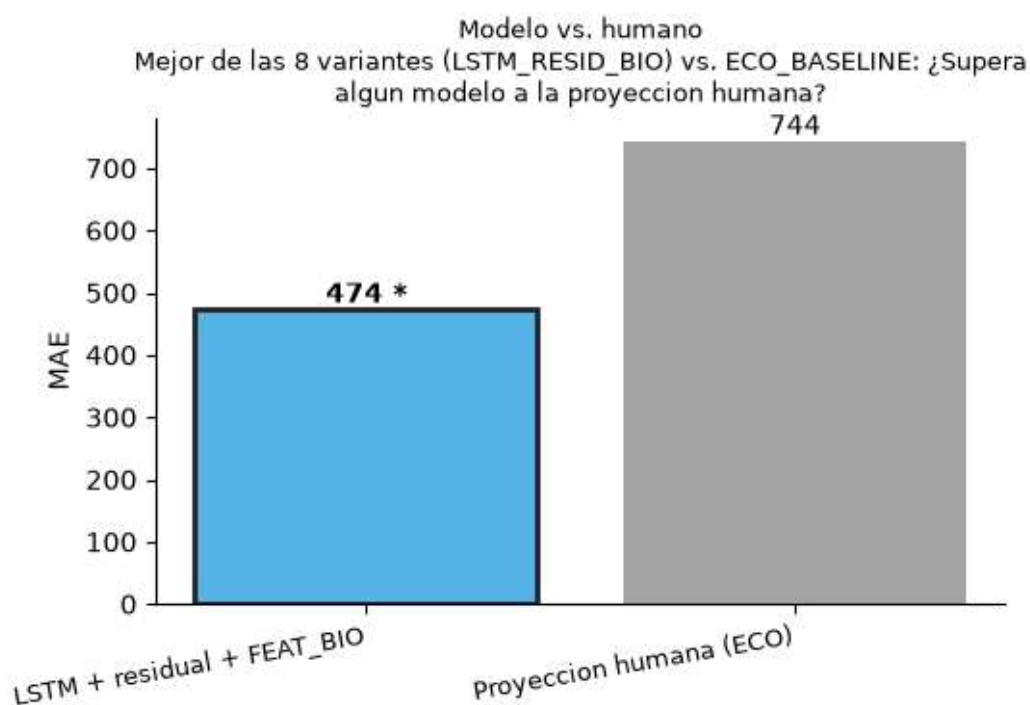


Figura 65 Comparativa de las variantes de la pregunta 6

4.6. Cuantificación de incertidumbre

Las proyecciones que actualmente se realizan son puntuales, por ejemplo, se predice que se van a cosechar 4200 tallos sin decir cuánta confianza se tiene en esa predicción. En términos de producción, es más útil especificar un intervalo, por ejemplo, se proyecta producir un intervalo de entre 3800 y 4600 tallos con un 90% de confianza. Esto último permite planificar no solo para el caso esperado sino para el peor y para el mejor.

4.6.1. Predicción conformal con recalibración rodante o expansiva

Este método consiste en que, para cada variante, se ordena cronológicamente sus predicciones y se recalcula cada cierto número de semanas el percentil 90 del error absoluto observado hasta ese momento, esto produce una ventana de calibración que crece con el tiempo y nunca mira hacia adelante, ese percentil 90 es el semiancho del intervalo aplicado a las predicciones del bloque de las semanas siguientes. Que el percentil 90 actúe como

semiancho significa que dicho margen de error cubre el 90% de los errores absolutos o residuos en una distribución empírica, de tal modo que las predicciones se expresan como la estimación de más menos ($\pm$) valor del percentil 90, garantizando que el 90% de las veces el valor real caiga dentro de ese rango.

Este es un método libre de distribución, es decir que no asume que el error sea normal ni de ninguna otra forma en particular y se adapta si el modelo empieza a cometer más errores con el tiempo en lugar de usar el semiancho fijo calculado una sola vez, por lo tanto, este semiancho es dinámico.

El tratamiento realizado a las variantes que no acumulan suficiente historia para la calibración y en consecuencia quedan sin un intervalo definido, es excluirlas del cálculo de cobertura. De esta forma el número de cobertura reportado no se beneficia de la calibración con datos insuficientes. La cantidad mínima de observaciones definida es de 20.

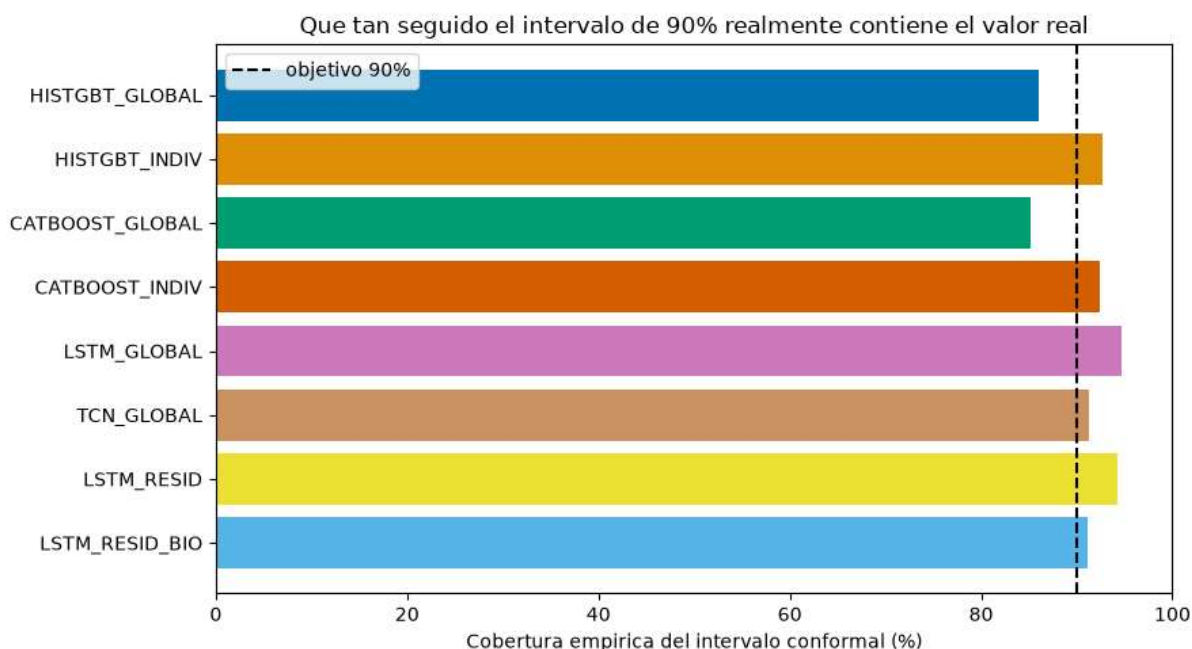


Figura 66 Nivel de confianza de las variantes estudiadas

En la Figura 66 se muestran la comparativa de los niveles de confianza obtenidos para cada variante. En este gráfico vemos que solo las variantes de CatBoost e HistGBT globales no

alcanzan el intervalo conformal o intervalo de predicción conforme del 90% establecido como objetivo.

4.6.2. Intervalo de predicción

Para ilustrar la incertidumbre seleccionamos dos de las variantes, CATBOOST_GLOBAL y LSTM_RESID_BIO, en la Figura 67 se grafica la predicción puntual junto con su banda de 90% de confianza, contra el valor real de una muestra de semanas de prueba, ordenadas por magnitud del valor real para que la banda se lea de forma más clara.

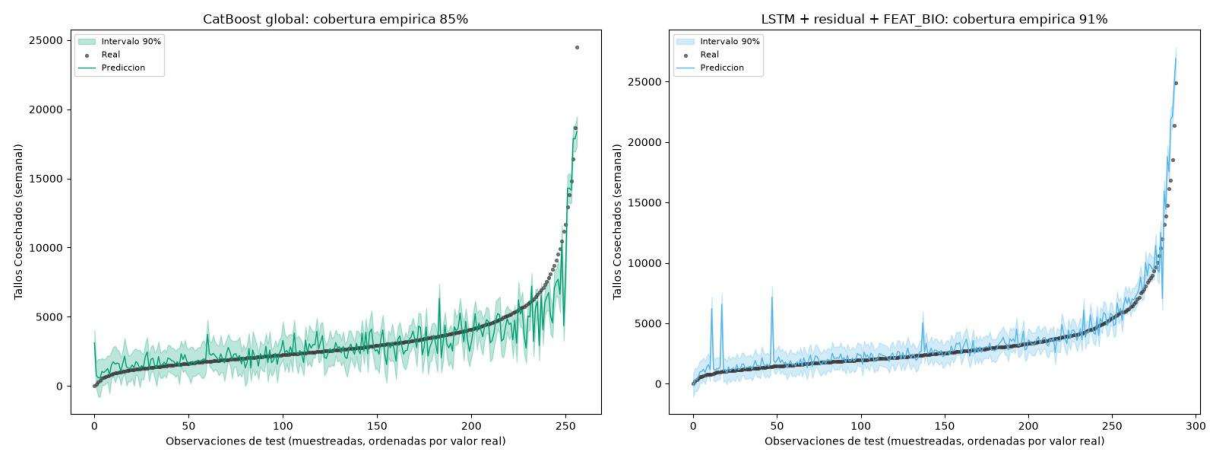


Figura 67 Predicción, su banda de confianza del 90% y valor de cosecha real

CAPITULO 5

5. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

5.1.1. Modelos que Superan a la Proyección Humana

El criterio seleccionado para comparar los modelos contra la proyección humana es el del error absoluto medio principalmente porque usa la misma métrica que la lógica del negocio, los tallos cosechados, esto lo hace directamente interpretable para la planificación. Adicionalmente al ser un valor entero y no sujeto a cálculos complejas, es menos sensible a valores extremos.

En la Figura 68 se muestran los porcentajes de reducción de *MAE* entre las diferentes variantes, tanto para la ventana amplia como para la ventana común estricta. Es interesante observar que para cada tipo de ventana tenemos un ganador diferente, para la ventana amplia gana *LSTM_RESID_BIO* y para la común estricta, *LSTM_RESID*. El tamaño de la muestra disponible para la ventana común es muy pequeño como para poder decidir entre una variante sobre otra con un buen grado de confiabilidad.

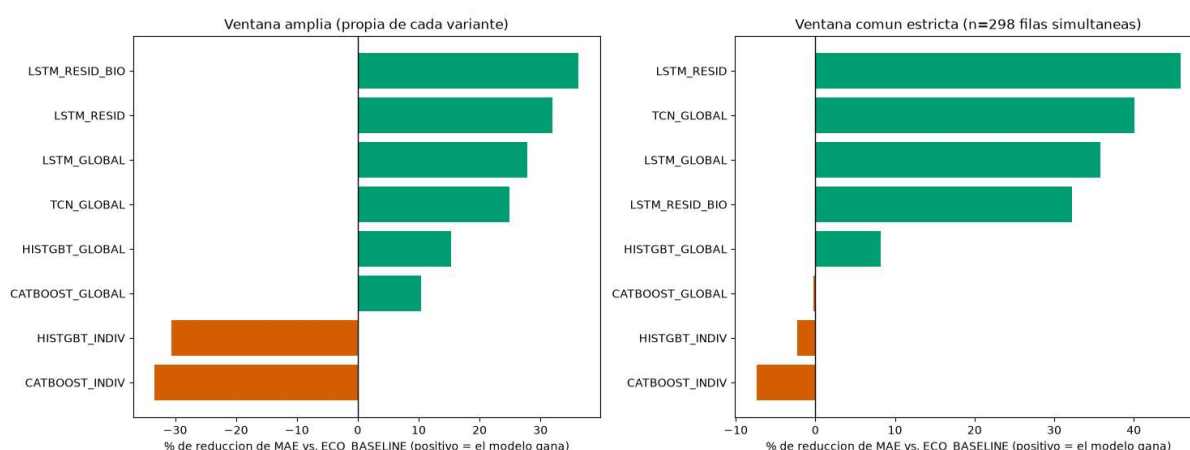


Figura 68 Comparativa de reducción de MAE por variante y tipo de ventana

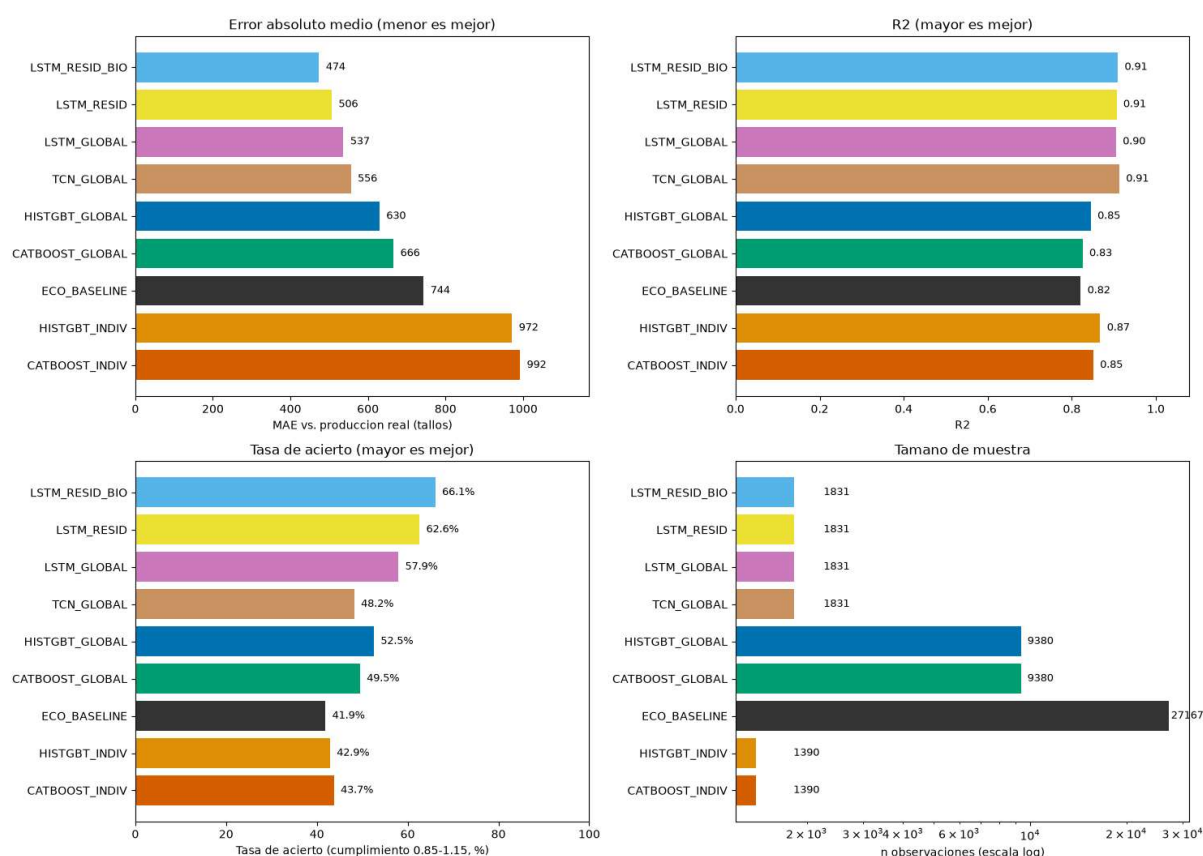


Figura 69 Ranking final de las variantes y el baseline

En la Figura 69 se muestra el ranking final de las métricas de las ocho variantes y del *baseline*. Aquí vemos que *LSTM_RESID_BIO* tiene el mejor *MAE*, iguala a los mejores R^2 , y presenta la mejor tasa de acierto alineada con las reglas del negocio, además, la selección de la variante *LSTM_RESID_BIO* se justifica al analizar las tres preguntas de diseño relacionadas, cada una de estas preguntas es progresiva en cuanto a mejor modelo, es decir, el ganador de la segunda descarta al de la primera y el ganador de la tercera descarta al de la segunda:

- Pregunta 3: *LSTM_GLOBAL* ($MAE = 536.7$) vs *TCN_GLOBAL* ($MAE = 558.1$)
- Pregunta 4: *LSTM_GLOBAL* ($MAE = 536.7$) vs *LSTM_RESID* ($MAE = 506.5$)
- Pregunta 5: *LSTM_RESID* ($MAE = 506.5$) vs *LSTM_RESID_BIO* ($MAE = 474.0$)

Los 3 pasos encadenan hacia el mismo ganador, cada paso usa como punto de partida al ganador del paso anterior, de esta forma es que *LSTM_RESID_BIO* termina acumulando las tres mejoras. Consistente con eso, tiene el MAE más bajo de las cuatro variantes secuenciales. Este comportamiento se explica porque cada comparación de variantes agrega una capacidad real sobre la anterior, la comparación de la pregunta tres elige la arquitectura secuencial que mejor aprovecha la ventana de historia diaria buscada, la comparación de la pregunta cuatro agrega una corrección liviana por variedad sobre esa arquitectura ya elegida y la comparación de la pregunta cinco le suma a esa corrección el rezago biológico explícito. Si cualquiera de las tres comparaciones hubiera revertido la secuencia de mejora, la recomendación final cambiaría, es por eso se explica la cadena completa, no solo el resultado final.

5.1.2. Riesgos y Limitaciones

El ganador de ventana amplia y el de ventana común estricta pueden diferir, como es en este caso, la ventana amplia le da a cada variante su propio período de prueba lo que resulta en más filas, pero no directamente comparable entre variantes; mientras que la ventana común restringe a las filas donde las nueve fuentes tienen predicción simultánea, esto resulta en menos filas, pero una comparación estrictamente igualada.

La disparidad de resultados entre los dos tipos de ventanas obliga a extender el tamaño de la muestra para la ventana común estricta, sin embargo, las decisiones tomadas para generarla se deben a limitaciones de poder cómputo más que a la disponibilidad de datos crudos.

Las únicas variantes que pueden competir por el pronóstico con granularidad diaria son las secuenciales, las otras fueron entrenadas para predecir de forma semanal, es decir, sus modelos no son capaces de producir una predicción diaria y no hay forma de usarlas para un pronóstico diario sin importar qué tan bajo sea su MAE semanal. Si bien esta limitación es

parte del diseño experimental, es importante de aclararlo para la justificación de la selección para este caso de uso.

La restricción de las variantes tabulares con pronóstico semanal nativo de no poder competir con el caso de uso de pronóstico diario no aplica al revés, es decir, las variantes secuenciales a pesar de ser entrenadas para la predicción diaria sí pueden realizar predicciones semanales al agregar las predicciones de los días de la semana objetivo. Por este motivo se puede comparar directamente a variante *LSTM_RESID_BIO*, recomendada arriba para grano diario, contra la variante *HISTGBT_GLOBAL*, la recomendada para pronóstico semanal en el mismo terreno, el error de predicción de la producción semanal.

Considerando lo anterior, de la Figura 70 se infiere que la variante *LSTM_RESID_BIO* puede ser utilizada tanto en un escenario de ventana común como de ventana amplia ya que resulta mejor que *HISTGBT_GLOBAL* en las dos métricas más relevantes, *MAE* y tasa de acierto.

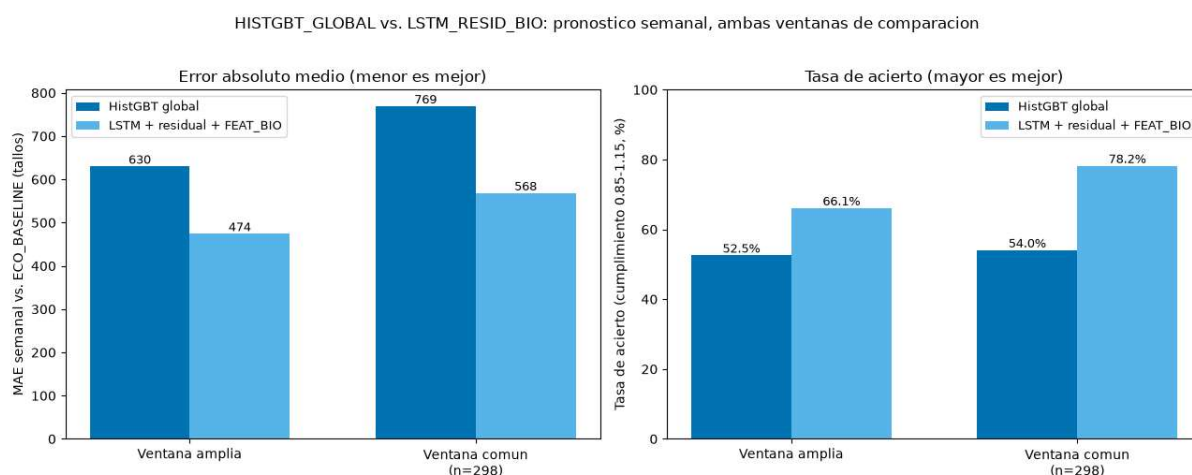


Figura 70 Comparación de métricas entre las dos mejores variantes de cada familia de modelos

Adicionalmente, es posible incluir en el análisis el nivel de cobertura empírica del intervalo conformal por cada variante, mostrado en la Figura 71, en donde se ha fijado como objetivo el 90%, es decir que al menos el 90% de las veces las predicciones caigan en dicho intervalo. Bajo este análisis, la variante *LSTM_RESID_BIO* también resulta ser la ganadora.

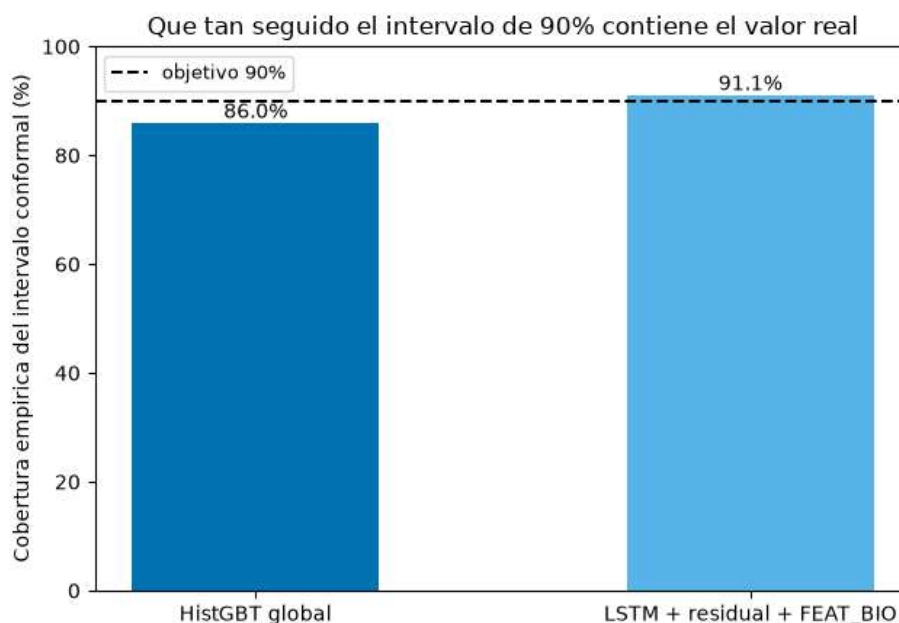


Figura 71 Cobertura empírica del intervalo conformal de las dos mejores variantes de cada familia de modelos

5.2. Recomendaciones

5.2.1. Cobertura de Datos Extendida

La limitación más concreta y accionable al final de este estudio no es de modelado, sino de volumen de datos disponibles para la comparación más estricta, la ventana común tiene solo 298 filas, concentradas en apenas dos semanas de calendario, desde 2026-04-20 hasta 2026-05-04. La razón por la que esta ventana es tan angosta es porque se construye considerando la intersección en donde las ocho variantes más el baseline, cada una de estas variantes tiene su propia ventana de fechas con predicción y no todas cubren las mismas semanas, la principal diferencia es que las variantes tabulares se validan sobre 12 folds de walk-forward, pero las variantes secuenciales solo 3 ya que demandan mayor poder de cómputo.

Para resolver este asunto se pueden adoptar una o varias de las siguientes estrategias:

1. Continuar recolectando historial operativo, cada mes adicional de datos de cosecha/pinch/clima/ECO ensancha automáticamente la ventana común, sin cambiar una sola línea de código ya que el corte no es una fecha fija, es la intersección real de

la cobertura.

2. Evaluar las variantes secuenciales sobre más folds walk-forward, no solo 3. Durante este estudio, la limitación es el poder de cómputo necesario para reentrenar una red por cada *fold*, no falta de datos crudos. Con más historial y/o con más capacidad cómputo, es posible subir de 3 a, por ejemplo, 6 u 8 fold. Esto angostaría mucho menos la ventana común y les daría más peso estadístico a los 3 contrastes de la cadena de *LSTM_RESID_BIO*.
3. Priorizar que la ventana común eventualmente cubra al menos un ciclo estacional completo, es decir, al menos un año, no solo un par de semanas. Esta es la única forma de que una comparación completa incluya tanto los períodos fáciles como los difíciles en vez de depender de qué tramo puntual quedó al final del historial disponible.
4. Extender el umbral de elegibilidad de variedades a más variedades a medida que se acumulan más datos históricos. En este estudio, 118 de las variedades totales califican; más datos por variedad amplían tanto la cobertura de las variantes globales como la posibilidad de sumar más variedades a las variantes tabulares.

5.2.2. Despliegue en Producción

5.2.2.1. Monitoreo

Para asegurar la robustez del modelo durante su despliegue en producción, se debe establecer un marco de evaluación continua sobre los nuevos flujos de datos. En particular, se recomienda calcular de manera semanal métricas de desempeño predictivo tales como el Error Absoluto Medio (MAE), el coeficiente de determinación (R^2) y la tasa de acierto o cumplimiento. Este procedimiento debe replicar la metodología analítica aplicada durante la validación inicial, con la salvedad de que los cálculos se deben aplicar sobre las observaciones de la semana más reciente en lugar del conjunto histórico de prueba.

En este contexto, es fundamental definir umbrales de alerta rigurosos para la degradación del modelo. Si la tasa de acierto experimenta una disminución sostenida durante un periodo de dos a tres semanas consecutivas, situándose por debajo de los niveles basales observados empíricamente durante la fase de prueba, esto constituye un indicador significativo de alteración en el proceso generador de datos. Dicha desviación estadística puede atribuirse a fenómenos exógenos, tales como la introducción de nuevas variedades carentes de registro histórico, modificaciones estructurales en el manejo agronómico o la incidencia de anomalías climáticas. Ante la presencia de estas señales, se hace necesaria una revisión exhaustiva de la arquitectura y los datos de entrenamiento antes de continuar fundamentando la toma de decisiones en las inferencias del modelo.

Adicionalmente, el esquema de monitoreo no debe limitarse exclusivamente al análisis de los residuos o métricas de error, sino que debe integrar una vigilancia proactiva de la explicabilidad algorítmica. El monitoreo continuo de la importancia de las variables es fundamental como mecanismo de alerta temprana. Una alteración significativa en sus ponderaciones históricas sugiere que el modelo podría estar adoptando atajos predictivos incorrectos en detrimento de los patrones causales reales del sistema. En consecuencia, identificar estas desviaciones permite anticipar la degradación estructural del algoritmo de forma proactiva, frecuentemente antes de que el deterioro sea evidente en las métricas de desempeño convencionales.

5.2.2.2. Reentrenamiento

Respecto a la metodología de actualización algorítmica en un entorno operativo, es fundamental precisar que no resulta necesario someter a un proceso de reentrenamiento a la totalidad de las arquitecturas evaluadas durante la fase de experimentación presentada en

este estudio. Por el contrario, los recursos computacionales deben focalizarse exclusivamente en la actualización del modelo específico que ha sido seleccionado y desplegado en producción. Si bien en el entorno de desarrollo el protocolo de validación *walk-forward multi-fold* dicta que evaluar y entrenar constituyeran una operación unificada para todas las variantes tabulares y secuenciales, la dinámica en producción se optimiza iterando únicamente sobre el algoritmo definitivo.

Considerando esta dinámica computacional, se sugiere establecer una cadencia de actualización estructurada, llevando a cabo un reentrenamiento completo del sistema una vez que se haya acumulado un volumen de datos de nueva generación correspondiente a un periodo de dos a tres meses. No obstante, este intervalo de tiempo deberá acortarse de forma preventiva, forzando un reentrenamiento anticipado, en caso de que los mecanismos de monitoreo establecidos en la sección anterior detecten anomalías y emitan una alerta de degradación predictiva o alteración estructural.

5.2.2.3. Estrategia de Rollout

Para la transición del modelo hacia su entorno operativo, se recomienda adoptar una estrategia de integración progresiva orientada a mitigar riesgos sistémicos. Como fase inicial, se recomienda ejecutar un *shadow deployment* por un lapso de cuatro a ocho semanas. En esta etapa, el algoritmo debe operar de forma paralela al proceso tradicional de proyección humana, sin que sus inferencias sean empleadas para la toma de decisiones. El propósito exclusivo de esta fase es validar empíricamente que el desempeño en producción se mantenga dentro de los parámetros de precisión observados durante la fase de experimentación. Una vez superada esta validación, se sugiere pasar a un piloto acotado, donde las predicciones algorítmicas se tornarán visibles, pero con carácter no vinculante, para un subconjunto específico de bloques y variedades. Esta configuración permitirá al equipo operativo contrastar

el rendimiento del modelo frente al criterio experto mediante un análisis casuístico pormenorizado.

Finalmente, la metodología contempla una fase de adopción generalizada bajo estricta supervisión, extendiendo la aplicación algorítmica a la totalidad de la operación. Es fundamental resaltar que la proyección humana debe preservarse como un mecanismo continuo de referencia y contraste empírico. Esta dualidad operativa resulta particularmente crítica para la detección de instancias atípicas donde el modelo experimenta una degradación en su precisión, tal como se evidenció en aquellas variedades donde el modelo base o ECO_BASELINE demuestra superioridad predictiva. Como principio rector del despliegue, en ninguna etapa del ciclo de vida del proyecto el algoritmo deberá sustituir íntegramente el juicio humano. El equipo de producción dispone de información cualitativa exógena, como el estado fenológico de los cultivos, dinámicas de pedidos o eventos estocásticos puntuales, que trasciende el espacio de variables integrado en la arquitectura del modelo actual.

Referencias

- Abdelmigid, H. M., Baz, M., AlZain, M. A., Al-Amri, J. F., Zaini, H. G., Abualnaja, M., . . . Alhumaidi, A. (2022). Spatiotemporal deep learning model for prediction of Taif rose phenotyping. *Agronomy*, 12(4), 807. Retrieved from <https://doi.org/10.3390/agronomy12040807>
- Abou-Mehdi-Hassani, F. (2025). Systematic literature review of smart greenhouse monitoring. 6(2), 95.
- Ahlawat, S., Nehra, V., Hasan, M., Singh, M. C., & Nehra, K. (2015). Fuzzy expert system for greenhouse rose yield prediction. *Ecology, Environment and Conservation*, 21(December Supplement), S347-S352. Retrieved from https://www.envirobiotechjournals.com/issues/article_abstract.php?aid=6437&iid=201&jid=3
- Alaoui, M. E., Chahidi, L. O., Rougui, M., Mechaqrane, A., & Allal, S. (2023, March). Evaluation of CFD and machine learning methods on predicting greenhouse microclimate parameters with the assessment of seasonality impact on machine learning performance. *Scientific African*, 19, e01578. <https://doi.org/10.1016/j.sciaf.2023.e01578>
- Albarico, J. (2023). Roses greenhouse cultivation classification using machine learning techniques. 218, 2163-2171.
- Allen, R. G. (Ed.). (2000). *Crop evapotranspiration: guidelines for computing crop water requirements* (repr ed.). Rome: Food and Agriculture Organization of the United Nations.
- Almas, H. K., Barad, A. V., & Madhuri, V. (2014). Effect of different levels of pruning and micronutrient (Fe) on growth, flowering and cut flower yield of Dutch rose (*Rosa hybrida* Linn.) cv. First Red under greenhouse condition. *The Asian Journal of Horticulture*, 9(2), 291-296. <https://doi.org/10.15740/HAS/TAJH/9.2/291-296>

- Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494-591.
- Arize. (2023). *Arize.com*. Obtenido de Mean Absolute Percentage Error (MAPE): What You Need To Know : <https://arize.com/blog-course/mean-absolute-percentage-error-mape-what-you-need-to-know/>
- Asociación de Exportadores de Flores de Ecuador. (2024). *Reporte Estadístico Anual 2024: Sector florícola*. Obtenido de https://d016f8b9-ce08-482e-82f7-0212ed37736f.filesusr.com/ugd/d0e607_4ce0f089cb2247959738716721e195cf.pdf
- Bai, S., Kolter, J. Z., & Koltun, V. (2018). *An empirical evaluation of generic convolutional and recurrent networks for sequence modeling*. arXiv.
- Banco Central del Ecuador. (2025). *Balanza Comercial por Producto*. Obtenido de https://contenido.bce.fin.ec/comercio-exterior/?bce_url=https%3A%2F%2Fcontenido.bce.fin.ec%2Fdocumentos%2FEstadisticas%2FSectorExterno%2FBalanzaPagos%2FbalanzaComercial%2Fbc202602.html&bce_path=5451%2C5452&bce_top=5451&bce_child=5452&bce_desc=SiRetorno
- Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2). <https://doi.org/10.1214/23-aos2276>
- Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021, May). Machine Learning in Agriculture: A Comprehensive Updated Review. *Sensors*, 21, 3758. <https://doi.org/10.3390/s21113758>
- Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191, 192-213.

- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281-305.
- Bhat, S. A., Huang, N.-F., Hussain, I., Bibi, F., Sajjad, U., Sultan, M., . . . Mahmoud, K. H. (2021, November). On the classification of a greenhouse environment for a rose crop based on AI-based surrogate models. *Sustainability*, 13(21), 12166. <https://doi.org/10.3390/su132112166>
- Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5.^a ed.). Wiley.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
- Brownlee, J. (27 de Abril de 2021). *MachineLearningMastery.com*. Obtenido de Histogram-Based Gradient Boosting Ensembles in Python: <https://machinelearningmastery.com/histogram-based-gradient-boosting-ensembles/>
- Bruce, A., Gedeck, P., & Bruce, P. (2020). *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*. O'Reilly Media.
- Castany, M. (18 de Abril de 2023). *QuantDemy*. Obtenido de El método Walk Forward: la guía definitiva: <https://quantdemy.com/metodo-walk-forward/>
- Castillo Santander, M. (2023). Optimización de operaciones en una florícola utilizando métodos de predicción y programación lineal. Quito, Ecuador: Universidad San Francisco de Quito. Obtenido de <http://repositorio.usfq.edu.ec/handle/23000/13266>
- Castillo, J. (2017). Modelo Matemático del Crecimiento de la Rosa Freedom a través de Redes Neuronales. 19(1), 11-20. Obtenido de <http://scielo.org.co>
- CatBoost. (2026). *CatBoost*. Obtenido de <https://catboost.ai/>

- Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109(11), 1997-2028. <https://doi.org/10.1007/s10994-020-05910-7>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc.
- Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business & Economic Statistics*, 11(2), 121-135.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (págs. 785-794). ACM. <https://doi.org/10.1145/2939672.2939785>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. <https://doi.org/10.1023/A:1022627411411>
- DataCamp. (2026). *Datacamp.com*. Obtenido de Error absoluto medio explicado: Medición de la precisión del modelo: <https://www.datacamp.com/es/tutorial/mean-absolute-error>
- Demir, S., Dedeoğlu, M., & Başayığit, L. (2024). Yield prediction models of organic oil rose farming with agricultural unmanned aerial vehicles (UAVs) images and machine learning algorithms. *Remote Sensing Applications: Society and Environment*, 33, 101131. Retrieved from <https://doi.org/10.1016/j.rsase.2023.101131>
- Dorogush, A. (2018). CatBoost: gradient boosting with categorical features support. *abs/1810.11363*. Obtenido de <http://arxiv.org/abs/1810.11363>
- Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177), 1-81.

- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232. <https://doi.org/10.1214/aos/1013203451>
- Friedman, J. H. (2001, October). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232. <https://doi.org/10.1214/aos/1013203451>
- Getachew, A., Kassa, M., & Mohammed, A. (2012). Quality of greenhouse roses (*Rosa hybrida* L.) as affected by height and stage of shoot bending and flower bud removal. *International Journal of Agricultural Research*, 7(2), 69-77. <https://doi.org/10.3923/ijar.2012.69.77>
- Gibbs, I., & Candès, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. 34, pp. 1660-1672. Curran Associates, Inc. Retrieved from <https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html>
- Gong, L., Yu, M., Jiang, S., Cutsuridis, V., & Pearson, S. (2021, July). Deep Learning Based Prediction on Greenhouse Crop Yield Combined TCN and RNN. *Sensors*, 21, 4537. <https://doi.org/10.3390/s21134537>
- Guo, S. (2020). Graphical Representation of Fourier Series from Fourier Transformation. 621-624. <https://doi.org/10.1109/ICAICA50127.2020.9182503>
- Gutiérrez Colomer, R. P., González-Real, M. M., & Baille, A. (2001). Influence of shoot bending on the photosynthetic capacity of a greenhouse rose (*Rosa hybrida*) canopy. *Acta Horticulturae*(559), 377-382. <https://doi.org/10.17660/ActaHortic.2001.559.56>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2.^a ed.). Springer.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (págs. 770-778).

- Herrera, D., Escudero-Villa, P., Cárdenas, E., Ortiz, M., & Varela-Aldás, J. (2024). Combining image classification and unmanned aerial vehicles to estimate the state of Explorer roses. *AgriEngineering*, 6(2), 1008-1021. Retrieved from <https://doi.org/10.3390/agriengineering6020058>
- Herrera, V. D., Lucero-Urresta, E., Ilvis, D. I., Mora, J. C., Chuchico, C. P., Espinel, K. A., . . . Garcia, M. V. (2024). Optimizing 'Explorer' rose production data with SVM in smart agriculture. *IFAC-PapersOnLine*, 58(8), 168-173. Retrieved from <https://doi.org/10.1016/j.ifacol.2024.08.115>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hooker, G., Mentch, L., & Zhou, S. (2021, October). Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(6), 82. <https://doi.org/10.1007/s11222-021-10057-z>
- Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3.rd ed.). OTexts.
- Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679-688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>
- Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679-688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>
- Integrated Ocean Observing System. (2014). *Manual for real-time quality control of wind data : a guide to quality control and quality assurance for coastal and oceanic wind observations*.
- Joseph, M., & Tackes, J. (2022). *Modern Time Series Forecasting with Python*. Birmingham: Packt Publishing Ltd.
- Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), 1-21.

- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* 30.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. 30, pp. 3146-3154. Curran Associates, Inc. Retrieved from <https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree>
- Lieth, J. H., & Pasian, C. C. (1991, February). A simulation model for the growth and development of flowering rose shoots. *Scientia Horticulturae*, 46(1-2), 109-128. [https://doi.org/10.1016/0304-4238\(91\)90097-I](https://doi.org/10.1016/0304-4238(91)90097-I)
- Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3.rd ed.). Wiley.
- López de Prado, M. (2018). *Advances in financial machine learning*. Wiley.
- Mancer, M., Terrissa, L. S., & Ayad, S. (2024). Machine Learning-Based Prediction of Tomato Yield in Greenhouse Environments. In C. A. Kerrache, A. E. Tahari, D. Kassimi, & C. Chakraborty (Eds.), *Proceedings of the International Conference on Emerging Intelligent Systems for Sustainable Development (ICEIS 2024)* (Vol. 184, pp. 117–128). Dordrecht: Atlantis Press International BV. https://doi.org/10.2991/978-94-6463-496-9_10
- Meyes, R., Lu, M., de Puiseau, C. W., & Meisen, T. (2019). *Ablation studies in artificial neural networks*. arXiv.
- Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter*, 3(1), 27-32.
- Montero-Manso, P., & Hyndman, R. J. (2021). Principles and algorithms for forecasting groups of time series: Locality and globality. *International Journal of Forecasting*, 37(4), 1632-1653.

Oracle. (2026). *RMSE (Error cuadrático medio)*. Obtenido de Oracle® Fusion Cloud EPM Trabajo con Planning:

https://docs.oracle.com/cloud/help/es/pbcs_common/PFUSU/insights_metrics_RMSE.htm#PFUSU-GUID-FD9381A1-81E1-4F6D-8EC4-82A6CE2A6E74

Park, H. B., & Kim, W. S. (2021, December). Validation and modification of the shoot growth model in cut roses. *Horticultural Science and Technology*, 39(6), 750-759.
<https://doi.org/10.7235/HORT.20210066>

Pasian, C. C., & Lieth, J. H. (1994, October). Prediction of flowering rose shoot development based on air temperature and thermal units. *Scientia Horticulturae*, 59(2), 131-145.
[https://doi.org/10.1016/0304-4238\(94\)90080-9](https://doi.org/10.1016/0304-4238(94)90080-9)

Paudel, D., Boogaard, H., De Wit, A., Janssen, S., Osinga, S., Pylaniadis, C., & Athanasiadis, I. N. (2021, February). Machine learning for large-scale crop yield forecasting. *Agricultural Systems*, 187, 103016. <https://doi.org/10.1016/j.agsy.2020.103016>

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31, 6638-6648.

Raviv, M., Medina, S., Wendin, A., & Lieth, J. H. (2010). Development of alternate cut-flower rose greenhouse temperature set-points based on calorimetric plant tissue evaluation. *Scientia Horticulturae*, 126(4), 454-461. <https://doi.org/10.1016/j.scienta.2010.08.011>

Sedig, V. (2019). Greenhouse Climate Optimization using Weather Forecasts and Machine Learning.

Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2021). Machine Learning Applications for Precision Agriculture: A Comprehensive Review. *IEEE Access*, 9, 4843–4873.
<https://doi.org/10.1109/ACCESS.2020.3048415>

- Shawon, S. M., Ema, F. B., Mahi, A. K., Niha, F. L., & Zubair, H. T. (2025, March). Crop yield prediction using machine learning: An extensive and systematic literature review. *Smart Agricultural Technology*, 10, 100718. <https://doi.org/10.1016/j.atech.2024.100718>
- Shin, J., Yeon, D., & Kim, W. (2023). Influence of suboptimal temperature on flower quality and floral organ development in spray-type cut rose 'Pink Shine'. *Horticulturae*, 9(8), 861. <https://doi.org/10.3390/horticulturae9080861>
- Silberbush, M., & Lieth, J. H. (2004). Nitrate and potassium uptake by greenhouse roses (*Rosa hybrida*) along successive flower-cut cycles: A model and its calibration. *Scientia Horticulturae*, 101(1-2), 127-141. <https://doi.org/10.1016/j.scienta.2003.10.009>
- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199-222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>
- Suarez León, D. (2025). Modelo machine learning para la predicción de la producción de rosas y su aplicación en la optimización logística e inventarios en la empresa Mountain Roses.
- Sujatha, K., Tejaswini, P., Laxman, R. H., & Smitha, G. R. (2024). Flower yield, physiology, nutrient acquisition and soil fertility variations in cut flower rose (*Rosa hybrida* L.) as influenced by rootstocks. *Journal of Plant Growth Regulation*, 43(11), 4230-4247. <https://doi.org/10.1007/s00344-024-11386-9>
- Sun, W., & Chang, F.-J. (2023, October). Empowering Greenhouse Cultivation: Dynamic Factors and Machine Learning Unite for Advanced Microclimate Prediction. *Water*, 15, 3548. <https://doi.org/10.3390/w15203548>
- Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4), 437-450. [https://doi.org/10.1016/S0169-2070\(00\)00065-0](https://doi.org/10.1016/S0169-2070(00)00065-0)

- Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4), 437-450. [https://doi.org/10.1016/S0169-2070\(00\)00065-0](https://doi.org/10.1016/S0169-2070(00)00065-0)
- van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., . . . Kavukcuoglu, K. (2016). *WaveNet: A generative model for raw audio*. arXiv.
- Van Klompenburg, T., Kassahun, A., & Catal, C. (2020, October). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177, 105709. <https://doi.org/10.1016/j.compag.2020.105709>
- Van Ploeg, D., & Heuvelink, E. (2005, January). Influence of sub-optimal temperature on tomato growth and yield: a review. *The Journal of Horticultural Science and Biotechnology*, 80, 652–659. <https://doi.org/10.1080/14620316.2005.11511994>
- Vasudevan, S., & Kannan, M. (2015). Effect of bending and plant growth regulators on maximizing the yield and quality of rose (*Rosa hybrida* var. Tajmahal) under greenhouse conditions. *Progressive Horticulture*, 47(2). <https://doi.org/10.5958/2249-5258.2015.00068.8>
- Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, (págs. 29-39).
- Zaffran, M., Feron, O., Goude, Y., Josse, J., & Dieuleveut, A. (2022, July). Adaptive conformal predictions for time series. *Proceedings of the 39th International Conference on Machine Learning*. 162, pp. 25834-25866. PMLR. Retrieved from <https://proceedings.mlr.press/v162/zaffran22a.html>