

Redes neuronales para la predicción de productividad de predios agrícolas y su aplicación como servicio “Agropredict”

Sebastian Lagos
sebastian.lagosgar@usm.cl

Gabriel Jara
Profesor Guía

Patricio Santander
Profesor Correferente

Resumen: La industria agrícola en Chile enfrenta incertidumbre operativa significativa debido a la variabilidad climática y la dependencia de métodos de estimación tradicionales poco precisos, lo que deriva en ineficiencias logísticas y en el uso de recursos. Este trabajo propone el desarrollo de un modelo pre-dictivo de productividad para predios frutícolas basado en Redes Neuronales Artificiales. La solución técnica se centra en el diseño, implementación y validación de una arquitectura de Machine Learning capaz de integrar y modelar las complejas interacciones no lineales entre datos climáticos históricos, registros catastrales y series temporales geoespaciales (Google Earth Engine). El modelo resultante se integrará como componente principal en el prototipo del servicio “AgroPredict”. Esta plataforma, diseñada para agrupar múltiples herramientas y componentes, demostrará una prueba de concepto escalable que busca reducir la incertidumbre en la toma de decisiones gerenciales y mejorar la eficiencia productiva en el sector.

Palabras Clave: Redes Neuronales, Agricultura de Precisión, Predicción de Rendimiento, Machine Learning, Long short-time memory.

1 Introducción

La agricultura, pilar de la economía chilena, se encuentra cada vez más expuesta a la variabilidad climática, lo que genera una alta incertidumbre operativa y estratégica para las empresas productoras. La incapacidad para anticipar con precisión y suficiente antelación los volúmenes de cosecha resulta en decisiones subóptimas, tales como la ineficiente aplicación de insumos (fertilizantes, riego) y dificultades en la planificación logística (negociación de contratos, embalaje y mano de obra). Actualmente, la gestión se basa mayormente en la experiencia histórica o en modelos estadísticos lineales, los cuales no logran capturar la complejidad no lineal de los ecosistemas productivos modernos.



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: Redes neuronales para la predicción de productividad de predios agrícolas y su aplicación como servicio "Agropredict"

Nombre del candidato(a): Sebastián Edgardo Lagos Garces

Carrera / Grado: Ingeniería en Informática

Campus: sede José Miguel Carrera **Departamento:** Electrotecnia e Informática

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Gabriel Jara Bulnes, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por (**marcar una opción**):

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 04/08/2026 **Firma:** 

Estudiante o Candidato(a):

Fecha: 04/08/2026 **Firma:** 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

En este contexto, surge la necesidad de herramientas avanzadas que permitan la anticipación de escenarios productivos. El objetivo general de esta tesina es abordar esta problemática a través de la inteligencia artificial, desarrollando un modelo robusto que integre las múltiples variables que afectan el rendimiento agrícola. Este modelo será la base del servicio "AgroPredict", una solución que busca transformar la gestión agrícola al proporcionar información predictiva, clara y accionable para perfiles gerenciales, permitiendo una mejor planificación y un uso más eficiente de los recursos.

2 Contexto

La agricultura, uno de los pilares fundamentales y junto a la minería sector clave de la economía chilena, se encuentra en una encrucijada entre la creciente demanda global de alimentos y la severa variabilidad climática (sequías, olas de calor) que azota al país. Este contexto genera una alta incertidumbre operativa y estratégica en el proceso de estimación de cosechas.

Actualmente, la gestión de la productividad se basa en dos pilares ineficientes para el entorno moderno:

- Observación y Experiencia: Métodos manuales y cualitativos que son subjetivos, difíciles de escalar y propensos a errores humanos.
- Modelos Estadísticos Simples: Aplicación de modelos lineales que no consiguen capturar las interacciones no lineales y dinámicas entre múltiples variables (clima histórico, condición del suelo, manejo del huerto) que determinan el rendimiento final.

Este escenario se traduce en un desfase crítico entre el volumen estimado y el volumen real de cosecha, que impacta negativamente la planificación.

3 Justificación del Problema: Consecuencias Operacionales y Estratégicas

La falta de una herramienta predictiva robusta tiene repercusiones directas en la eficiencia y el riesgo financiero:

Ineficiencia Operacional: Una predicción inexacta del volumen conduce a decisiones subóptimas en la aplicación de insumos (agua, fertilizantes), afectando los costos de producción y la sostenibilidad ambiental. Sin una estimación anticipada y precisa, el agricultor pierde la ventana de oportunidad para realizar ajustes correctivos durante los períodos fenológicos claves.

Riesgo Logístico y de Contrato: La incertidumbre dificulta la planificación logística a medio y largo plazo. Las empresas deben comprometerse con recursos caros (contratos de packing, embalaje, mano de obra temporal) con meses de antelación. Un error en la estimación puede resultar en pérdida de contratos, exceso de capacidad contratada o penalizaciones por incumplimiento de volumen.

El problema fundamental que aborda esta tesina no es solo la incertidumbre, sino la carencia de una herramienta digital, escalable y técnicamente avanzada capaz de transformar grandes volúmenes de datos geoespaciales y climáticos en una predicción precisa y accionable.

4 Solución General

La propuesta por este trabajo de título es el desarrollo del servicio “AgroPredict”. Este servicio busca transformar la gestión agrícola chilena al entregar una predicción robusta y escalable del rendimiento de predios frutícolas, sustentada en la Inteligencia Artificial.

El eje central de esta solución es el diseño y validación de un modelo predictivo basado en Redes Neuronales Artificiales (RNA). Estas redes están específicamente diseñadas para:

- **Integrar Múltiples Variables:** Consolidar y analizar la interacción entre datos de diversa naturaleza: series temporales climáticas, registros catastrales del predio y datos espaciales de teledetección (como índices de vegetación satelital).
- **Escalabilidad no lineal:** Capturar los patrones complejos y no lineales que rigen la productividad de la fruta, algo inaccesible para los modelos estadísticos simples.

5 Descripción del Problema Técnico

El desafío fundamental de este trabajo de título radica en la complejidad técnica asociada a la construcción, escalabilidad y precisión de un modelo predictivo orientado al sector agrícola.

Desde la perspectiva de la Ciencia de Datos, la construcción de un sistema predictivo para la agricultura presenta tres barreras técnicas fundamentales que este trabajo busca resolver:

- **Insuficiencia de Modelos Lineales ante la Complejidad Biológica:** Los modelos estadísticos tradicionales (como la regresión lineal múltiple) asumen que las variables afectan el resultado de manera aditiva y constante. Sin embargo, la fisiología del Palto responde a patrones no lineales y umbrales críticos (ej. el daño por heladas no es gradual, sino que ocurre drásticamente al cruzar cierta temperatura). Resolver este problema técnico exige la evaluación e implementación de modelos predictivos flexibles de alta dimensionalidad. Entre las alternativas idóneas a considerar se encuentran las Redes Neuronales Artificiales, debido a su capacidad para modelar funciones de mapeo complejas, capturar interacciones no lineales entre variables y detectar los puntos de saturación fisiológica que los enfoques clásicos ignoran.
- **Integración de Datos Heterogéneos y Multidimensionales:** El modelo debe ingerir y armonizar datos de naturaleza dispar. Por un lado, existen datos estáticos y tabulares que describen el predio (catastro, variedad, tipo de suelo); por otro, se manejan datos geoespaciales temporales provenientes de modelos de reanálisis climático (ERA5-Land). El desafío técnico reside en consolidar estas series temporales climáticas (grillas tridimensionales de latitud, longitud y tiempo) con los registros puntuales de producción, creando un conjunto de datos unificado y coherente para el entrenamiento.
- **Ingeniería de Características Fenológica (Alineación Temporal):** El impacto de una variable climática no es uniforme a lo largo del año calendario. Por ejemplo, el estrés térmico es crítico durante la floración, pero menos relevante en la dormancia. El reto de la Ingeniería de Características es

abandonar la agregación mensual rígida (ej. "Enero") en favor de una alineación basada en ventanas fenológicas. Esto implica diseñar algoritmos que agreguen las variables climáticas según las etapas biológicas del cultivo, asegurando que la red neuronal aprenda la relación causa-efecto en el momento exacto en que ocurre el fenómeno biológico

6 Propuesta de Solución Técnica

La solución propuesta se articula en tres etapas secuenciales de desarrollo, diseñadas para abordar la complejidad de los datos agroclimáticos y garantizar la escalabilidad del sistema.

Ingeniería y Preparación de Datos (Data Engineering): Se abordará la heterogeneidad de las fuentes mediante la consolidación de registros climáticos de reanálisis global (ERA5-Land), datos catastrales y series históricas de rendimiento. La incorporación de ERA5-Land es fundamental para este proyecto, ya que proporciona una malla continua de datos meteorológicos de alta resolución horaria, resolviendo la falta de estaciones de monitoreo físico en zonas agrícolas aisladas y garantizando la consistencia histórica necesaria para el modelo. El componente central será un proceso avanzado de Ingeniería de Características, el cual sustituirá la agregación mensual tradicional por la generación de variables basadas en ventanas fenológicas dinámicas. Esto permitirá alinear los datos climáticos con los momentos biológicos críticos del cultivo (como floración o cuaja), creando un conjunto de datos optimizado que resalta las señales de estrés ambiental para el entrenamiento de los modelos predictivos.

Diseño y Validación de la Arquitectura Neuronal: El trabajo se centrará en el diseño e implementación de arquitecturas de Aprendizaje Profundo especializadas en regresión de series temporales, específicamente modelos de Memoria a Corto y Largo Plazo (LSTM). El desempeño de este modelo se contrastará frente a redes neuronales estáticas (Perceptrón Multicapa - MLP) y modelos estadísticos de referencia (baselines). La validación aplicará un esquema estrictamente cronológico (respetando la causalidad temporal) y utilizará métricas de evaluación robustas ante valores atípicos, específicamente el Error Absoluto Medio (MAE) y la Raíz del Error Cuadrático Medio (RMSE).

7 Objetivos del Trabajo de Título

7.1 Objetivo General

Desarrollar, validar y prototipar un modelo predictivo de productividad para predios frutícolas en Chile, utilizando Redes Neuronales y un dataset integrado de datos climáticos y catastrales, para sustentar la creación del servicio "AgroPredict".

7.2 Objetivos Específicos

- **Implementación de un Flujo de Trabajo Integral:**

Diseñar e implementar un flujo de trabajo automatizado de Machine Learning que abarque desde la ingesta de datos climáticos heterogéneos hasta la generación de inferencias. Este desarrollo servirá como una Prueba de Concepto (PoC) funcional y escalable, validando la arquitectura técnica necesaria para soportar el servicio "AgroPredict".

- **Evaluación de Desempeño y Superación de la Línea Base:**

Evaluar cuantitativamente la capacidad predictiva del modelo neuronal propuesto (LSTM) utilizando métricas de error robustas: Error Absoluto Medio (MAE) y Raíz del Error Cuadrático Medio (RMSE). El objetivo es demostrar una mejora estadística significativa en la estimación de productividad (ton/ha) en comparación con los modelos de referencia tradicionales, tales como la regresión lineal múltiple.

- **Formalización de la Arquitectura y Hoja de Ruta Tecnológica:**

Documentar el diseño técnico de la solución (gestión de datos, arquitectura del modelo y despliegue) para asegurar su replicabilidad. Asimismo, definir una Hoja de Ruta Tecnológica que establezca los lineamientos para futuras iteraciones del servicio, priorizando la integración de nuevas fuentes de datos de alta resolución (satelitales) y variables biológicas complejas.

8 Metodología y plan de trabajo

La ejecución del trabajo de título se fundamenta en una metodología híbrida que integra dos marcos de trabajo complementarios, el estándar CRISP-DM (Cross-Industry Standard Process for Data Mining) para estructurar el ciclo de vida científico del modelo de datos, y los principios de Gestión Ágil para la administración operativa de las tareas de investigación.

Justificación del Enfoque Híbrido: La adopción de este esquema responde a la naturaleza intrínsecamente experimental y exploratoria del Aprendizaje Automático (Machine Learning). El desarrollo de redes neuronales (RNA) no es lineal; requiere un proceso iterativo de ensayo y error donde los resultados obtenidos en la etapa de Validación (Fase 3) frecuentemente obligan a refinar las estrategias de Preprocesamiento e Ingeniería de Características (Fase 2).

9 Fases de Desarrollo e Investigación

9.1 Fase 1: Definición del Problema y Adquisición de Datos

Esta etapa establece los cimientos del estudio, definiendo las variables físicas y consolidando la base de datos necesaria para el entrenamiento.

Definición de la Tarea de Regresión: El objetivo se formaliza como la predicción del rendimiento comercial final (ton/ha) de predios frutícolas, modelado como una función dependiente de la historia climática acumulada y las características físicas del huerto. Esta formulación abstracta permite que la metodología sea extensible a distintas familias de cultivos, aunque para efectos de validación de este trabajo se aplicará sobre un caso de estudio específico.

Adquisición de Datos Climáticos: Extracción automatizada de variables termodinámicas (temperatura del aire, punto de rocío) y edafoclimáticas (temperatura y humedad del suelo) provenientes del modelo de reanálisis ERA5-Land de Copernicus, garantizando continuidad temporal y espacial.

Recopilación de Etiquetas: Construcción del vector objetivo mediante la recopilación de registros históricos de rendimiento productivo, segmentados en bloques temporales o hitos de consolidación histórica.

9.2 Fase 2: Ingeniería de Datos y Características

Etapa centrada en la transformación de datos crudos en tensores aptos para el aprendizaje profundo, alineando la información climática con la biología del cultivo.

Preprocesamiento y Limpieza: Implementación de rutinas de limpieza para la gestión de valores atípicos y la estandarización de formatos, asegurando la calidad de las series temporales.

Ingeniería de Características Fenológica: Transformación de las series diarias mediante agregación temporal inteligente. En lugar de meses calendario, las variables se agrupan según ventanas fenológicas críticas, calculando perfiles estadísticos complejos (media, dispersión, asimetría y curtosis) para capturar eventos de estrés.

Exploración de Datos: Análisis visual y estadístico para identificar correlaciones entre las nuevas variables sintéticas y el rendimiento, validando la hipótesis de que los extremos climáticos influyen en la productividad.

9.3 Fase 3: Modelado Experimental y Validación Adaptativa

Debido a la restricción en el volumen de datos, esta fase adapta los protocolos estándar de validación para maximizar la capacidad de aprendizaje del modelo.

Implementación de Modelos de Referencia (Línea Base): Desarrollo de modelos estándar, específicamente Regresión Lineal Múltiple y Perceptrón Multicapa (MLP), para establecer un umbral mínimo de rendimiento y evidenciar la no linealidad del problema.

Diseño de Arquitectura Recurrente (LSTM): Implementación y ajuste fino de una arquitectura profunda Long Short-Term Memory (LSTM). Se configura una topología "Muchos a Uno" con capas de enmascaramiento y regularización para procesar la secuencia temporal completa y mitigar el so-sobre ajuste.

Estrategia de Validación "Maximización del Aprendizaje": En lugar de utilizar una partición de datos tradicional, se entrena con la totalidad del historial disponible bajo un régimen estocástico. Esta decisión metodológica prioriza maximizar la exposición del modelo a las dinámicas agroclimáticas históricas completas, asumiendo la contrapartida de no contar con un conjunto de prueba estático para medir la variabilidad futura. Para subsanar esta limitación y garantizar la fiabilidad arquitectónica del Modelo Agropredict, se implementa una estrategia de validación basada en Evaluación Comparativa y pruebas de estrés mediante análisis de sensibilidad. Este procedimiento evalúa directamente la respuesta del modelo ante perturbaciones climáticas simuladas,

validando la consistencia de las proyecciones a largo plazo (2024-2033) frente a modelos estadísticos de referencia y criterios agronómicos expertos.

10 Marco Conceptual y Estado del Arte

10.1 Fundamentos Agronómicos y Climáticos

Esta sección establece las bases conceptuales sobre la interacción entre la variabilidad climática y la respuesta productiva de los frutales, justificando la necesidad de herramientas predictivas avanzadas.

10.2 Agricultura de Precisión y Modelado de Cultivos

Agricultura de Precisión: La Agricultura de Precisión es una estrategia de gestión que utiliza datos temporales y espaciales para optimizar el uso de recursos y la productividad. A diferencia de la agricultura tradicional, gestiona la variabilidad intrapredial para reducir la incertidumbre operativa mediante decisiones basadas en datos cuantificables.

Fenología Vegetal y Tiempo Biológico: La fenología estudia la ocurrencia de eventos biológicos cíclicos, como la floración y maduración, y su dependencia directa de los factores climáticos. Es la base para entender el ritmo de desarrollo del cultivo más allá del calendario.

Palto (*Persea americana*): es una especie arbórea subtropical de hoja perenne originaria de Mesoamérica, cuyo fruto es una baya comestible de pulpa cremosa y alto contenido de aceite que contiene una única semilla de gran tamaño. Se caracteriza por ser un cultivo fisiológicamente sensible que no madura en el árbol (climatérico) y que depende críticamente de la estabilidad ambiental, ya que su sistema radicular superficial lo hace vulnerable al estrés hídrico y a las fluctuaciones térmicas extremas, factores que impactan directamente su potencial productivo.

Copernicus Climate Data Store (CDS): es la infraestructura digital centralizada implementada por el programa de observación de la Tierra de la Unión Europea, diseñada para fungir como un punto de acceso unificado a la información climática global. Esta plataforma proporciona acceso abierto, gratuito y estandarizado a un vasto repositorio de datos geofísicos, integrando series históricas de reanálisis, monitoreo en tiempo real y proyecciones futuras, con el objetivo de sustentar la investigación científica y la toma de decisiones estratégicas frente a la variabilidad climática.

10.3 Marco teórico: CRISP-DM

CRISP-DM por sus siglas en inglés Cross-Industry Standard Process for Data Mining es un modelo de proceso que describe el ciclo de vida completo de un proyecto de ciencia de datos. Se ha consolidado como el estándar de referencia en la industria debido a su enfoque independiente respecto a las herramientas tecnológicas y a su énfasis en la aplicación práctica de los modelos predictivos en entornos de negocio reales.

A diferencia de las metodologías de desarrollo de software tradicionales (como la cascada), CRISP-DM se define por su naturaleza iterativa y cíclica. El modelo no impone

una secuencia lineal rígida, sino que reconoce la necesidad de retroalimentación constante entre sus fases, permitiendo volver a etapas anteriores para refinar la estrategia a medida que se adquiere nuevo conocimiento sobre los datos.

El modelo estructura el proyecto en seis fases interconectadas:

- 1 **Comprensión del Negocio:** Definición de los objetivos del proyecto y requerimientos desde una perspectiva agronómica.
- 2 **Comprensión de los Datos:** Recolección inicial, descripción y exploración de la calidad de los datos climáticos y productivos.
- 3 **Preparación de los Datos:** Etapa crítica que incluye la limpieza, transformación y la ingeniería de características para construir el conjunto final de entrenamiento.
- 4 **Modelado:** Selección y aplicación de técnicas de modelado (como redes LSTM), calibración de parámetros y optimización.
- 5 **Evaluación:** Validación rigurosa del modelo no sólo en términos estadísticos, sino verificando si cumple con los objetivos de negocio planteados.
- 6 **Despliegue:** Implementación del modelo en un entorno productivo para que el usuario final pueda utilizar las predicciones.

10.4 Marco Teórico: Ingeniería de Datos y Series Temporales

Formato NetCDF (Network Common Data Form): El formato NetCDF es un estándar de software y modelo de datos autodescriptivo diseñado específicamente para el almacenamiento y acceso eficiente a información científica orientada a matrices. El formato NetCDF es multidimensional, lo que permite encapsular en un único archivo binario la complejidad espacial y temporal de las variables climáticas, integrando dimensiones de latitud, longitud y tiempo.

Serie de Tiempo: En el contexto del aprendizaje automático, una serie de tiempo se define como una secuencia de datos discretos indexados en un orden cronológico estricto. La premisa fundamental que distingue a esta estructura de los conjuntos de datos estáticos es la existencia de una dependencia temporal intrínseca, donde el valor actual de una variable se encuentra estadísticamente correlacionado con sus estados históricos previos, estableciendo así una continuidad secuencial en la evolución de la información.

Tensores: Desde una perspectiva matemática y computacional, un tensor constituye la generalización algebraica de escalares, vectores y matrices hacia un número arbitrario de dimensiones. En el ámbito del Aprendizaje Profundo, esta estructura multidimensional opera como la unidad fundamental para la representación de datos, permitiendo organizar la información más allá de las limitaciones bidimensionales tradicionales. Su arquitectura faculta el procesamiento simultáneo de múltiples características y secuencias temporales integradas dentro de una única entidad lógica coherente.

10.5 Marco Teórico: Inteligencia Artificial y Aprendizaje Profundo

Aprendizaje Supervisado y Regresión

Regresión vs. Clasificación: En el aprendizaje supervisado, la regresión es la tarea de predecir una variable de salida continua basándose en variables de entrada. A diferencia de la clasificación, que asigna categorías discretas, el objetivo en esta investigación es estimar un valor numérico específico: el rendimiento en toneladas por hectárea.

Regresión Lineal Múltiple (Baseline): Es un método estadístico que modela la relación entre una variable dependiente y múltiples variables independientes mediante una ecuación lineal. Se utiliza como modelo de referencia (baseline) para demostrar la

existencia de no linealidades en el sistema; su incapacidad para capturar puntos de saturación fisiológica o interacciones complejas justifica la necesidad de arquitecturas neuronales avanzadas.

10.6 Redes Neuronales Artificiales (ANN)

Perceptrón Multicapa (MLP): Es una clase de red neuronal artificial de alimentación hacia adelante compuesta por múltiples capas de nodos. Si bien son aproximadores universales de funciones capaces de modelar relaciones no lineales, se consideran modelos estáticos que procesan las entradas de forma simultánea, careciendo de memoria temporal para interpretar la secuencia o causalidad de los eventos históricos.

Funciones de Activación (ReLU): Son componentes matemáticos que introducen no linealidad en la red, determinando si una neurona debe activarse. La Unidad Lineal Rectificada (ReLU) es la función estándar en modelos profundos; permite el paso directo de valores positivos mientras anula los negativos, facilitando un entrenamiento eficiente y mitigando problemas de gradiente en capas ocultas.

10.7 Redes Neuronales Recurrentes (RNN) y LSTM

El Problema de la Dependencia a Largo Plazo: Las RNN estándar enfrentan la dificultad técnica del "desvanecimiento del gradiente", donde la información de los primeros pasos de tiempo se diluye a medida que se propaga por la red. Esto impide que el modelo aprenda correlaciones en secuencias largas, provocando una pérdida de "memoria" sobre eventos climáticos ocurridos al inicio de la temporada.

Long Short-Term Memory (LSTM): Es una arquitectura recurrente especializada diseñada para preservar información durante intervalos prolongados. Su estructura interna sustituye la neurona tradicional por una celda de memoria (Cell State) regulada por mecanismos de compuertas (Gates), lo que le permite retener o descartar información selectivamente. Esto faculta al modelo para capturar la Dependencia de Trayectoria (Path Dependence), entendiendo el rendimiento final como el resultado acumulativo e interdependiente de la historia climática completa.

Topología "Muchos a Uno" (Many-to-One): Configuración arquitectónica en la que el modelo ingesta una secuencia temporal de múltiples pasos (historial climático de la temporada) para producir un único vector de salida (predicción de cosecha final). Esta estructura condensa la evolución temporal de las variables en una estimación puntual de productividad.

10.8 Regularización y Optimización

Dropout (Regularización): Técnica estocástica de regularización que consiste en la desactivación aleatoria de un porcentaje de neuronas durante el proceso de entrenamiento. Su propósito es prevenir el sobreajuste (overfitting) evitando que la red "memorice" el ruido de los datos de entrenamiento, forzándola a desarrollar representaciones internas redundantes y robustas.

Descenso del Gradiente Estocástico: Estrategia de optimización donde los pesos del modelo se actualizan tras procesar cada muestra individualmente, en lugar de por lotes grandes. En contextos de escasez de datos, esta configuración maximiza la frecuencia

de aprendizaje por época e introduce un ruido estocástico beneficioso que ayuda al algoritmo a escapar de mínimos locales y mejorar la generalización

11 Análisis e Ingeniería de datos

En este apartado se explicará el proceso de obtención de los datos, su análisis y transformación pertinente en estructuras aptas para el aprendizaje profundo del modelo.

11.1 Adquisición de los datos

Para la construcción de data set de entrenamiento, se implementó un proceso de extracción de datos climáticos históricos provenientes del Servicio de Cambio Climático de Copernicus (C3S), Climate Data Store.

Específicamente, se optó por utilizar el conjunto de datos ERA5-Land post-processed daily statistics, el cual proporciona una visión consistente de la evolución de las variables atmosféricas terrestres desde 1950 hasta la actualidad.

A diferencia de la propuesta inicial basada en agregaciones espaciales en Google Earth Engine o estaciones meteorológicas locales, la selección de este data set específico mediante la API directa de Copernicus (Climate Data Store - CDS). los siguientes criterios se usaron como justificación a la elección del set de datos.

Continuidad Estructural (Reanálisis): A diferencia de los sensores remotos ópticos (afectados por la nubosidad) o las estaciones meteorológicas in-situ (expuestas a fallas operativas), ERA5-Land opera como una simulación global construida a partir de la asimilación de secuencias de datos reales. Al combinar observaciones físicas con modelos de pronóstico numérico basados en leyes dinámicas, este set de datos permite tener una visión espacial y temporal continua de toda la superficie terrestre. Esto proporciona series temporales consistentes y sin valores nulos (gaps), garantizando la estabilidad del gradiente durante el entrenamiento del Modelo Agropredict y evitando que la arquitectura predictiva pierda consistencia temporal debido a datos faltantes.

Optimización Computacional: Para optimizar el flujo de datos, se seleccionó el producto ERA5-Land post-processed daily statistics. A diferencia de la descarga masiva de series horarias brutas, este servicio permite delegar el cálculo de las estadísticas de agregación diaria (media, máximo y mínimo) al sistema CDS durante la etapa de recuperación (retrieval). Esta estrategia reduce drásticamente la carga computacional en la descarga y los requisitos de almacenamiento del pipeline ETL, facilitando el procesamiento eficiente de décadas de historia climática. Esta eficiencia en el manejo de recursos constituye un pilar fundamental para la futura escalabilidad del sistema AgroPredict.

Detección Eficiente de Umbrales Térmicos: La estrategia de extracción de datos prioriza el acceso a extremos diarios pre-calculados en origen (mínimas y máximas). Esta decisión es crítica para garantizar la detección precisa de eventos de estrés fenológico (como heladas o golpes de calor) eliminando el costo computacional que implicaría iterar localmente sobre ciclos horarios de 24 puntos. De esta forma, el modelo captura la señal de eventos extremos sin comprometer la eficiencia del procesamiento.

Incorporación de Variables Edafoclimáticas (Suelo-Atmósfera): A diferencia de los modelos limitados a variables atmosféricas, la utilización de ERA5-Land permite integrar métricas exclusivas de la interacción suelo-atmósfera, tales como la temperatura del suelo y el contenido volumétrico de agua, con una resolución de 9 km. La inclusión de estos factores dota al modelo de la capacidad para inferir directamente el estrés hídrico a nivel radicular, constituyendo un predictor de rendimiento agronómico significativamente más robusto y fisiológicamente relevante que la simple precipitación pluvial.

11.2 Adquisición de la Variable Objetivo y Definición de Bloques Productivos

La variable dependiente del estudio corresponde al Rendimiento Comercial Promedio (ton/ha) del cultivo de Palto en la zona de estudio. La construcción de este vector objetivo presentó un desafío mayor debido a la asimetría de información presente en la industria agrícola chilena.

11.3 Fuentes de Información:

A diferencia de las variables climáticas, que son de acceso abierto y resolución diaria (ERA5-Land), los registros de productividad real constituyen información estratégica y privada de las empresas agrícolas, protegida por secreto comercial. No existen repositorios públicos de acceso abierto que detallen la producción anual continua a nivel predial.

En consecuencia, para la construcción de la data set de entrenamiento, se debió recurrir a la recopilación manual de datos públicos agregados, provenientes de informes de gestión y censos productivos esporádicos disponibles en plataformas gubernamentales (como ODEPA y CIREN). Esta restricción estructural en la disponibilidad de datos obligó a definir la variable objetivo no como una serie temporal continua anual, sino como un conjunto discreto de Hitos de Consolidación Histórica.

11.4 Estructuración del Data set de Etiquetas

Derivado de la naturaleza de las fuentes públicas disponibles, la serie de tiempo de 25 años (1999-2023) se segmentó en cinco bloques productivos. Estos bloques no poseen una duración uniforme, sino que corresponden a los periodos intercensales o ventanas de reporte disponibles públicamente.

No obstante, es imperativo reconocer que esta severa restricción en el volumen de registros históricos constituye una condicionante estructural para el resto de la investigación. La escasez de muestras no solo determinó la elección de la arquitectura, sino que impondrá límites estrictos sobre las estrategias de validación y la capacidad estadística de generalización del modelo. Estas implicancias, junto con los desafíos asociados a Dimensionalidad, serán analizadas en la sección de Limitaciones del Diseño Experimental.

11.5 Selección de Variables

La adquisición de datos se automatizó mediante el desarrollo de un script en Python utilizando la librería cliente "cdsapi". El diseño del algoritmo se estructuró para descargar y procesar las siguientes variables, seleccionadas por su impacto directo en la fisiología vegetal, a continuación, se describen las variables escogidas.

Temperatura del punto de rocío a 2 metros (2m dewpoint temperature)

Unidad: K (Kelvin)

Descripción: Es la temperatura a la cual el aire, situado a 2 metros sobre la superficie de la Tierra, tendría que enfriarse para que ocurra la saturación. Es una medida de la humedad del aire. Combi-nada con la temperatura y la presión, puede utilizarse para calcular la humedad relativa. La temperatura del punto de rocío a 2m se calcula interpolando entre el nivel más bajo del modelo y la superficie de la Tierra, tomando en cuenta las condiciones atmosféricas.

Temperatura a 2 metros (2m temperature)

Unidad: K (Kelvin)

Descripción: Es la temperatura del aire a 2 metros sobre la superficie de la tierra, el mar o las aguas continentales. La temperatura a 2m se calcula interpolando entre el nivel más bajo del modelo y la superficie de la Tierra, tomando en cuenta las condiciones atmosféricas. La temperatura medida en Kelvin puede convertirse a grados Celsius (°C) restando 273.15.

Temperatura del suelo - Nivel 1 (Soil temperature level 1)

Unidad: K (Kelvin)

Descripción: Temperatura del suelo en la capa 1 (0 - 7 cm) del Sistema Integrado de Predicción (IFS) del ECMWF. La superficie se sitúa a los 0 cm. La temperatura del suelo se establece en el punto medio de cada capa, y la transferencia de calor se calcula en las interfaces entre ellas. Se asume que no existe transferencia de calor hacia el exterior desde el fondo de la capa más profunda (condición de borde adiabática). La temperatura medida en Kelvin puede convertirse a grados Celsius (°C) restando 273.15.

Contenido volumétrico de agua en el suelo - Capa 1 (Volumetric soil water layer 1)

Unidad: Metros cúbicos por metro cúbico

Descripción: Volumen de agua presente en la capa de suelo 1 (0 - 7 cm) del Sistema Integrado de Predicción (IFS) del ECMWF. La superficie se sitúa a los 0 cm. El contenido volumétrico de agua está intrínsecamente asociado con la textura del suelo (o su clasificación edafológica), la profundidad del perfil del suelo y el nivel freático (aguas subterráneas) subyacente.

Presión superficial (Surface pressure)

Unidad: Pa (Pascales)

Descripción: Presión (fuerza por unidad de área) ejercida por la atmósfera sobre la superficie terrestre, marina y de aguas continentales. Es una medida del peso de toda la columna de aire situada verticalmente sobre un punto fijo de la superficie de la Tierra. La presión superficial se utiliza frecuentemente en combinación con la temperatura para calcular la densidad del aire. Debido a la fuerte variación de la presión con la altitud, resulta difícil identificar sistemas de alta y baja presión sobre zonas montañosas utilizando esta variable; por ello, para análisis sinópticos generales se suele utilizar la presión media al nivel del mar. Las unidades de esta variable son Pascales (Pa), aunque a menudo se mide en hectopascales (hPa) o en la unidad antigua de milibares (mb).

Se priorizaron las variables que constituyen los principios básicos del desarrollo vegetal (Temperatura y Agua). Estas cinco variables representan el núcleo biofísico necesario para que el modelo aprenda. Validar la arquitectura con este set reducido permite establecer una línea base de rendimiento (baseline) con el menor costo computacional posible.

Esta restricción es una medida de diseño inicial y no una limitación permanente de la arquitectura, el diseño contempla la incorporación progresiva de variables secundarias (como Radiación Solar o Velocidad del Viento) en iteraciones futuras del servicio "AgroPredict".

Para la implementación y validación experimental del prototipo "Agropredict", se seleccionó como cultivo objetivo exclusivo el Palto (*Persea americana*). Esta decisión se fundamenta en la primacía productiva de esta especie en la Provincia de Petorca, lo que la convierte en el activo agrícola de ma-yor interés económico y estratégico para la región. Si bien la zona presenta una matriz productiva diversa, el estudio se restringió deliberadamente a una única especie siguiendo un criterio de escalabilidad incremental. Esta acotación permite validar la arquitectura en un entorno controlado (monovarietal), estableciendo una prueba de concepto sólida antes de escalar la complejidad del sistema hacia un modelo multiproducto en iteraciones futuras. No obstante, cabe destacar que el diseño e investigación subyacentes mantienen un carácter genérico; por consiguiente, la metodología desarrollada y las conclusiones científicas a las que se arribe en este trabajo son perfectamente replicables y transferibles para múltiples especies frutícolas con dinámicas de desarrollo similares.

11.6 Definición Espacial del Estudio

La delimitación geográfica del estudio se fundamentó en la ubicación estratégica de los huertos de la especie objetivo: el Palto (*Persea americana*).

Para capturar fielmente las condiciones ambientales que afectan a este cultivo, se estableció un Bounding Box centrado en la zona núcleo de producción de la Provincia de Petorca (zona históricamente con una de las mayores producciones de esta especie), Región de Valparaíso.

El cuadrante de extracción se configuró con las coordenadas:

Norte: -32.4
Oeste: -71.2
Sur: -32.5
Este: -71.1.

Se definió un área de estudio alineada con la resolución de 0.1 grados de latitud y longitud, consistente con la fuente de datos meteorológicos. Esta precisión geográfica es clave para capturar el en-torno real del cultivo, minimizando la interferencia de dinámicas atmosféricas de terrenos vecinos que no son productivos.

11.7 Estructura de Descarga y Organización de Archivos

Tras definir el subconjunto de variables climáticas y de características de suelo a utilizar, se ejecutó el protocolo de extracción automatizada, diseñando una arquitectura de almacenamiento que responde a las características estructurales del servicio de recuperación de datos del API. Debido a la organización nativa del data set, el sistema genera y descarga tres paquetes comprimidos independientes por cada mes, donde cada archivo ZIP encapsula exclusivamente una de las tres estadísticas diarias disponibles: Media, Mínima y Máxima.

Esta segmentación operativa es necesaria para gestionar correctamente la solicitud de datos al servidor, evitando conflictos en la estructura de los archivos resultantes. A su

vez, esta triplicidad resulta fundamental para el posterior análisis agronómico, ya que garantiza la disponibilidad de series temporales separadas que permiten capturar el rango térmico completo y la amplitud de las variaciones diarias, sin perder la información de la tendencia central.

Al interior de cada contenedor comprimido, los datos se encuentran en formato NetCDF. La elección de este estándar binario y multidimensional responde a su capacidad para preservar la estructura matricial original, integrando las dimensiones espaciales y temporales en un solo archivo, lo cual facilita su posterior ingesta y transformación a tensores numéricos mediante librerías de procesamiento científico. En términos de volumen operativo, esta metodología implica el procesamiento de treinta y seis archivos por cada ciclo anual. Considerando que la ventana temporal del estudio abarca 25 años continuos (desde 1999 hasta 2023), el repositorio de datos brutos se constituyó finalmente por un total de 900 paquetes comprimidos (.zip). Al interior de cada uno, la información reside en formato NetCDF, preservando la estructura matricial original para su posterior transformación a tensores.

```
# Dataset e Estadísticas
DATASET = "derived-era5-land-daily-statistics"
DAILY_STATISTICS = ["daily_mean", "daily_minimum", "daily_maximum"]

# Estructura base de la solicitud
BASE_REQUEST = {
    "variable": [
        "2m_dewpoint_temperature",
        "2m_temperature",
        "soil_temperature_level_1",
        "surface_pressure",
        "volumetric_soil_water_layer_1",
    ],
    # Incluimos hasta el día 31. CDS gestionará automáticamente los días no existentes.
    "day": [
        f"{d:02d}" for d in range(1, 32)
    ],
    "time_zone": "utc-04:00",
    "frequency": "6_hourly",
    "data_format": "netcdf",
    "download_format": "zip",
    "area": [-32.4, -71.2, -32.5, -71.1], # [Norte, Oeste, Sur, Este]
}
```

Figura 1 Definición de parámetros de descarga para el dataset ERA5-Land

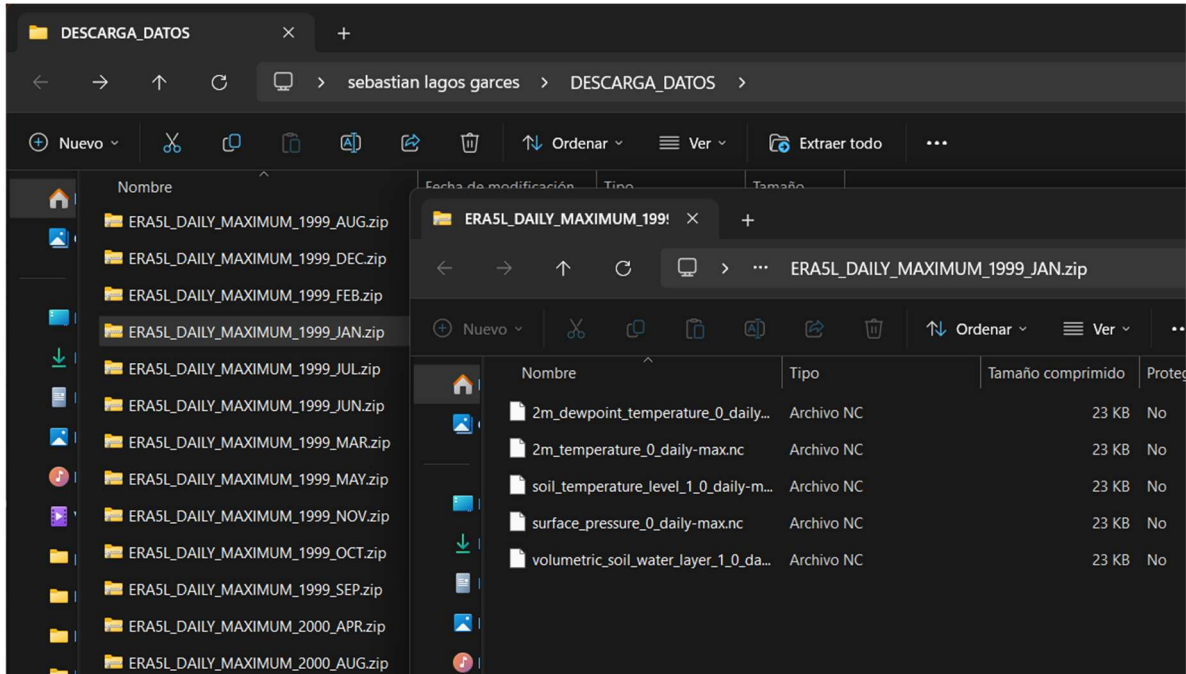


Figura 2 Carpeta de todos los paquetes descargados y sus archivos correspondientes

11.8 Preprocesamiento y Transformación de Datos (Pipeline ETL Automatizado)

La conversión de los datos climáticos crudos en una estructura optimizada para el aprendizaje profundo se orquestó mediante un pipeline ETL (Extract, Transform, Load) completamente automatizado. Este flujo de trabajo fue ejecutado por tres módulos de software secuenciales `AutoProcesamiento.py`, `Fusionador_csv.py` y `Generador_DF.py` diseñados para manejar la masividad de los registros históricos sin comprometer la memoria del sistema

11.9 De Archivos Comprimidos a Vectores Estadísticos

El procesamiento inicia con la ingesta de los archivos comprimidos .zip obtenidos de la descarga mediante la solicitud del api, los cuales contienen las series temporales en formato matricial NetCDF. Para gestionar este volumen de información de manera eficiente, se implementó el script `AutoProcesamiento.ipynb` bajo una lógica de procesamiento por lotes.

Este algoritmo realiza una descompresión dinámica de cada archivo en directorios temporales volátiles y utiliza la librería `xarray` para extraer los tensores climáticos preservando su estructura original.

En esta etapa, en lugar de reducir la información a promedios simples, el sistema calcula un perfil estadístico exhaustivo para cada mes, generando nueve métricas clave:

- medidas de tendencia central (Promedio, Mediana),
- dispersión (Desviación Estándar, Rango Inter cuartil Q25-Q75),
- extremos absolutos (Mínimo, Máximo)
- métricas de forma (Asimetría y Curtosis).

La decisión de generar un perfil estadístico de nueve dimensiones en lugar de utilizar promedios mensuales simples responde a la naturaleza no lineal de la respuesta biológica de los cultivos. Desde la comprensión de las variables climáticas, comprimir la variabilidad de 30 días en un solo valor (la media) provoca una pérdida crítica de información que impide al modelo detectar situaciones de estrés.

Esta expansión de características se justifica en tres niveles:

Detección de umbrales críticos (Mínimos y Máximos): En fisiología vegetal, el daño no es gradual, sino que ocurre al cruzar umbrales específicos.

Un mes puede tener una temperatura media de 15°C resultado de días estables de 15°C, o de una alternancia violenta entre -2°C y 32°C. Para el promedio, ambos escenarios son idénticos. Al incluir los extremos Absolutos, la red neuronal recibe la señal explícita de si se violaron los límites de congelación (daño por heladas) o de cierre estomático (estrés térmico), permitiéndole penalizar la predicción de rendimiento acorde a la realidad biológica.

Captura de la Volatilidad (Dispersión): La estabilidad climática es clave para procesos como la cuajada y el llenado de fruto. Métricas como la Desviación Estándar y el Rango Inter cuartil (IQR) informan al modelo sobre la "turbulencia" del mes. Un rango amplio indica un mes inestable, lo que fuerza a la planta a gastar energía en adaptación en lugar de producción, un patrón que la red LSTM puede aprender a asociar con menores rendimientos.

11.10 Justificación Técnica de las Métricas de Forma

La inclusión de métricas de forma responde a la necesidad agronómica de capturar eventos que los promedios tienden a ocultar.

la Asimetría (Skewness) permite identificar la dirección de las anomalías climáticas. Un valor positivo podría revelar por ejemplo picos de calor inusuales, mientras que uno negativo señala eventos de frío extremo, permitiendo a la red neuronal diferenciar la naturaleza del estrés en las variables.

Por otro lado, la Curtosis (Kurtosis) mide la presencia de "colas pesadas" o eventos extremos puntuales. En el contexto agrícola, una curtosis alta alerta sobre meses aparentemente estables que esconden eventos violentos de corta duración, como una helada negra de un solo día, los cuales suelen ser los determinantes del daño productivo.

```
def generar_resumen_estadistico(df_final: pd.DataFrame, nombre_zip: str, directorio_salida: Path) -> None:
    """
    Realiza el análisis estadístico y guarda el resultado como CSV.
    """
    print(" [INFO] Generando análisis estadístico...")

    df_datos = df_final.iloc[:, 0:]
    df_resumen_completo = df_datos.describe().T

    # Añadir Skewness y Kurtosis
    df_resumen_completo['Asimetría (Skew)'] = df_datos.skew()
    df_resumen_completo['Curtosis (Kurtosis)'] = df_datos.kurtosis()

    # Renombrar columnas
    df_resumen_completo = df_resumen_completo.rename(columns={
        'mean': 'Promedio',
        'std': 'Desviación Estándar',
        'min': 'Mínimo',
        'max': 'Máximo',
        '25%': 'Q25',
        '50%': 'Mediana (Q50)',
        '75%': 'Q75',
    })

    # Columnas ordenadas (excluyendo 'count'/'Conteo', como se solicitó)
    columnas_ordenadas = ['Mínimo', 'Máximo', 'Promedio', 'Mediana (Q50)',
        'Desviación Estándar', 'Q25', 'Q75',
        'Asimetría (Skew)', 'Curtosis (Kurtosis)']

    df_resumen_final = df_resumen_completo[columnas_ordenadas]

    # Guardar el resumen con el nombre del ZIP
    nombre_csv_salida = directorio_salida / f"{nombre_zip}_resumen.csv"
```

Figura 3 Fragmento de código del módulo de Auto Procesamiento encargado de generar el resumen estadístico mensual

```
=====
INICIO DEL PROCESAMIENTO POR LOTES: 180 archivos .zip encontrados.
Resultados guardados en: C:\Users\lagos\OneDrive\Desktop\AUTOPROCESAMIENTO\resumenes_csv_finales
=====

[1/180] Procesando ZIP: ERA5L_DAILY_MAXIMUM_2014_APR.zip
[OK] Archivo 'ERA5L_DAILY_MAXIMUM_2014_APR.zip' descomprimido en carpeta temporal.
[INFO] Encontrados 5 archivos .nc para procesar.
[OK] Variable 'd2m' extraída y limpia.
[OK] Variable 't2m' extraída y limpia.
[OK] Variable 'stl1' extraída y limpia.
[OK] Variable 'sp' extraída y limpia.
[OK] Variable 'swvl1' extraída y limpia.
[OK] DataFrame final unido en memoria para ERA5L_DAILY_MAXIMUM_2014_APR.
[INFO] Generando análisis estadístico...
=====
✅ RESUMEN COMPLETO GUARDADO EN: ERA5L_DAILY_MAXIMUM_2014_APR_resumen.csv
```

Figura 4 Registro de ejecución (log) del módulo de Auto Procesamiento

Se puede observar en la figura 4 como se realiza la extracción exitosa del archivo .zip, la extracción de las variables .nc y su conversión en un archivo .CSV.

jupyter ERA5L_DAILY_MAXIMUM_1999_APR_resumen.csv Last Checkpoint: last month

File Edit View Settings Help

Delimiter: .

		Mínimo	Máximo	Promedio	Mediana (Q50)	Desviación Estándar	Q25	Q75
1	d2m	278.820068359375	287.16278076171875	284.6029052734375	284.9562530517578	1.4557297229766846	283.658447265625	285.680206298828
2	t2m	287.745849609375	300.2353820800781	294.0091857910156	293.5424346923828	2.255267858505249	292.5826644897461	295.325172424316
3	stt1	294.40960693359375	305.1739501953125	300.49041748046875	300.5717468261719	2.156745433807373	299.26416015625	302.0570678710937
4	sp	93156.5	98213.375	96007.703125	96324.9375	1555.9708251953125	95181.703125	97062.984375
5	swvl1	0.1847991943359375	0.4305877685546875	0.25259044766426086	0.233306884765625	0.05757914483547211	0.21204757690429688	0.27914428710937

Figura 5 transformación de archivo .nc a .csv con nuevas variables estadísticas diarias máximas para el mes de abril 1999

11.11 Consolidación y Aplanamiento Vectorial (Módulos "Fusionador" y "Generador")

La etapa final de integración de datos es ejecutada secuencialmente por los módulos de software `Fusionador_csv.py` y `Generador_DF.py`. El procedimiento inicia con una operación de fusión vertical, diseñada para consolidar la información que inicialmente se encuentra dispersa en archivos inde-pendientes. En esta instancia, los registros correspondientes a las tres perspectivas diarias descargadas (Mínima, Media y Máxima) se agrupan y unifican en una sola estructura tabular maestra por cada mes, centralizando así toda la variabilidad climática del periodo en una entidad lógica coherente.

Posteriormente, para adaptar estos datos a la arquitectura de entrada de la red neuronal, se aplica una técnica de transformación estructural conocida como vectorización o flattening.

Este proceso convierte la tabla mensual multidimensional en un vector lineal continuo de características. La densidad de información de este vector resultante es producto de una combinatoria exhaustiva de todos los factores monitoreados: para cada una de las cinco variables biofísicas seleccionadas, se integran sus tres perspectivas de observación diaria, y a cada una de estas combinaciones se le calculan los nueve descriptores estadísticos y de forma previamente definidos.

Como consecuencia de este diseño matricial, cada mes de historia climática queda representado matemáticamente por un vector denso constituido por 135 características únicas, encapsulando la totalidad de la señal ambiental con alta resolución estadística antes de ingresar a la secuencia temporal.

```
Buscando archivos consolidados en: ./tablas_completas_por_mes_f/
Total de archivos encontrados: 300
Estructura descubierta: 5 variables x 27 columnas consolidadas = 135 features por mes.
-----
PROCESO FINALIZADO CON ÉXITO.
El DataFrame Base Cronológico (2D) se guardó en: ./df_base_cronologico.csv
Forma final del DataFrame (Meses x Características): (300, 136)
Total de características generadas: 135
-----

=== MAPEO DINÁMICO DE CARACTERÍSTICAS (feature_XX) ===
NOTA: El orden es (Variable Climática) primero, luego (Columna Consolidada)

feature_00: [D2M] - Valor para la columna consolidada: 'MIN_Mínimo'
feature_01: [D2M] - Valor para la columna consolidada: 'MIN_Máximo'
feature_02: [D2M] - Valor para la columna consolidada: 'MIN_Promedio'
feature_03: [D2M] - Valor para la columna consolidada: 'MIN_Mediana (Q50)'
feature_04: [D2M] - Valor para la columna consolidada: 'MIN_Desviación Estándar'
feature_05: [D2M] - Valor para la columna consolidada: 'MIN_Q25'
feature_06: [D2M] - Valor para la columna consolidada: 'MIN_Q75'
feature_07: [D2M] - Valor para la columna consolidada: 'MIN_Asimetría (Skew)'
feature_08: [D2M] - Valor para la columna consolidada: 'MIN_Curtosis (Kurtosis)'
feature_09: [D2M] - Valor para la columna consolidada: 'MEAN_Mínimo'
feature_10: [D2M] - Valor para la columna consolidada: 'MEAN_Máximo'
feature_11: [D2M] - Valor para la columna consolidada: 'MEAN_Promedio'
feature_12: [D2M] - Valor para la columna consolidada: 'MEAN_Mediana (Q50)'
feature_13: [D2M] - Valor para la columna consolidada: 'MEAN_Desviación Estándar'
feature_14: [D2M] - Valor para la columna consolidada: 'MEAN_Q25'
feature_15: [D2M] - Valor para la columna consolidada: 'MEAN_Q75'
feature_16: [D2M] - Valor para la columna consolidada: 'MEAN_Asimetría (Skew)'
feature_17: [D2M] - Valor para la columna consolidada: 'MEAN_Curtosis (Kurtosis)'
feature_18: [D2M] - Valor para la columna consolidada: 'MAX_Mínimo'
feature_19: [D2M] - Valor para la columna consolidada: 'MAX_Máximo'
feature_20: [D2M] - Valor para la columna consolidada: 'MAX_Promedio'
feature_21: [D2M] - Valor para la columna consolidada: 'MAX_Mediana (Q50)'
feature_22: [D2M] - Valor para la columna consolidada: 'MAX_Desviación Estándar'
feature_23: [D2M] - Valor para la columna consolidada: 'MAX_Q25'
feature_24: [D2M] - Valor para la columna consolidada: 'MAX_Q75'
feature_25: [D2M] - Valor para la columna consolidada: 'MAX_Asimetría (Skew)'
feature_26: [D2M] - Valor para la columna consolidada: 'MAX_Curtosis (Kurtosis)'
feature_27: [T2M] - Valor para la columna consolidada: 'MIN_Mínimo'
```

Figura 6 output de script Generador DF

Se puede apreciar en la imagen como el script realiza el aplanamiento correcto de la variable "D2M" para cada estadística diaria.

	Date	feature_00	feature_01	feature_02	feature_03	feature_04	feature_05
1	1999-01-01	282.0132141113281	287.34234619140625	285.0533752441406	284.9986267089844	1.1227641105651855	284.1763610839844
2	1999-02-01	279.690673828125	288.7528991699219	285.9590148925781	286.25196838378906	2.0017659664154053	284.8326187133789
3	1999-03-01	277.31707763671875	288.7469177246094	284.0157775878906	284.17774963378906	2.1524782180786133	282.69285583496094
4	1999-04-01	269.853759765625	285.7685546875	281.71099853515625	282.17901611328125	2.8514420986175537	280.6376342773437
5	1999-05-01	272.05908203125	285.32049560546875	281.005859375	281.1816711425781	2.69463324546814	279.63153076171875
6	1999-06-01	268.4417724609375	283.8860168457031	278.6275939941406	279.33673095703125	3.1035375595092773	276.5426025390625
7	1999-07-01	266.113525390625	281.3168334960937	275.087646484375	275.3759155273437	3.4647130966186523	272.6688461303711
8	1999-08-01	268.2205810546875	282.9522705078125	278.3536071777344	278.9647521972656	3.201211214065552	276.67539978027344
9	1999-09-01	271.67724609375	285.1470947265625	279.5221252441406	280.3819274902344	3.1348326206207275	276.9465026855469
10	1999-10-01	277.244873046875	285.9320983886719	282.5061340332031	282.3070373535156	1.7665032148361206	281.15840911865234
11	1999-11-01	279.25103759765625	285.6275939941406	283.0483093261719	283.44944763183594	1.6656484603881836	282.1021423339844
12	1999-12-01	277.3538818359375	286.4159545898437	283.11529541015625	283.46702575683594	1.7781449556350708	282.3386764526367
13	2000-01-01	279.1947021484375	286.7247009277344	282.9222106933594	282.8461608886719	1.4918882846832275	281.7589492797852
14	2000-02-01	275.7011108398437	287.3017883300781	283.5058288574219	283.8224182128906	2.472888708114624	282.17923736572266
15	2000-03-01	274.7353515625	286.4804992675781	283.07208251953125	284.0706481933594	2.6884305477142334	282.3504638671875
16	2000-04-01	279.257568359375	285.71649169921875	283.2574768066406	283.69921875	1.5235601663589478	282.3390579223633
17	2000-05-01	265.5015869140625	283.5482177734375	278.8950500488281	279.5466613769531	3.5417613983154297	276.9879150390625
18	2000-06-01	273.14404296875	282.860595703125	278.60211181640625	278.75665283203125	2.5308756828308105	277.0750427246094
19	2000-07-01	263.5865478515625	281.2182312011719	273.2222351074219	275.0111694335937	3.664708375930786	271.76593017578125
20	2000-08-01	270.37890625	283.0581359863281	278.4059753417969	278.78564453125	2.804781436920166	276.28257751464844
21	2000-09-01	273.4981689453125	284.5936889648437	279.6951599121094	279.66168212890625	2.6902916431427	277.9933242797852
22	2000-10-01	277.28741455078125	285.93194580078125	282.1200256347656	282.587646484375	2.0445895195007324	280.60040283203125
23	2000-11-01	275.7384033203125	286.593505859375	282.8268127441406	283.33290100097656	2.4441850185394287	281.1025161743164
24	2000-12-01	279.1729736328125	287.5077514648437	283.8672180175781	284.0670928955078	1.8277279138565063	282.4783630371094
25	2001-01-01	277.8724365234375	287.7447509765625	284.686279296875	285.1079406738281	1.9434270858764648	283.3861389160156
26	2001-02-01	282.5340270996094	289.03094482421875	286.4462280273437	286.5147399902344	1.6852411031723022	285.12530517578125
27	2001-03-01	278.8155517578125	287.80401611328125	285.1409606933594	285.5204772949219	1.9203596115112305	284.3361282348633
28	2001-04-01	268.5245361328125	287.87408447265625	281.9052124023437	283.2271423339844	4.332112312316895	280.0554962158203
29	2001-05-01	272.57666015625	285.561279296875	279.7866516113281	279.75115966796875	2.7564260959625244	278.0138397216797
30	2001-06-01	263.94287109375	283.25201416015625	275.53289794921875	275.9263610839844	4.229854583740234	273.2557067871094

Figura 7 Resultado del proceso de consolidación y vectorización temporal

Salida del módulo Generador DF.py, mostrando la integración exitosa de los 25 años de historia climática dispersa en una única estructura tabular ordenada estrictamente por la columna Date. En la imagen se aprecian las primeras 5 de las 135 características aplanadas (feature 00 a feature 04) junto con la continuidad temporal del índice.

11.12 Metodología de Ventaneo Dinámico

En contraste con los enfoques tradicionales de series de tiempo basados en ventanas de calendario fijas, se implementó una estrategia de ventanas de tamaño dinámico. Esta decisión de diseño responde fundamentalmente a la plasticidad fenológica del cultivo, donde los ciclos de formación del fruto no se rigen por fechas de calendario rígidas, sino por la acumulación de condiciones climáticas.

El periodo de desarrollo de la cosecha es intrínsecamente variable: las fluctuaciones en la temperatura y la precipitación pueden acelerar o retrasar la maduración, desplazando las fechas de cosecha año tras año. Por lo tanto, imponer una ventana estática (ej. "de enero a diciembre") podría fragmentar etapas críticas del desarrollo o incluir ruido de ciclos adyacentes. La estrategia ajusta la ventana de observación para capturar el "año biológico" real, asegurando que la red neuronal procese la historia climática exacta que dio origen a cada evento de cosecha, respetando su duración natural (ya sea más corta o extensa según el clima de esa temporada).

Complementariamente, esta arquitectura flexible aborda un desafío práctico de escalabilidad: la disponibilidad de datos de rendimiento no siempre es uniforme ni estrictamente anual. Al desacoplar el modelo de una frecuencia temporal rígida, el

sistema se vuelve capaz de procesar registros de rendimiento irregulares o esporádicos. Esto permite escalar la solución a diferentes predios o realidades operativas sin necesidad de imputar datos faltantes, trabajando simplemente con la ventana de tiempo disponible asociada a cada registro de producción existente.

Para gestionar estas secuencias de entrada de longitud variable, la arquitectura LSTM incorpora capas de enmascaramiento. Este mecanismo permite al modelo aprender eficazmente de la heterogeneidad temporal sin la introducción de rellenos artificiales (padding) que distorsionarían la señal climática acumulada.

	Rendimiento	Fecha_Inicio_Bloque	Fecha_Fin_Bloque
1	9.4	1999-01-01	2002-09-04
2	9.3	2002-09-05	2008-09-03
3	10.0	2008-09-04	2014-10-30
4	11.7	2014-10-31	2017-11-07
5	10.2	2017-11-08	2023-10-02

Figura 8 Ventanas de tiempo y su correspondiente Rendimiento

11.13 Diseño de la arquitectura de Red Neuronal

Dado el contexto del problema se selecciona una arquitectura basada en memoria a corto-largo plazo (Long Short-Term Memory) ya que responde a la necesidad crítica de modelar la productividad agrícola como una Serie de Tiempo Dinámica y no como un conjunto estático de características.

Mientras que los modelos tradicionales (como Regresión Lineal o Perceptrón Multicapa) procesan las entradas de forma simultánea e independiente (donde el orden de los factores no altera el resultado), el cultivo del Palto es intrínsecamente un proceso secuencial dependiente de la trayectoria histórica.

Esta distinción se justifica en dos puntos clave.

Preservación de la Estructura Secuencial: A diferencia de los modelos tradicionales que interpretan el conjunto de datos como un vector de características independientes y permutables ("Bolsa de Características"), la arquitectura LSTM procesa la información respetando estrictamente su tipología temporal. Este enfoque permite que el algoritmo asigne un significado y un peso diferencial a una misma variable biofísica en función exclusiva de su posición dentro de la secuencia, reconociendo que la relevancia de un dato de entrada está intrínsecamente condicionada por su ubicación cronológica y no solo por su magnitud numérica.

Modelado de Causalidad Acumulativa y Dependencia de Trayectoria: El fenómeno productivo se conceptualiza no como una suma lineal de eventos discretos, sino como el resultado de una integración progresiva de estados a lo largo del tiempo. Mediante su mecanismo interno de Celda de Memoria (Cell State), la LSTM tiene la capacidad de retener y transportar información latente a través de la ventana temporal. Esto faculta al modelo para capturar relaciones de dependencia de trayectoria (path dependence), donde las perturbaciones ocurridas en las etapas iniciales de la serie modulan la

respuesta del sistema ante estímulos futuros, replicando la naturaleza acumulativa e interdependiente del desarrollo biológico que un análisis estático es incapaz de representar.

En resumen, se seleccionó LSTM porque permite preservar la integridad cronológica de los datos, asumiendo que la secuencia de los eventos es tan importante como la magnitud de estos, a continuación, se describe y la topología y la justificación de cada capa que conformarán el modelo.

11.14 Topología del Modelo (Stacked LSTM)

La arquitectura implementada corresponde a una topología de tipo "Muchos a Uno", diseñada para procesar una secuencia histórica de datos climáticos y producir un único valor estimado de rendimiento (ton/ha). El diseño consta de cinco capas:

1. **Capa de Enmascaramiento (Masking Layer):** El punto de entrada del modelo incorpora un mecanismo de normalización estructural que aplica técnicas de relleno con valores neutros (padding). Su función es procesar los ciclos fenológicos de duración variable, estandarizando las entradas en estructuras rectangulares uniformes.
2. **Primera Capa LSTM:** El primer bloque de procesamiento profundo consta de sesenta y cuatro unidades recurrentes configuradas para preservar la dimensión temporal completa de los datos. Esta capa opera como un extractor de características de bajo nivel. Su objetivo es procesar la secuencia paso a paso y transmitir la totalidad de la historia temporal a la siguiente capa. La elección de este número de unidades busca un equilibrio técnico entre la capacidad de representación (poder modelar patrones complejos) y la capacidad computacional, evitando una red excesivamente ancha que incremente el número de parámetros entrenables más allá de lo sostenible para un entorno de datos limitados.
3. **Capa de Regularización (Dropout):** Entre las capas recurrentes se insertó un mecanismo de regularización estructural con una tasa de desconexión del veinte por ciento. Ante la restricción severa de muestras históricas disponibles, el riesgo de que la red "memorice" los datos en lugar de aprender de ellos es extremo. Esta capa mitiga dicho riesgo apagando aleatoriamente neuronas durante el entrenamiento, lo que fuerza al modelo a desarrollar representaciones distribuidas y robustas, impidiendo que la predicción dependa excesivamente de nodos específicos.
4. **Segunda Capa LSTM:** El segundo nivel recurrente reduce la complejidad a treinta y dos unidades y está configurado para condensar la secuencia temporal. Esta arquitectura apilada permite descomponer la complejidad de la señal climática en niveles de abstracción. Mientras la primera capa analiza la secuencia completa, esta segunda capa integra esas características para generar patrones. Su función final es comprimir toda la historia climática procesada en un único vector de contexto semántico que resume la temporada completa.

5. Cabezal de Regresión (Dense Output): La etapa final consta de una capa densa intermedia con activación no lineal, seguida de una proyección lineal unitaria que entrega la estimación del rendimiento.

12 Restricciones en la Disponibilidad de Datos Históricos (Escasez de Registros de Rendimiento)

La principal limitante del diseño experimental radica en el reducido volumen de muestras históricas de entrenamiento (5 ciclos productivos). Esta restricción no es una decisión metodológica arbitraria, sino la consecuencia directa de una barrera en la disponibilidad de la información.

Mientras que los datos climáticos son abundantes y de acceso público gracias a servicios satelitales globales como Copernicus, los datos de rendimiento real son propiedad privada, dispersa y frecuentemente no digitalizada. La obtención de series históricas largas y fidedignas de producción (kg/ha) enfrenta dos obstáculos críticos:

Inexistencia de Datos Públicos: A diferencia de otros sectores, no existen repositorios abiertos o censos anuales públicos que vinculen rendimiento georreferenciado a nivel predial con variables climáticas.

Hermetismo y Falta de Estandarización Privada: La recolección de datos privados depende de la colaboración de productores individuales, cuyos registros históricos suelen ser fragmentarios, inconsistentes o celosamente guardados por razones competitivas.

En consecuencia, la capacidad del modelo para generalizar situaciones climáticas atípicas se encuentra condicionada por esta "ceguera histórica", obligando al sistema a aprender patrones complejos a partir de una ventana de observación estrecha.

13 Estrategia de Entrenamiento y Validación Adaptativa

Para el proceso de aprendizaje de la arquitectura de memoria secuencial apilada descrita anteriormente, se definió una estrategia de entrenamiento diseñada específicamente para mitigar las limitaciones impuestas por la escasez de registros históricos.

En el ámbito de las redes neuronales aplicadas a series temporales, el estándar para la validación es la estrategia de Validación de Ventana Deslizante (o validación progresiva). Este método se fundamenta en el respeto estricto a la cronología de los eventos: el modelo se entrena utilizando datos pasados y se prueba prediciendo el futuro inmediato mediante ventanas temporales iterativas. Por ejemplo, se entrena con los años uno al cinco para predecir el seis, luego se incorpora el seis al entrenamiento para predecir el siete, y así sucesivamente. Este esquema asegura que nunca se utilice información futura para realizar predicciones (preservando la causalidad temporal) y simula fielmente las condiciones reales de un entorno de producción.

Sin embargo, su aplicación rigurosa requiere un historial de datos extenso. Dado que este estudio cuenta con un volumen crítico de registros, aplicar un particionamiento tradicional como la ventana deslizante o la separación estática de prueba resultaba inviable tanto técnica como estadísticamente. Segregar un porcentaje significativo de los datos (entre el 20% y el 30%) habría impedido que el modelo aprendiera patrones climáticos extremos esenciales para su capacidad de generalización.

Por consiguiente, se optó por una estrategia de ingeniería denominada "Maximización del Aprendizaje". En lugar de reservar datos para un conjunto de validación en tiempo real, se utilizó la totalidad del historial disponible para el ajuste de los pesos sinápticos.

Esta decisión prioriza que la red neuronal capture la variabilidad climática completa del conjunto de datos. Para compensar la falta de validación interna y evitar la memorización pura (exceso de ajuste), el control de calidad se trasladó a dos frentes de regularización:

- **Regularización Estructural:** Se implementaron capas de desactivación aleatoria (Dropout) y un régimen de aprendizaje estocástico con un tamaño de lote unitario. Estas técnicas introducen un ruido controlado durante la optimización, lo que impide que la red dependa excesivamente de neuronas específicas.
- **Validación Posterior al Entrenamiento:** La verificación de la coherencia y robustez del modelo se realiza mediante pruebas de inferencia posteriores, tales como el análisis de escenarios y la evaluación de volatilidad, en lugar de monitorear el error durante el proceso de descenso del gradiente.

14 Justificación del Régimen Estocástico

Se configuró un tamaño de lote unitario (batch size de tamaño 1) para operar bajo un régimen de aprendizaje estocástico puro. Esta decisión es crítica en este entorno de datos limitados, ya que el uso de lotes estándar resultaría en una única actualización de pesos por época, ralentizando el aprendizaje.

Al procesar muestra a muestra, se fuerza a la red a recalcular el error y ajustar sus parámetros cinco veces por época, maximizando la frecuencia de convergencia. Además, la varianza introducida por este cálculo individual inyecta un ruido estocástico beneficioso que actúa como regularizador, ayudando al modelo a escapar de mínimos locales y evitando el sobreajuste determinista.

15 Validación por Evaluación Comparativa y Análisis de Comportamiento

Considerando la naturaleza secuencial de los datos y la limitación en la cantidad de registros, se descartaron los métodos de validación que fragmentan la serie temporal, tales como la validación cruzada, debido a que infringen la causalidad cronológica intrínseca del fenómeno productivo. Como alternativa, se implementó un protocolo de validación basado en una evaluación comparativa.

Bajo este esquema, la validez de la arquitectura propuesta de memoria secuencial no se demuestra mediante un conjunto de prueba aislado, sino a través de su superioridad relativa y robustez de comportamiento frente a modelos de referencia estándar.

La estrategia metodológica consiste en someter simultáneamente al modelo propuesto, a una regresión lineal y a una red neuronal de perceptrón multicapa a los mismos escenarios de predicción futura para el periodo comprendido entre los años 2024 y 2033, con el fin de evaluar la coherencia de sus respuestas. La hipótesis de validación establece que, si la arquitectura de memoria recurrente ha asimilado correctamente la dinámica del sistema en lugar de simplemente memorizar el ruido debe exhibir comportamientos que los modelos básicos no logran replicar.

Específicamente, el modelo debe demostrar estabilidad ante las perturbaciones, manteniendo proyecciones constantes y sin fluctuaciones erráticas frente a variaciones

aleatorias del clima simulado, diferenciándose así de la inestabilidad de la red neuronal simple. Asimismo, debe evidenciar coherencia agronómica, evitando proyectar tendencias constantes hacia el colapso productivo como ocurre con la regresión lineal, y en su lugar, capturar la capacidad de resiliencia y los ciclos de recuperación naturales propios del cultivo. Para operativizar y medir esta coherencia agronómica, se implementará un protocolo de pruebas de estrés donde se contrastará el gradiente de las predicciones frente a las curvas teóricas de desarrollo del cultivo; el cumplimiento de este criterio se evaluará mediante la tasa de convergencia del modelo hacia valores de rendimiento biológicamente viables tras la simulación de eventos climáticos extremos.

16 Descripción de los Modelos de Referencia

Para establecer un punto de comparación técnico, se implementaron dos modelos de control:

- **Regresión Lineal Múltiple:** Se utilizó un enfoque clásico de mínimos cuadrados ordinarios. Este modelo intenta explicar el rendimiento productivo como una suma ponderada simple de las variables climáticas. Asume una relación rígida y proporcional. Para la regresión, un aumento de temperatura siempre suma o resta en la misma medida, siendo incapaz de detectar puntos de saturación fisiológica, es decir, situaciones donde el calor excesivo deja de ser beneficioso y pasa a ser dañino para la planta.
- **Perceptrón Multicapa (Referencia No Lineal):** Se implementó una Red Neuronal Artificial clásica de tipo "alimentación hacia adelante", compuesta por capas densas de neuronas conectadas entre sí. A diferencia de la regresión, este modelo puede representar curvas y relaciones complejas. Es un modelo estático y sin memoria. Procesa la información climática como un conjunto desordenado de datos, ignorando la secuencia temporal de los eventos. El modelo no distingue si el estrés hídrico ocurrió antes o después de la floración, perdiendo el contexto de causalidad acumulativa.

```
# B. MLP (Simple)
tf.random.set_seed(42)
mlp_model = Sequential([
    Dense(64, activation='relu', input_dim=X_scaled.shape[1]),
    Dense(32, activation='relu'),
    Dense(1, activation='linear')
])
```

Figura 9 Capas del Modelo MLP

17 Métricas de Evaluación de Desempeño

Para cuantificar la precisión de los modelos y diagnosticar su capacidad de generalización, se seleccionaron dos métricas estadísticas estándar en problemas de regresión:

- **Error Absoluto Medio (MAE - Mean Absolute Error):** El MAE representa la magnitud promedio de los errores en un conjunto de predicciones, sin considerar su dirección (positiva o negativa). Matemáticamente, se define como el promedio de las diferencias absolutas entre los valores predichos y los valores observados.

Esta métrica proporciona una evaluación lineal de la precisión del modelo. Al no elevar los residuos al cuadrado, otorga un peso proporcional a todas las desviaciones, lo que la convierte en un indicador robusto y directamente interpretable en las mismas unidades físicas de la variable objetivo (ton/ha).

Un MAE bajo indica que las predicciones del modelo se mantienen cercanas a la realidad observada.

- **Raíz del Error Cuadrático Medio (RMSE - Root Mean Squared Error):** El RMSE corresponde a la raíz cuadrada del promedio de los errores al cuadrado. Esta métrica mide la desviación estándar de los residuos de predicción. A diferencia del MAE, el RMSE penaliza de manera desproporcionada los errores de gran magnitud debido a la función cuadrática. Esto la convierte en una métrica de alta sensibilidad ante valores atípicos o fallos catastróficos en la predicción. En el contexto de este estudio, el RMSE es fundamental para evaluar la estabilidad del modelo un valor de RMSE significativamente superior al MAE indicaría una alta varianza en los residuos, sugiriendo que el modelo comete errores puntuales severos.

18 Diagnóstico de Problemas de Aprendizaje mediante Métricas

El análisis conjunto de estas métricas sobre el conjunto de entrenamiento permite evidenciar fenómenos críticos en el comportamiento de las redes neuronales:

- **Detección de Sobreajuste:** La obtención de métricas de error cercanas a cero absolutos durante la fase de entrenamiento demuestra una anomalía de aprendizaje conocida como memorización. Esto indica que el modelo ha ajustado sus parámetros para replicar el ruido estadístico de la muestra limitada en lugar de aprender la función subyacente, perdiendo su capacidad de generalización ante datos futuros.
- **Validación de Generalización:** La presencia de un error residual bajo pero existente ($MAE > 0$) en modelos recurrentes es un indicador positivo de generalización. Demuestra que los mecanismos de regularización han impedido la memorización exacta, forzando al modelo a aproximar la dinámica general del sistema tolerando un margen de error intrínseco.

19 Análisis de Robustez y Estabilidad a Largo Plazo

Para verificar si el modelo aprendió la dinámica estructural del cultivo o si simplemente “memorizo” datos recientes, se realizó una prueba de estrés a largo plazo. Se proyectó el rendimiento para la próxima década (2024-2033) bajo un escenario de cambio climático simulado, introduciendo ruido estocástico anual en las variables ambientales.

19.1 Protocolo de Inyección de Ruido y Análisis de Sensibilidad

Para la generación de las proyecciones a largo plazo, se diseñó un mecanismo de aleatoriedad controlada que introduce un factor de ruido aleatorio, consistente en una variación porcentual aleatoria entre 2% y 5% aplicada independientemente a cada variable en cada año simulado; esto permite generar escenarios sintéticos ante la

imposibilidad de conocer con certeza los valores futuros de las 135 variables climáticas, facilitando así la evaluación de la sensibilidad del modelo.

19.2 Justificación de la Prueba de Sensibilidad

Es fundamental aclarar que este ruido matemático no pretende modelar la complejidad real de fenómenos meteorológicos caóticos. Sin embargo, su implementación cumple un rol crítico como Prueba de Sensibilidad. El objetivo de inyectar este ruido no es acertar al clima del futuro, sino evaluar la estabilidad interna del modelo. Si la red neuronal (como el Perceptrón Multicapa) cambia drásticamente su predicción de rendimiento ante una variación mínima del 2% en el clima (ruido), se considera hipersensible y poco confiable. Por el contrario, si el modelo (como la arquitectura de Memoria Secuencial propuesta) es capaz de absorber estas perturbaciones y mantener una trayectoria de predicción coherente, constituye un fuerte indicio de que la arquitectura asimiló las dinámicas estables del cultivo, sustentando su capacidad de generalización ante condiciones de alta variabilidad.

TABLA DE PROYECCIONES FUTURAS			
Año	LR	MLP	LSTM
2024	6.05	11.40	9.41
2025	5.97	5.91	9.64
2026	5.08	10.59	8.91
2027	4.66	7.72	9.12
2028	4.94	8.26	9.45
2029	3.35	9.10	8.89
2030	3.29	8.92	9.08
2031	3.27	9.02	9.15
2032	3.31	8.19	9.10
2033	1.72	11.28	8.80

Figura 10 tabla de resultados de proyección a 10 años

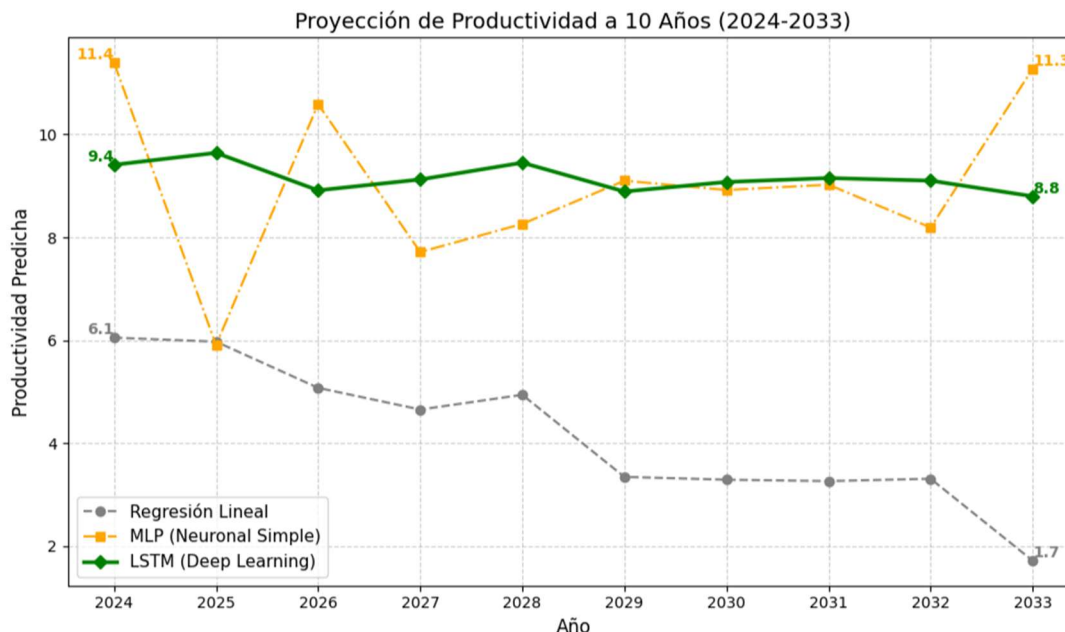


Figura 11 Gráfico de proyección a 10 años con los 3 modelos

Volatilidad del Modelo Simple (Línea Naranja - Perceptrón): El modelo estático exhibe un comportamiento de "dientes de sierra", con fluctuaciones violentas e injustificadas (saltando de 11.4 a 5.9 ton/ha en un solo ciclo). Esta alta volatilidad confirma que el modelo carece de inercia; al no tener memoria, reacciona de forma hipersensible al ruido climático anual, lo que lo vuelve inutilizable para la planificación estratégica.

Colapso de la Regresión (Línea Gris): La proyección lineal muestra una degradación monótona y constante hacia la improductividad (llegando a 1.7 ton/ha en 2033). Esto demuestra que el método estadístico clásico extrapola ciegamente la tendencia negativa sin considerar los mecanismos de adaptación biológica de la especie.

Robustez de la Arquitectura Propuesta (Línea Verde): El modelo seleccionado demuestra una estabilidad superior. Mantiene las proyecciones dentro de un rango acotado y realista (8.8 - 9.6 ton/ha), exhibiendo una suave tendencia de ajuste a la baja acorde al estrés climático, pero sin caer en el colapso ni en la volatilidad. Esta capacidad para filtrar el ruido y mantener una trayectoria coherente confirma que el modelo ha capturado la "inercia" del sistema productivo.

19.3 Valor de la Validación Agronómica Basada en Teoría

Para trascender la mera evaluación de estabilidad estadística y asegurar la coherencia agronómica sugerida por la revisión, las trayectorias resultantes de la prueba de sensibilidad se contrastarán directamente con estimaciones basadas en la teoría fisiológica del cultivo. específicamente, se utilizarán modelos de umbrales críticos bioclimáticos establecidos en la literatura agrícola (tales como los límites de tolerancia térmica a heladas y los requerimientos de acumulación de días grado). Esta validación cruzada es fundamental, ya que permite verificar si las proyecciones estables del modelo ante el ruido meteorológico se mantienen dentro de los rangos de viabilidad biológica real. De este modo, la prueba no solo certifica la robustez algorítmica del sistema, sino

que aporta un indicador objetivo de su utilidad y fidelidad conceptual para modelar las respuestas del ecosistema agrícola en escenarios de alta incertidumbre.

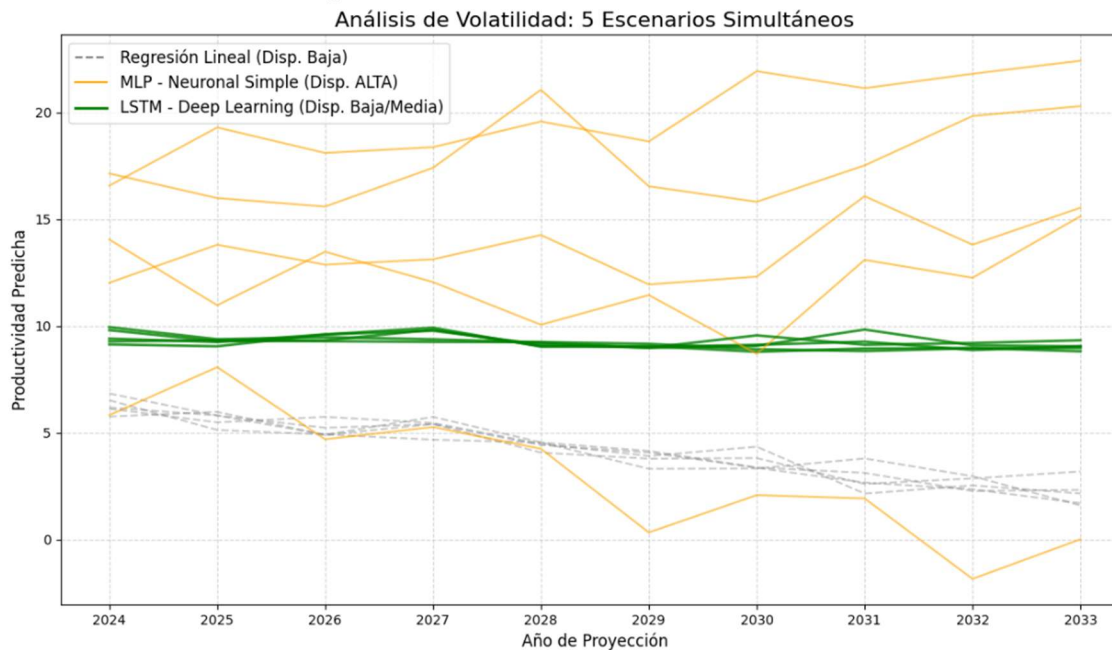


Figura 12 Gráfico del experimento repetido 10 veces

19.4 Predicción Operativa para el Ciclo 2024

Basado en la robustez demostrada anteriormente, el sistema entrega como resultado oficial para la toma de decisiones la proyección del modelo de memoria secuencial. Valor Proyectado: 9.40 toneladas por hectárea.

1. Regresión Lineal: 6.122
2. MLP (Neuronal Simple): 10.612
3. LSTM (Deep Learning): 9.399

Figura 13 Resultados de la comparación de la predicción para 2024

20 Conclusiones

El desarrollo de esta investigación permitió validar la viabilidad técnica de aplicar modelos de aprendizaje profundo en contextos de agricultura de precisión caracterizados por una severa limitación de registros históricos. Las principales conclusiones derivadas del estudio experimental son:

- **Superioridad de la Arquitectura Recurrente:** Se demostró empíricamente que el modelo seleccionado supera significativamente a los enfoques tradicionales (Regresión Lineal) y a las redes neuronales estáticas (Perceptrón Multicapa). La capacidad de la arquitectura propuesta para procesar la historia climática como una secuencia evolutiva, y no como un promedio estático.
- **Efectividad de la Estrategia para "Datos Escasos":** La implementación de la estrategia de ingeniería denominada "Maximización del Aprendizaje"

(entrenamiento con el 100% del conjunto de datos complementado con fuerte regularización estructural y aprendizaje estocástico) resultó efectiva. El modelo logró aprender la topología general del fenómeno sin caer en la memorización pura (sobreajuste), tal como evidencian las pruebas de estrés a 10 años donde la red no colapsó ante datos nuevos y ruidosos.

20.1 Consideraciones de Escalamiento y Trabajo Futuro

Si bien el prototipo actual está limitado a las restricciones de un conjunto de datos limitado (5 ventanas temporales), la arquitectura fue concebida para escalar hacia un entorno de datos ideal sin requerir modificaciones estructurales mayores. Se identifican las siguientes líneas de desarrollo para la evolución del sistema:

- **Transición del Protocolo de Validación:** Actualmente, el sistema utiliza una estrategia de maximización de datos debido a la escasez de muestras. En un escenario futuro, con la disponibilidad de un historial más extenso (idealmente), el protocolo de entrenamiento deberá transicionar hacia una validación de Ventana Deslizante. Este método estándar permitirá simular rigurosamente el despliegue productivo, reentrenando el modelo iterativamente a medida que nuevos años de cosecha se incorporen al registro histórico.
- **Incorporación del Factor Biológico (Alternancia Productiva):** El modelo actual se centra exclusivamente en el potencial climático-productivo. Para futuras versiones, se propone integrar el rendimiento del año anterior como una variable de entrada explícita. Esto permitirá modelar el fenómeno de "añerismo" o alternancia productiva propia de los frutales, donde una alta carga frutal en un año tiende a reducir fisiológicamente la producción del siguiente, independientemente de las condiciones climáticas favorables.
- **Ajuste de Hiperparámetros para Grandes Volúmenes:** Con el incremento futuro de la base de datos, la estrategia de optimización deberá evolucionar. Por ejemplo, aumentar el tamaño del grupo de muestras procesadas por iteración (actualmente unitario) a bloques más grandes (ej. 4 u 8). Esta modificación estabilizará el cálculo matemático del error, acelerando la velocidad de aprendizaje del modelo y aprovechando mejor la capacidad de procesamiento paralelo, lo que permitirá reducir el ruido aleatorio que hoy resulta indispensable debido a la escasez de registros.

Agradecimientos. Agradezco sinceramente a mi profesor guía, Gabriel Jara, por su invaluable orientación, su apoyo técnico y por responder cada duda en mi trayecto universitario desde el primer día en esta sede. Asimismo, extendo mi gratitud a los ingenieros y profesionales que me antecedieron, quienes me brindaron las herramientas indispensables que hicieron posible este trabajo. Agradezco también a mis compañeros por su compañerismo y labor compartida y su visión y fe en este proyecto. Finalmente, doy las gracias a mi familia por creer en mí, ser pacientes y aguantar en este camino, pero, sobre todo, por su respaldo incondicional, su preocupación y todo su amor.

21 Referencias

- [5] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780.
- [6] Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz-Sabater, J., ... & Thépaut, J. N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999-2049.
- [7] Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (pp. 29-39).
- [8] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagation errors. *Nature*, 323(6088), 533-536.
- [9] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1), 1929-1958.
- [10] Whiley, A. W., Schaffer, B., & Wolstenholme, B. N. (Eds.). (2013). *The avocado: botany, production and uses*.