Maestría en

CIENCIA DE DATOS Y MÁQUINAS DE APRENDIZAJE MENCIÓN EN INTELIGENCIA ARTIFICIAL

Trabajo previo a la obtención de título de Magister en Ciencia de Datos y Máquinas de Aprendizaje con mención en Inteligencia Artificial

AUTORES:

Santiago Gabriel Aguilera Sarmiento

Yanela Stefany Angamarca Castillo

Hugo Javier Erazo Granda

Bryan Andrés Quimbita Carrera

TUTORES:

Navas Castellanos Andrés Roberto

Avalos Silva Héctor Guillermo

Modelado predictivo de la producción agrícola de papa, maíz y arveja basado en variabilidad climática y factores productivos en la Sierra del Ecuador

Quito – Ecuador

junio - 2026

REINVENTEMOS EL FUTURO

CERTIFICACIÓN DE AUTORÍA

Nosotros, **Santiago Gabriel Aguilera Sarmiento, Yanela Stefany Angamarca Castillo, Hugo Javier Erazo Granda, Bryan Andrés Quimbita Carrera**, 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.

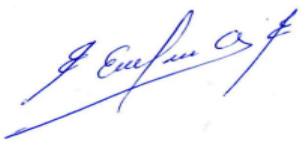
Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

Firma del graduado

Santiago Gabriel Aguilera Sarmiento

Firma del graduado

Yanela Stefany Angamarca Castillo



Firma del graduado

Hugo Javier Erazo Granda

Firma del graduado

Bryan Andrés Quimbita Carrera

AUTORIZACIÓN DE DERECHOS DE PROPIEDAD INTELECTUAL

Nosotros, **SANTIAGO GABRIEL AGUILERA SARMIENTO,YANELA STEFANY ANGAMARCA CASTILLO, HUGO JAVIER ERAZO GRANDA, BRYAN ANDRÉS QUIMBITA CARRERA**, en calidad de autores del trabajo de investigación titulado ***Modelado predictivo de la producción agrícola de papa, maíz y arveja basado en variabilidad climática y factores productivos en la Sierra del 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, junio 2026

Firma del graduado

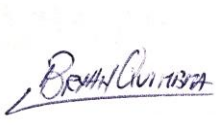
Santiago Gabriel Aguilera Sarmiento

Firma del graduado

Yanela Stefany Angamarca Castillo

Firma del graduado

Hugo Javier Erazo Granda

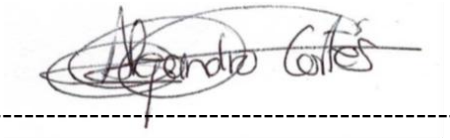


Firma del graduado

Bryan Andrés Quimbita Carrera

APROBACIÓN DE DIRECCIÓN Y COORDINACIÓN DEL PROGRAMA

Nosotros, **Alejandro Cortés López Director EIG y Karla Estefanía Mora Cajas Coordinadora UIDE,** declaramos que: **SANTIAGO GABRIEL AGUILERA SARMIENTO, YANELA STEFANY ANGAMARCA CASTILLO, HUGO JAVIER ERAZO GRANDA, BRYAN ANDRÉS QUIMBITA CARRERA**, 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 Ciencia de datos y máquinas de aprendizaje con mención en Inteligencia Artificial

Karla Estefanía Mora Cajas

Coordinadora de la Maestría en Ciencia de datos y máquinas de aprendizaje con mención en Inteligencia Artificial

DEDICATORIA

A nuestra familia y amigos...

AGRADECIMIENTOS

Expresamos nuestro más sincero agradecimiento a nuestros docentes, familiares y amigos, quienes con su apoyo, orientación, confianza y constante motivación nos acompañaron durante este proceso académico. Su respaldo fue fundamental para superar los desafíos presentados y hacer posible la culminación de este proyecto de titulación de maestría.

RESUMEN

La producción agrícola en la región Sierra del Ecuador está influenciada por la variabilidad climática, especialmente por factores como la temperatura y la precipitación, los cuales afectan el desarrollo de los cultivos y dificultan la planificación agrícola y la toma de decisiones. El objetivo de esta investigación fue desarrollar una herramienta predictiva basada en técnicas de aprendizaje automático para estimar la producción de papa, maíz suave seco y arveja tierna. Para ello, se utilizaron datos históricos de producción agrícola y variables climáticas correspondientes al período 2002 - 2023, obtenidos de fuentes oficiales. Los datos fueron sometidos a procesos de limpieza, tratamiento de valores faltantes, generación de variables derivadas y análisis exploratorio. Posteriormente, se implementaron y compararon modelos de Regresión Lineal, Random Forest, XGBoost y Redes Neuronales Artificiales, evaluados mediante las métricas RMSE, MAE y R^2. Los resultados evidenciaron que la Regresión Lineal obtuvo el mejor desempeño para la predicción de papa y arveja, mientras que Random Forest presentó la mayor capacidad predictiva para el maíz, demostrando que la selección del modelo depende de las características de cada cultivo. Finalmente, se desarrolló una interfaz interactiva que permite estimar la producción agrícola a partir de variables climáticas y productivas. Se concluye que la herramienta constituye un apoyo para la planificación agrícola y la toma de decisiones basadas en datos frente a la variabilidad climática.

Palabras clave

Aprendizaje automático, producción agrícola, variabilidad climática, predicción agrícola.

ABSTRACT

Agricultural production in the Ecuadorian Highlands (Sierra region) is strongly influenced by climate variability, particularly by factors such as temperature and precipitation, which affect crop development and make agricultural planning and decision-making more challenging. The objective of this study was to develop a predictive tool based on machine learning techniques to estimate the production of potatoes, dry soft corn, and green peas. Historical agricultural production data and climatic variables for the period 2002–2023 were obtained from official sources. The dataset was processed through data cleaning, missing value treatment, feature engineering, and exploratory data analysis. Subsequently, Linear Regression, Random Forest, XGBoost, and Artificial Neural Network models were implemented and compared using RMSE, MAE, and R^2 as evaluation metrics. The results showed that Linear Regression achieved the best predictive performance for potatoes and green peas, whereas Random Forest provided the highest predictive accuracy for dry soft corn, demonstrating that model selection depends on the characteristics of each crop. Finally, an interactive interface was developed to estimate agricultural production based on climatic and production-related variables. It is concluded that the proposed tool supports agricultural planning and data-driven decision-making under conditions of climate variability.

Keywords

Machine learning, agricultural production, climate variability, agricultural prediction.

TABLA DE CONTENIDOS

CERTIFICACIÓN DE AUTORÍA i

AUTORIZACIÓN DE DERECHOS DE PROPIEDAD INTELECTUAL ii

ACUERDO DE CONFIDENCIALIDAD ..... **¡Error! Marcador no definido.**

APROBACIÓN DE DIRECCIÓN Y COORDINACIÓN DEL PROGRAMA iii

DEDICATORIA iv

AGRADECIMIENTOS v

RESUMEN vi

ABSTRACT vii

TABLA DE CONTENIDOS viii

LISTA DE TABLAS xii

LISTA DE FIGURAS xiv

CAPÍTULO 1 1

1. INTRODUCCIÓN 1

1.1. Introducción 1

1.2. Definición del proyecto 3

1.3. Justificación e importancia del trabajo de investigación 4

1.4. Alcance 5

1.5. Objetivos 6

1.5.1. Objetivo general 6

1.5.2. Objetivo específico 6

CAPÍTULO 2 8

2. REVISIÓN DE LITERATURA 8

2.1. Estado del Arte 8

2.1.1. Modelado Predictivo en Agricultura 8

2.1.2. Estudios Internacionales 9
2.1.3. Estudios Relacionados en Ecuador 10
2.1.4. Síntesis crítica del estado del arte 11
2.1.5. Vacío de investigación 11
2.2. Marco Teórico 12
2.2.1. Modelos estadísticos tradicionales 13
2.2.2. Fundamentos del aprendizaje automático 14
2.2.3. Aplicaciones del aprendizaje automático en el sector agrícola 15
2.2.4. Algoritmos utilizados para la predicción 16
2.2.5. Evaluación de modelos predictivos 20
2.2.6. Variabilidad Climática 22
2.2.7. Producción Agrícola y Factores Determinantes 31
2.2.8. Factores Climáticos y Productivos 33
2.2.9. Influencia de variables temporales en la agricultura 35
2.2.10. Vulnerabilidad agrícola frente a variabilidad climática 38
CAPÍTULO 3 43
3. DESARROLLO 43
3.1. Metodología 43
3.2. Desarrollo del Trabajo 46
3.2.1. Obtención y Selección de las Fuentes de Información 46
3.2.2. Análisis Exploratorio de Datos Productivos 49
3.2.3. Análisis Exploratorio de Datos Climáticos 51
3.2.4. Tratamiento de Valores Faltantes 53
3.2.5. Integración de Información Agroclimática 56
3.2.6. Construcción del dataset final 57

3.2.6.1. Incorporación del índice ENSO 57
3.2.6.1.1. Variables de interacción 57
3.2.6.2. Ingeniería de variables 58
3.2.6.2.1. Indicadores de variabilidad climática 58
3.2.6.2.2. Variables temporales.....59
3.2.7. Dataset final 61
3.2.8. Análisis Exploratorio de Datos (EDA) del dataset integrado.....63
3.2.9. Tratamiento de Datos y Calidad de la Información.....63
3.2.10. Caracterización Estadística Segmentada.....64
3.2.11. Análisis de Correlación e Interacciones No Lineales.....66
3.2.12. Concentración Geográfica y Vulnerabilidad 66
CAPÍTULO 4.....68
4. ANÁLISIS DE RESULTADOS.....68
4.1. EVALUACIÓN DE MODELOS PREDICTIVOS 68
4.1.1. Validación de Modelos y Estrategia Experimental.....68
4.1.2. Métricas de Desempeño 68
4.1.3. Desempeño del Modelado Predictivo en Papa (Tubérculo Fresco): 68
4.1.4. Desempeño del Modelado Predictivo en Maíz Suave Seco 71
4.1.5. Desempeño del Modelado Predictivo en Arveja.....73
4.2. COMPARACIÓN DE MODELOS 75
4.3. INTERPRETABILIDAD Y ANÁLISIS DE IMPORTANCIA DE VARIABLES 78
4.4. APLICACIÓN (SISTEMA INTELIGENTE DE PREDICCIÓN AGRÍCOLA) 79

4.4.1. Inicialización y Gestión de Modelos (Capa de Entrenamiento)...80
4.4.2. Interfaz y Lógica de Inferencia 80
4.4.3. Interpretabilidad y Validación de Resultados 83
4.4.4. Análisis de Resultados de Inferencia: Caso de Estudio (Cultivo: Papa, Chimborazo, 2019)....83
4.4.5. Análisis de Resultados de Inferencia: Caso de Estudio (Cultivo: Maiz, Cotopaxi, 2011)....88
4.4.6. Análisis de Resultados de Inferencia: Caso de Estudio (Cultivo: Arveja, Carchi, 2009)....92
CAPÍTULO 5....98
5. CONCLUSIONES Y RECOMENDACIONES 98
5.1. Conclusiones 98
5.2. Recomendaciones 99
6. REFERENCIAS BIBLIOGRÁFICAS....100
7. APÉNDICE 104

LISTA DE TABLAS

Tabla 1 Características de la variabilidad climática y el cambio climático.....23

Tabla 2 Categorías del índice ENSO y su efecto típico en la Sierra del Ecuador 27

Tabla 3 Impacto del fenómeno ENSO en la producción promedio de los cultivos analizados ...28

Tabla 4 Rangos óptimos de temperatura e impactos potenciales por temperaturas extremas en los cultivos analizados 29

Tabla 5 Variables de interacción entre el índice ENSO y las condiciones climáticas locales.....30

Tabla 6 Comparación entre el rendimiento agrícola y la producción total como indicadores de evaluación.....32

Tabla 7 Correlación entre las variables predictoras y la producción agrícola 34

Tabla 8 Cálculo de variables lag. 36

Tabla 9 Promedios móviles de tres años (MA3) de las variables analizadas.....38

Tabla 10 Análisis de vulnerabilidad de las provincias de la región Sierra del Ecuador a partir de variables productivas y climáticas 40

Tabla 11 Variables utilizadas para representar la vulnerabilidad en el modelo de aprendizaje automático 42

Tabla 12 Resumen metodológico de la investigación.....44

Tabla 13 Fuentes de información y variables utilizadas en el estudio 48

Tabla 14 Características del conjunto de datos productivos original 51

Tabla 15 Características del conjunto de datos climáticos original.....52

Tabla 16 Métodos de imputación de datos evaluados para el tratamiento de valores faltantes.54

Tabla 17 Resultados de la evaluación de los métodos de imputación de datos 55

Tabla 18 Variables de interacción entre el índice ENSO y las condiciones climáticas locales...58

Tabla 19 Variables derivadas para representar la variabilidad climática 59

Tabla 20 Cálculo de variables temporales lag. 60

Tabla 21 Promedios móviles de tres años (MA3) de las variables analizadas.....61

Tabla 22 Clasificación de las variables del dataset final utilizadas para el entrenamiento de los modelos predictivos 62

Tabla 23 Métricas de evaluación de los modelos de aprendizaje automático para la predicción de la producción de papa (2004 - 2023).....69

Tabla 24 Métricas de evaluación de los modelos de aprendizaje automático para la predicción de la producción de maíz suave seco (2004–2023) 71

Tabla 25 Métricas de evaluación para el cultivo de Arveja (2004-2023). 73

Tabla 26 Métricas comparativas por cultivos (mejores modelos).....75

Tabla 27 Interpretación de los Pesos del Modelo (Coeficientes β). 87

Tabla 28 Jerarquía de importancia de variables en el modelo Random Forest. 91

Tabla 29 Análisis de pesos del modelo de regresión lineal (Coeficientes β).....95

LISTA DE FIGURAS

Figura 1 Temperatura promedio y precipitación acumulada por provincia (2002–2023).....24

Figura 2 Variabilidad climática interanual por provincia (coeficiente de variación, CV).....26

Figura 3 Pipeline metodológico del sistema de predicción 45

Figura 4 Flujo general de construcción del conjunto de datos.....49

Figura 5 Evolución temporal de la producción agrícola por cultivo (2002–2023) 64

Figura 6 Distribución de la producción agrícola por cultivo en la Sierra ecuatoriana (2002-2023) 65

Figura 7 Matriz de correlación entre las variables del conjunto de datos.....67

Figura 8 Contraste entre la producción observada y la producción predicha mediante Regresión Lineal 70

Figura 9 Contraste entre la producción observada y la producción predicha mediante Random Forest (Maíz) 72

Figura 10 Contraste entre la producción observada y la producción predicha mediante Regresión Lineal (Arveja).....74

Figura 11 Desempeño del coeficiente de determinación (R^2) por modelo y cultivo.....76

Figura 12 Ajuste del modelo: valores observados frente a valores predichos (Maíz) 77

Figura 13 Importancia de las variables en el modelo Random Forest para el cultivo de maíz...78

Figura 14 Arquitectura de la aplicación y flujo de inferencia. 81

Figura 15 Interfaz de usuario del sistema de predicción agrícola 82

Figura 16 Resultado de la inferencia para el cultivo de papa 83

Figura 17 Influencia de las variables en la predicción del modelo para el cultivo de papa.....85

Figura 18 Resultado de inferencia para el cultivo de Maiz. 89

Figura 19 Influencia de las variables en la predicción del modelo para el cultivo de maíz.....90

Figura 20 Resultado de la inferencia para el cultivo de arveja.....93

Figura 21 Influencia de las variables en la predicción del modelo para el cultivo de arveja.....94

CAPÍTULO 1

1. INTRODUCCIÓN

1.1. Introducción

La agricultura constituye una de las actividades económicas y sociales más importantes para el desarrollo de los países, debido a su contribución a la seguridad alimentaria, la generación de empleo y el crecimiento económico. En Ecuador, la región Sierra concentra una parte significativa de la producción agrícola nacional, destacándose cultivos como la papa, el maíz suave seco y la arveja tierna, los cuales representan una fuente de ingresos para miles de productores y un componente esencial de la cadena alimentaria del país. La Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO, 2022).

Sin embargo, la producción agrícola se encuentra expuesta a diversos factores que influyen directamente en su comportamiento y rendimiento. Entre ellos, la variabilidad climática ocupa un papel fundamental, ya que cambios en la temperatura y la precipitación pueden afectar los ciclos de crecimiento de los cultivos, la disponibilidad de recursos hídricos y la productividad agrícola en general (IPCC, 2023). A estos factores se suman variables productivas, como la superficie cultivada, así como dinámicas temporales que reflejan el comportamiento histórico de la producción.

En los últimos años, el avance de la ciencia de datos y las técnicas de aprendizaje automático ha permitido desarrollar herramientas capaces de analizar grandes volúmenes de información y generar modelos predictivos con altos niveles de precisión. Estos enfoques facilitan la identificación de patrones complejos y relaciones no lineales entre variables climáticas y productivas, contribuyendo a mejorar los procesos de planificación y toma de decisiones en el sector agrícola (James et al, 2021).

En este contexto, la presente investigación propone el desarrollo de un modelo predictivo para estimar la producción agrícola de los cultivos de papa, maíz suave seco y arveja tierna en provincias de la región Sierra del Ecuador. Para ello, se integran variables climáticas, productivas y temporales obtenidas de fuentes oficiales, aplicando técnicas de aprendizaje automático que permitan analizar su influencia sobre la producción agrícola y generar estimaciones confiables. Como resultado, se busca aportar una herramienta basada en datos que contribuya a fortalecer la planificación agrícola y la gestión de los riesgos asociados a la variabilidad climática.

A pesar de la importancia estratégica de la producción agrícola para la región Sierra del Ecuador, aún persisten limitaciones para estimar de manera precisa el comportamiento futuro de cultivos como la papa, el maíz suave seco y la arveja tierna frente a la variabilidad climática. Si bien existen investigaciones que analizan la influencia de variables climáticas sobre la producción agrícola o que aplican técnicas de aprendizaje automático en contextos específicos, la evidencia muestra que son escasos los estudios desarrollados para el contexto ecuatoriano que integren simultáneamente variables climáticas históricas, factores productivos y componentes temporales en un mismo modelo predictivo. Esta limitación dificulta la generación de herramientas confiables que apoyen la planificación agrícola y la toma de decisiones basada en evidencia frente a escenarios de incertidumbre climática. En este contexto, la presente investigación aborda esta problemática mediante el desarrollo y evaluación de modelos de aprendizaje automático orientados a mejorar la predicción de la producción agrícola de cultivos estratégicos de la región Sierra del Ecuador.

1.2. Definición del proyecto

El presente proyecto tiene como finalidad desarrollar un modelo predictivo de la producción agrícola de tres cultivos estratégicos: arveja tierna, maíz suave seco y papa. En provincias seleccionadas de la región Sierra del Ecuador, considerando la influencia de la variabilidad climática, factores productivos y dinámicas temporales, mediante técnicas de aprendizaje automático.

La producción agrícola en la región Sierra está condicionada por múltiples factores, entre los que destacan las variables climáticas, tales como la temperatura y la precipitación, así como factores productivos como la superficie cultivada (FAO, 2022; IPCC, 2023). Adicionalmente, el comportamiento de los cultivos presenta dinámicas temporales, donde los efectos del clima pueden manifestarse de forma retardada en el tiempo.

En este contexto, el proyecto propone un enfoque multivariable que integra datos históricos de la producción agrícola y variables climáticas, permitiendo construir un dataset consolidado a nivel provincia, año y cultivo. Sobre esta base, se incorporan variables derivadas que capturan la dinámica temporal, como rezagos climáticos y tasas de crecimiento interanual de la producción.

A partir de este conjunto de datos, se implementarán modelos de aprendizaje automático orientados a predecir la producción agrícola y analizar la influencia relativa de las variables consideradas, con el fin de generar conocimiento útil para la planificación agrícola y la toma de decisiones basadas en datos (James et al, 2021; Machuca et al., 2025).

1.3. Justificación e importancia del trabajo de investigación

La producción agrícola es un componente fundamental para la seguridad alimentaria y el desarrollo económico de la región Sierra del Ecuador (FAO, 2022; MAG, 2025). Cultivos como la arveja tierna, el maíz suave seco y la papa representan una parte significativa de la actividad agrícola, siendo esenciales tanto para el consumo interno como para la economía de pequeños y medianos productores.

La variabilidad climática, caracterizada por cambios en los patrones de precipitación y temperatura, constituye uno de los principales factores de incertidumbre en la producción agrícola (IPCC, 2023). Sin embargo, la producción no depende exclusivamente de las condiciones climáticas, sino también de factores productivos como la superficie cultivada y de dinámicas temporales asociadas al comportamiento de los cultivos a lo largo del tiempo.

En este contexto, resulta necesario adoptar enfoques integradores que permitan analizar de manera conjunta estos factores. El presente estudio propone un modelo multivariable que combina variables climáticas, productivas y temporales, con el fin de capturar de forma más realista la complejidad del sistema agrícola.

Desde una perspectiva metodológica, el uso de técnicas de aprendizaje automático permite modelar relaciones no lineales y explorar patrones complejos en los datos, superando las limitaciones de enfoques tradicionales (James et al, 2021). Asimismo, la incorporación de técnicas de interpretabilidad facilita la comprensión del impacto de cada variable en la producción agrícola. Los resultados de este estudio pueden contribuir a la planificación agrícola, la gestión del riesgo climático y la toma de decisiones basada en evidencia, tanto a nivel técnico como institucional.

Desde el ámbito académico, el proyecto se alinea con los objetivos de formación en ciencia de datos y aprendizaje automático, integrando procesos de análisis exploratorio, ingeniería de variables, modelado predictivo y evaluación de modelos.

1.4. Alcance

- *Ámbito geográfico:* Se considerarán las siguientes provincias de la región Sierra del Ecuador: Carchi, Imbabura, Pichincha, Cotopaxi, Tungurahua, Chimborazo, Cañar, Azuay y Loja, seleccionadas en función de la disponibilidad y consistencia de datos climáticos y productivos.
- *Delimitación por disponibilidad de datos:* Se excluyen provincias con ausencia significativa de datos en el periodo de estudio, con el fin de garantizar la calidad y coherencia del análisis.
- *Unidad de análisis:* Provincia, Año, Cultivo.
- *Periodo de estudio:* 2002–2023.
- *Cultivos analizados:* Arveja tierna, Maíz suave seco y Papa, seleccionados por su relevancia productiva y disponibilidad de información.
- *Variable dependiente:* Producción agrícola (tonelada).
- *Variables independientes:* Temperatura promedio, precipitación acumulada, superficie plantada, variables temporales (lags, promedios móviles, crecimiento interanual) y tipo de cultivo.
- *Enfoque metodológico:* Modelado predictivo supervisado mediante técnicas de aprendizaje automático, con análisis comparativo de modelos e interpretación de variables.

1.5. Objetivos

1.5.1. Objetivo general

Evaluar modelos de aprendizaje automático para determinar el modelo con mejor desempeño predictivo de la producción agrícola de los cultivos de arveja tierna, maíz suave seco y papa en las provincias de la región Sierra del Ecuador, integrando variables climáticas, factores productivos y dinámicas temporales.

1.5.2. Objetivo específico

1. Integrar las bases de datos climáticas y productivas mediante un proceso de depuración para construir un conjunto de datos consolidado a nivel de provincia, año y cultivo.
2. Evaluar la calidad del conjunto de datos y aplicar técnicas de tratamiento de valores faltantes para garantizar su consistencia.
3. Construir variables representativas de la variabilidad climática y variables derivadas que capturen dinámicas temporales, tales como rezagos (lags), promedios móviles y crecimiento interanual de la producción.
4. Realizar un análisis exploratorio de datos para identificar patrones, tendencias y relaciones entre variables climáticas, productivas y la producción agrícola.
5. Implementar modelos de aprendizaje automático supervisado para la predicción de la producción agrícola.
6. Comparar el desempeño de los modelos mediante RMSE, MAE y R^2 para seleccionar el modelo con mejor capacidad predictiva.
7. Analizar la importancia y la contribución de las variables climáticas y productivas mediante técnicas de interpretabilidad.

8. Desarrollar una interfaz interactiva para estimar la producción agrícola utilizando el modelo predictivo seleccionado.

CAPÍTULO 2

2. REVISIÓN DE LITERATURA

2.1. Estado del Arte

2.1.1. Modelado Predictivo en Agricultura

La transformación digital del sector agrícola ha impulsado el desarrollo de metodologías orientadas a mejorar la predicción de variables productivas mediante el uso de información climática, ambiental y agrícola. El incremento en la disponibilidad de datos provenientes de estaciones meteorológicas, sensores remotos, imágenes satelitales y registros históricos de producción ha favorecido la incorporación de técnicas de inteligencia artificial y aprendizaje automático para fortalecer la planificación agrícola, optimizar el uso de recursos y apoyar la toma de decisiones (Food and Agriculture Organization [FAO], 2022; Liakos et al., 2018; Kamilaris & Prenafeta-Boldú, 2018).

Tradicionalmente, la predicción de la producción y del rendimiento agrícola se ha realizado mediante modelos estadísticos basados en regresión y series temporales. Aunque estos enfoques permiten identificar tendencias y relaciones entre variables, presentan limitaciones para representar la naturaleza compleja de los sistemas agrícolas, donde interactúan simultáneamente factores climáticos, ambientales y productivos. Como consecuencia, el aprendizaje automático ha adquirido mayor relevancia debido a su capacidad para modelar relaciones no lineales y mejorar la precisión de las predicciones.

2.1.2. Estudios Internacionales

La literatura internacional evidencia un crecimiento significativo en la aplicación de técnicas de aprendizaje automático para la predicción agrícola.

Van Klompenburg et al. (2020), mediante una revisión sistemática de la literatura, identificaron que los algoritmos Random Forest, Support Vector Machine, Artificial Neural Networks y Gradient Boosting son los métodos más utilizados para la predicción del rendimiento agrícola debido a su capacidad para representar relaciones complejas entre variables climáticas, edáficas y productivas.

Por su parte, Crane-Droesch (2018) analizó la aplicación de modelos de aprendizaje automático para evaluar el impacto de la variabilidad climática sobre la producción agrícola. Sus resultados evidenciaron que estos modelos superan en diversos escenarios a los métodos estadísticos tradicionales al capturar relaciones no lineales entre las condiciones climáticas y la respuesta de los cultivos.

Asimismo, Shahhosseini et al. (2021) desarrollaron modelos híbridos que integran simulación agrícola y aprendizaje automático para estimar el rendimiento del maíz. Los autores concluyeron que la incorporación conjunta de variables meteorológicas, información histórica y variables de manejo agrícola incrementa significativamente la precisión de las predicciones.

La FAO (2022) destaca que la inteligencia artificial y las tecnologías digitales constituyen herramientas estratégicas para fortalecer la resiliencia de los sistemas agroalimentarios frente a la variabilidad climática, facilitando la planificación agrícola y la toma de decisiones basada en evidencia.

2.1.3. Estudios Relacionados en Ecuador

En Ecuador, la aplicación del aprendizaje automático al sector agropecuario ha mostrado un crecimiento durante los últimos años; sin embargo, la mayoría de las investigaciones se orientan al análisis de datos meteorológicos o a aplicaciones específicas.

Machuca Catagua et al. (2025) evaluaron la eficacia de diferentes algoritmos de aprendizaje automático, entre ellos Random Forest, Gradient Boosting, Support Vector Machine y XGBoost, para apoyar la toma de decisiones mediante el análisis de datos meteorológicos relacionados con la producción agrícola. Los resultados mostraron que XGBoost presentó el mejor desempeño predictivo. No obstante, el estudio se centra principalmente en la comparación del rendimiento de los algoritmos y no desarrolla un modelo orientado específicamente a la predicción de la producción agrícola de cultivos estratégicos de la Sierra ecuatoriana, ni incorpora variables temporales, indicadores ENSO e información histórica de producción.

Mendoza Cuzme et al. (2025) desarrollaron modelos predictivos basados en inteligencia artificial para estimar la producción agrícola utilizando variables climáticas. Los autores concluyen que estas herramientas fortalecen la planificación agrícola y reducen la incertidumbre asociada a la variabilidad climática. Sin embargo, el estudio no integra variables productivas históricas ni analiza específicamente los cultivos predominantes de la región Sierra.

De forma complementaria, Bravo Guzmán et al. (2024) aplicaron modelos de aprendizaje automático para la predicción de eventos climáticos en Ecuador. Aunque su investigación se orienta al pronóstico de variables meteorológicas y no directamente a la producción agrícola, demuestra la viabilidad del uso de técnicas de inteligencia artificial para el análisis de información climática en el contexto nacional.

2.1.4. Síntesis crítica del estado del arte

La revisión de la literatura evidencia que existe consenso respecto al potencial del aprendizaje automático para mejorar la precisión de las predicciones agrícolas frente a los métodos estadísticos tradicionales. Sin embargo, también se identifican diferencias metodológicas importantes entre los estudios analizados.

Los trabajos internacionales integran múltiples fuentes de información, incluyendo variables climáticas, propiedades del suelo, registros históricos de producción, prácticas de manejo agrícola y modelos de simulación, lo que permite construir modelos predictivos con mayor capacidad de generalización (Crane-Droesch, 2018; Shahhosseini et al., 2021; Van Klompenburg et al., 2020). En contraste, las investigaciones desarrolladas en Ecuador se concentran principalmente en la evaluación del desempeño de algoritmos de aprendizaje automático o en el análisis de variables meteorológicas específicas (Bravo Guzmán et al., 2024; Machuca Catagua et al., 2025; Mendoza Cuzme et al., 2025).

Asimismo, los estudios internacionales fueron desarrollados para cultivos y condiciones agroclimáticas diferentes a las de la región Sierra del Ecuador, lo que limita la transferencia directa de sus resultados al contexto nacional. Esta diferencia evidencia la necesidad de desarrollar modelos adaptados a las características climáticas y productivas propias del país.

2.1.5. Vacío de investigación

A partir de la revisión de la literatura se identifica una limitada aplicación de modelos de aprendizaje automático que integren simultáneamente variables climáticas históricas, indicadores asociados al fenómeno El Niño–Oscilación del Sur (ENSO), variables productivas y componentes temporales para la predicción de la producción agrícola de cultivos estratégicos de la región Sierra del Ecuador.

Aunque los estudios internacionales demuestran la eficacia de integrar múltiples fuentes de información para mejorar la precisión de los modelos predictivos, estos fueron desarrollados en contextos agroclimáticos distintos al ecuatoriano. Por otra parte, las investigaciones nacionales evidencian avances en el uso de inteligencia artificial aplicada al sector agrícola; sin embargo, se enfocan principalmente en la comparación de algoritmos o en el análisis de variables meteorológicas, sin desarrollar modelos que integren información agroclimática, variables históricas de producción, indicadores ENSO y componentes temporales.

En este contexto, la presente investigación propone desarrollar y comparar modelos de aprendizaje automático para estimar la producción agrícola de papa, maíz suave seco y arveja tierna mediante la integración de variables climáticas, productivas y temporales correspondientes a las provincias de la región Sierra del Ecuador. De esta manera, se busca contribuir al fortalecimiento de la planificación agrícola y de la toma de decisiones basada en datos, proporcionando una herramienta predictiva adaptada a las condiciones agroclimáticas del país.

2.2. Marco Teórico

El marco teórico presenta los fundamentos conceptuales que sustentan la presente investigación, abordando los principales enfoques estadísticos y de aprendizaje automático utilizados para la predicción de variables agrícolas. Asimismo, se describen los algoritmos de regresión considerados en este estudio y las métricas empleadas para evaluar su desempeño, proporcionando la base teórica para el desarrollo del modelo predictivo propuesto.

2.2.1. Modelos estadísticos tradicionales

Los modelos estadísticos tradicionales constituyen una de las principales herramientas utilizadas para describir relaciones entre variables y generar predicciones a partir de información histórica. Estas metodologías permiten analizar patrones presentes en los datos y establecer relaciones matemáticas entre variables explicativas y variables de respuesta (Montgomery et al., 2021).

Uno de los métodos más utilizados es la regresión lineal, la cual busca modelar la relación entre una variable dependiente y una o varias variables independientes mediante una función matemática lineal. Este enfoque ha sido ampliamente aplicado en estudios agrícolas para analizar la influencia de factores climáticos y productivos sobre el rendimiento y la producción de los cultivos (James et al, 2021).

Otra técnica ampliamente empleada corresponde a los modelos de series temporales, los cuales permiten identificar tendencias, ciclos y patrones estacionales presentes en los datos históricos para realizar proyecciones futuras (Hastie et al., 2009). Estos modelos resultan particularmente útiles cuando las observaciones presentan dependencia temporal.

No obstante, los sistemas agrícolas suelen estar influenciados por múltiples factores que interactúan de manera compleja y no lineal. Debido a ello, los modelos estadísticos tradicionales pueden presentar limitaciones para representar adecuadamente dichas relaciones, afectando la precisión de las predicciones en escenarios de alta variabilidad climática y productiva.

2.2.2. Fundamentos del aprendizaje automático

El aprendizaje automático o Machine Learning es una rama de la inteligencia artificial orientada al desarrollo de algoritmos capaces de aprender patrones a partir de datos históricos para realizar predicciones o clasificaciones sin necesidad de reglas explícitamente programadas (Russell & Norvig, 2021; Goodfellow et al., 2016).

Dentro del aprendizaje supervisado, el algoritmo recibe ejemplos en los cuales se conocen tanto las variables predictoras como la variable objetivo. A partir de esta información, el modelo aprende relaciones existentes entre las variables y posteriormente utiliza dicho conocimiento para generar estimaciones sobre nuevas observaciones (Bishop, 2006).

Cuando la variable objetivo corresponde a un valor numérico continuo, el problema se clasifica como una tarea de regresión. Entre los algoritmos más utilizados para este tipo de problemas se encuentran la Regresión Lineal, los árboles de decisión, Random Forest y XGBoost.

Random Forest es un método de ensamble basado en múltiples árboles de decisión construidos mediante técnicas de muestreo aleatorio. Este algoritmo mejora la capacidad de generalización del modelo al combinar múltiples predicciones individuales, reduciendo la varianza y aumentando la estabilidad de los resultados (Breiman, 2001).

Por su parte, XGBoost (Extreme Gradient Boosting) es un algoritmo basado en Gradient Boosting que construye secuencialmente árboles de decisión optimizando una función de pérdida mediante métodos de descenso por gradiente y técnicas de regularización. Estas características permiten obtener modelos altamente precisos y robustos frente al sobreajuste (Chen & Guestrin, 2016).

La evaluación de los modelos predictivos suele realizarse mediante métricas estadísticas como el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE) y el coeficiente de determinación (R^2), las cuales permiten cuantificar la precisión de las predicciones y comparar objetivamente diferentes algoritmos (James et al, 2021).

2.2.3. Aplicaciones del aprendizaje automático en el sector agrícola

El aprendizaje automático ha encontrado múltiples aplicaciones dentro del sector agrícola debido a su capacidad para procesar grandes volúmenes de información y detectar patrones complejos en los datos. Entre las principales aplicaciones destacan la predicción de rendimientos agrícolas, la estimación de producción, la detección de enfermedades y plagas, la optimización del riego y el desarrollo de sistemas de agricultura de precisión (FAO, 2022).

La integración de información climática, geográfica y productiva ha permitido desarrollar modelos predictivos capaces de apoyar la toma de decisiones en diferentes etapas del proceso agrícola. Estos modelos facilitan la planificación de siembras, la gestión de recursos hídricos y la evaluación de riesgos asociados a fenómenos climáticos extremos (FAO, 2022).

En este contexto, el concepto de agricultura inteligente (*Smart Farming*) integra tecnologías digitales, análisis de grandes volúmenes de datos (*Big Data*) y técnicas de aprendizaje automático para optimizar la productividad, mejorar la eficiencia en el uso de los recursos y fortalecer la toma de decisiones basada en datos en los sistemas agrícolas (Wolfert et al., 2017).

Diversas investigaciones han demostrado que algoritmos como Random Forest y XGBoost presentan un elevado desempeño en tareas de predicción agrícola debido a su capacidad para modelar relaciones complejas entre variables ambientales y productivas (Kamilaris & Prenafeta-Boldú, 2018; Van Klompenburg et al., 2020; Shahhosseini et al., 2021).

2.2.4. Algoritmos utilizados para la predicción

Regresión Lineal

La Regresión Lineal es uno de los métodos más utilizados en análisis predictivo debido a su simplicidad, interpretabilidad y facilidad de implementación.

Ecuación:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

Donde:

- y = variable dependiente o variable objetivo.
- β_0 = intercepto del modelo.
- β_i = coeficientes asociados a cada variable independiente.
- x_i = variables predictoras.
- ε = término de error aleatorio.

En investigaciones agrícolas recientes, la Regresión Lineal sigue siendo utilizada como línea base para evaluar el desempeño de algoritmos más avanzados, especialmente cuando se busca interpretar el efecto individual de los factores climáticos y productivos sobre los cultivos (Kuradusenge et al., 2023).

Random Forest

Random Forest es un algoritmo de aprendizaje supervisado propuesto por Breiman (2001), basado en la construcción de múltiples árboles de decisión que trabajan de manera conjunta para generar una predicción final. Cada árbol se construye utilizando muestras aleatorias de los datos y subconjuntos aleatorios de variables, lo que permite reducir la dependencia de un único modelo y mejorar la estabilidad de las predicciones.

La principal fortaleza de Random Forest radica en su capacidad para modelar relaciones complejas y no lineales entre las variables. Esto resulta especialmente útil en aplicaciones agrícolas, donde factores climáticos, productivos y temporales suelen interactuar de forma simultánea y con distintos niveles de influencia.

Además de su capacidad predictiva, Random Forest ofrece una característica de gran valor para la investigación científica: la estimación de la importancia de las variables. Esta funcionalidad permite identificar cuáles factores contribuyen en mayor medida a la predicción de la producción agrícola, facilitando la interpretación de los resultados obtenidos.

Entre sus principales ventajas destacan su capacidad para modelar relaciones no lineales, la robustez frente al ruido y valores atípicos, el menor riesgo de sobreajuste respecto a árboles individuales y la posibilidad de estimar la importancia relativa de las variables.

Diversos estudios han reportado resultados favorables utilizando Random Forest para la estimación del rendimiento de cultivos, el análisis de riesgos climáticos y la evaluación de la productividad agrícola. Kuradusenge et al. (2023) identificaron que este algoritmo alcanzó elevados niveles de precisión en la predicción del rendimiento agrícola utilizando variables meteorológicas y productivas.

XGBoost

XGBoost (*Extreme Gradient Boosting*) es una implementación optimizada del algoritmo *Gradient Boosting* desarrollada por Chen y Guestrin (2016). Su funcionamiento se basa en la construcción secuencial de árboles de decisión, donde cada nuevo árbol tiene como objetivo corregir los errores cometidos por los árboles generados previamente.

Actualmente, XGBoost es considerado uno de los algoritmos más utilizados en problemas de predicción debido a su elevada precisión y eficiencia computacional. A diferencia

de otros métodos basados en árboles, incorpora mecanismos de regularización que ayudan a controlar el sobreajuste y mejorar la capacidad de generalización del modelo.

Entre sus principales ventajas se encuentran su alta capacidad predictiva, la posibilidad de modelar relaciones complejas y no lineales entre variables, el manejo eficiente de valores faltantes, la escalabilidad para grandes conjuntos de datos y la optimización del tiempo de entrenamiento mediante procesamiento paralelo. Estas características han convertido a XGBoost en una de las técnicas más empleadas en aplicaciones de aprendizaje automático, especialmente en problemas donde intervienen múltiples variables con relaciones complejas.

La literatura reciente evidencia que XGBoost ha sido ampliamente utilizado en aplicaciones agrícolas relacionadas con la predicción de rendimientos, la estimación de cosechas y el análisis de variables climáticas. Jabed et al. (2024) reportan que los modelos basados en *Gradient Boosting* y XGBoost alcanzan un desempeño altamente competitivo cuando integran variables climáticas y factores productivos, evidenciando su capacidad para mejorar la precisión de las predicciones en sistemas agrícolas.

Redes Neuronales Artificiales (MLP)

Las Redes Neuronales Artificiales (Artificial Neural Networks, ANN) son modelos computacionales inspirados en el funcionamiento del sistema nervioso biológico. Estas redes están compuestas por unidades de procesamiento denominadas neuronas, organizadas en capas interconectadas que permiten aprender patrones complejos a partir de los datos (Goodfellow et al., 2016).

Dentro de las arquitecturas más utilizadas para problemas de regresión se encuentra el Perceptrón Multicapa (Multi-Layer Perceptron, MLP). Este modelo está formado por una capa de entrada, una o más capas ocultas y una capa de salida, donde cada neurona transforma la información recibida mediante funciones de activación no lineales.

El proceso de aprendizaje se realiza mediante la optimización de los pesos de conexión entre neuronas utilizando algoritmos como el descenso por gradiente (Gradient Descent) y el mecanismo de retropropagación del error (Backpropagation) (Goodfellow et al., 2016).

De manera general, la salida de una neurona puede expresarse como:

$$y = f(\Sigma(wi \cdot xi) + b)$$

Donde:

- *xi* = variables de entrada.
- *wi* = pesos asociados a cada entrada.
- *b* = es el sesgo (bias)
- *f* = función de activación.
- *y* = salida de la neurona.

Una de las principales ventajas de las redes neuronales es su capacidad para modelar relaciones altamente no lineales entre variables, permitiendo capturar patrones complejos que pueden no ser detectados mediante técnicas tradicionales. Esta característica resulta especialmente relevante en aplicaciones agrícolas, donde la producción puede verse afectada simultáneamente por múltiples factores climáticos, productivos y temporales.

Sin embargo, las redes neuronales presentan algunas limitaciones relacionadas con una menor interpretabilidad, mayores requerimientos computacionales y la necesidad de ajustar diversos hiperparámetros durante el entrenamiento.

2.2.5. Evaluación de modelos predictivos

La evaluación de modelos predictivos constituye una etapa fundamental del aprendizaje automático, ya que permite determinar qué algoritmo presenta el mejor desempeño para resolver un problema específico. En tareas de regresión supervisada, como la predicción de la producción agrícola, esta evaluación se realiza mediante métricas que cuantifican la diferencia entre los valores observados y los valores estimados por el modelo, permitiendo comparar objetivamente diferentes algoritmos (James et al., 2021).

Las métricas más utilizadas para evaluar modelos de regresión son el Error Absoluto Medio (Mean Absolute Error, MAE), la Raíz del Error Cuadrático Medio (Root Mean Squared Error, RMSE) y el coeficiente de determinación (R^2). Estas medidas proporcionan información complementaria sobre la precisión y la capacidad explicativa de los modelos, por lo que constituyen un estándar ampliamente aceptado en la evaluación del desempeño de técnicas de aprendizaje automático aplicadas a problemas de regresión (Chai & Draxler, 2014; James et al., 2021).

En las siguientes subsecciones se describen las principales características y fundamentos matemáticos de cada una de estas métricas.

Error Absoluto Medio (MAE)

El Error Absoluto Medio (Mean Absolute Error, MAE) mide el promedio de las diferencias absolutas entre los valores observados y los predichos por un modelo. Un menor valor de MAE indica una mayor precisión en las predicciones, al expresar el error promedio en las mismas unidades de la variable analizada (Willmott & Matsuura, 2005).

Ecuación:

$$MAE = \frac{1}{n}\sum |y_i - \hat{y}_i|$$

Donde:

- *n* = número total de observaciones.
- *yi* = valor real observado.
- *ŷi* = valor predicho por el modelo.
- |*yi* − *ŷi*| = error absoluto de cada observación.

Raíz del Error Cuadrático Medio (RMSE)

La Raíz del Error Cuadrático Medio (Root Mean Squared Error, RMSE) mide la magnitud promedio de los errores de predicción, otorgando mayor peso a los errores de gran magnitud. Por ello, resulta útil para identificar modelos que presentan desviaciones importantes respecto a los valores reales (Chai & Draxler, 2014).

Ecuación:

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

Donde:

- *n* = número total de observaciones.
- *yi* = valor real observado.
- *ŷi* = valor predicho por el modelo.
- *(yi − ŷi)²* = cuadrado del error de cada observación.

Coeficiente de Determinación (R^2)

El coeficiente de determinación (R^2) mide la proporción de la variabilidad de la variable objetivo que es explicada por el modelo predictivo. Valores cercanos a uno indican un mejor ajuste y una mayor capacidad explicativa del modelo (Glen, 2016).

Ecuación:

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

Donde:

- *yi* = valor real observado.
- *ŷi* = valor predicho por el modelo.
- $\bar{y}$ = promedio de los valores observados.
- $\Sigma(yi - \hat{y}i)^2$ = suma de cuadrados de los errores de predicción.
- $\Sigma(yi - \bar{y})^2$ = suma total de cuadrados de la variable observada.

2.2.6. Variabilidad Climática

Concepto y diferencias con cambio climático

Concepto Variabilidad Climática

La variabilidad climática se refiere a las fluctuaciones naturales del clima alrededor de sus valores promedio históricos, que ocurren de un año a otro, de una década a otra o incluso entre estaciones del mismo año. Estas oscilaciones forman parte del comportamiento natural del sistema climático y pueden modificar los patrones de temperatura y precipitación. En el sector agrícola, estas variaciones influyen directamente en el desarrollo de los cultivos, la disponibilidad de recursos hídricos y la productividad, generando incertidumbre en la planificación y la toma de decisiones (Rosenzweig et al., 2001; IPCC, 2023).

Cambio Climático

El cambio climático es una alteración sistemática y sostenida de los promedios climáticos a lo largo de décadas, principalmente causada por la actividad humana (emisión de gases de efecto invernadero). No es una oscilación temporal, sino una tendencia de largo plazo que desplaza el promedio histórico hacia nuevos valores.

Tabla 1

Características de la variabilidad climática y el cambio climático

Característica	Variabilidad Climática	Cambio Climático
Escala temporal	Años o décadas	Décadas o siglos
Causa principal	Procesos naturales (ENSO, volcanes...)	Actividad humana (GEI)
Efecto	Fluctuación alrededor del promedio	Desplazamiento del promedio
Reversibilidad	Sí, natural	No, persistente
Ejemplo en el dataset	temp_anomalia varía +4.2°C a -3.7°C por año	Tendencia al alza en temp_ma3 a largo plazo

Nota. Elaboración propia a partir de la literatura revisada. La tabla compara las principales diferencias entre la variabilidad climática y el cambio climático en función de su escala temporal, causas y efectos.

Indicadores de variabilidad climática

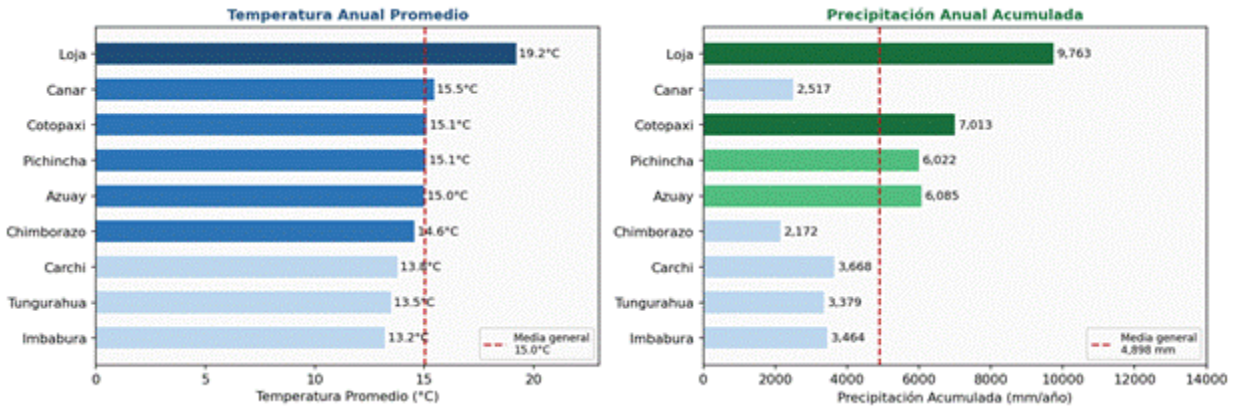
Para cuantificar cuánta variabilidad climática existe, se utilizan indicadores estadísticos que permiten comparar entre provincias y años. El dataset incorpora los siguientes indicadores, cada uno con un significado específico:

Temperatura promedio y precipitación acumulada (línea base)

Antes de medir la variabilidad, necesitamos conocer el valor "normal" de referencia para cada provincia. Estas columnas representan el clima base del año analizado.

Figura 1

Temperatura promedio y precipitación acumulada por provincia (2002–2023)



Nota. La figura muestra los valores promedio de temperatura y precipitación acumulada registrados entre 2002 y 2023 para las provincias de la región Sierra del Ecuador. Elaboración propia con datos del INAMHI (2023).

Observación: Loja es la más cálida (19.24°C) y Carchi la más fría (13.82°C). Loja también registra la mayor precipitación (9 763 mm/año).

Desviación estándar (temp_std, precip_std)

La desviación estándar mide cuánto varían los valores mensuales de temperatura o precipitación dentro de un mismo año. Un valor alto significa que el clima fue muy irregular durante ese año (meses muy fríos alternados con meses muy cálidos, o temporadas de lluvias intensas seguidas de sequías prolongadas).

Ecuación**:**

$$std = \sqrt{\frac{\sum(valor\_mes - promedio\_anual)^2}{12}}$$

Ejemplo: En Chimborazo, temp_std promedio = 0.97°C (muy estable durante el año), mientras que en Cañar = 3.31°C (mucho más variable mes a mes). Esta diferencia se debe a la altitud y la influencia de las corrientes de aire.

Coeficiente de variación (temp_cv, precip_cv)

El coeficiente de variación (CV): es la desviación estándar dividida por el promedio. Su ventaja es que permite comparar la variabilidad entre provincias, aunque tengan escalas de temperatura muy diferentes.

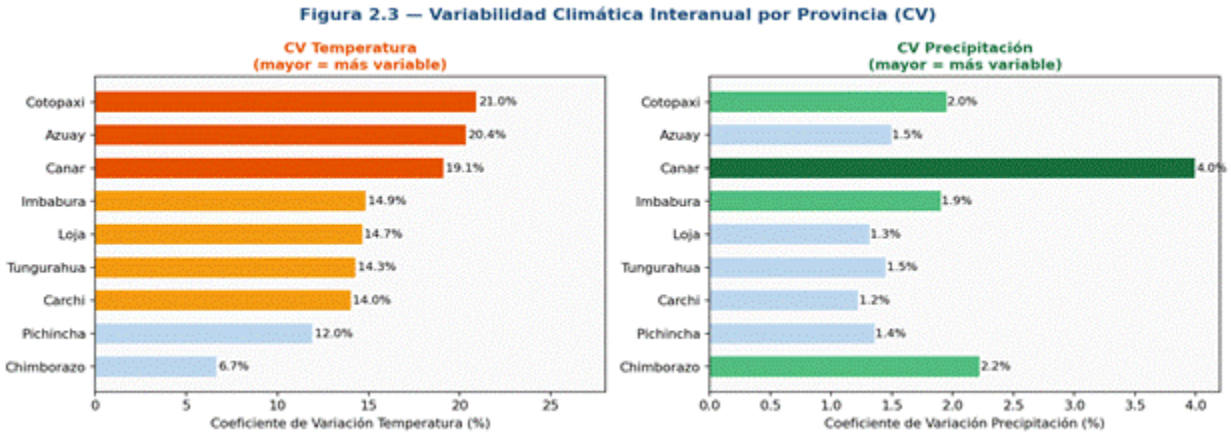
Ecuación**:**

$$CV = \frac{Desviación\ estándar}{Media}$$

Un CV de 0.20 significa que la variación típica es el 20% del promedio.

Figura 2

Variabilidad climática interanual por provincia (coeficiente de variación, CV)



Nota. La figura presenta el coeficiente de variación (CV) de la temperatura promedio y la precipitación acumulada por provincia durante el período 2002–2023. Elaboración propia con datos del INAMHI (2023).

Anomalías climáticas (temp_anomalia, precip_anomalia)

Las anomalías son la diferencia entre el valor del año actual y el promedio histórico de la provincia. Son el indicador más directo de variabilidad climática: indican si un año fue más frío/caliente o más húmedo/seco de lo habitual.

Temp_anomalia mide la variabilidad climática: cuánto se aleja la temperatura de un año respecto al promedio histórico de cada provincia. El rango observado es de -3.71°C (Cañar 2021) a +4.20°C (Azuay 2016), lo que refleja oscilaciones importantes.

Temp_ma3 (promedio móvil de 3 años) permite detectar si existe una tendencia de cambio climático subyacente.

Ecuación*:*

$$temp\_anomalia(t) = temp(t) - promedio\_historico\_provincia$$

$$precip\_anomalia(t) = precip(t) - promedio\_historico\_provincia$$

ENSO y su rol en la variabilidad climática

Uno de los principales mecanismos que genera variabilidad climática interanual en Ecuador es el fenómeno ENSO (El Niño–Oscilación del Sur). Aunque tiene origen en el océano Pacífico ecuatorial, su influencia modifica los patrones de temperatura y precipitación en amplias regiones del mundo, incluyendo el territorio ecuatoriano, afectando directamente las condiciones de producción agrícola (National Oceanic and Atmospheric Administration [NOAA], 2023; IPCC, 2023).

Tabla 2

Categorías del índice ENSO y su efecto típico en la Sierra del Ecuador

Categoría	Rango del índice	Efecto típico
El Niño	> +0.5	Sequías en sierra, lluvias excesivas en costa
Neutral	-0.5 a +0.5	Condiciones climáticas normales
La Niña	< -0.5	Más lluvias en sierra, posibles heladas

Nota. Elaboración propia a partir de la clasificación del índice ENSO. La tabla presenta las categorías del fenómeno El Niño–Oscilación del Sur (ENSO), sus rangos de clasificación y los efectos climáticos típicos asociados en la región Sierra del Ecuador.

En el dataset, de los 594 registros:

- 243 registros (40.9%) corresponden a años Neutrales
- 216 registros (36.4%) corresponden a años El Niño
- 135 registros (22.7%) corresponden a años La Niña

Impacto cuantificado del ENSO en la producción

Tabla 3

Impacto del fenómeno ENSO en la producción promedio de los cultivos analizados

Cultivo	Media con El Niño	Media con La Niña	Media Neutral	¿Quién produce más?
Papa	42 620 t	37 404 t	47 156 t	Neutral
Maíz suave seco	3 424 t	2 796 t	3 680 t	Neutral
Arveja tierna	1 217 t	1 440 t	1 375 t	La Niña (leve)

Nota. Elaboración propia con base en los datos históricos de producción y la clasificación de las fases del ENSO para el período de estudio. La tabla presenta la producción promedio de papa, maíz suave seco y arveja tierna durante las fases El Niño, La Niña y condiciones neutrales, permitiendo comparar la respuesta de cada cultivo frente a la variabilidad climática asociada al ENSO.

Impactos en sistemas agrícolas

La variabilidad climática no afecta a todos los cultivos ni a todas las provincias de la misma forma. Los impactos dependen de la intensidad del fenómeno, la altitud del área, el tipo de cultivo y la resiliencia de los sistemas agrícolas locales.

Impacto en la temperatura

Los cultivos de la Sierra tienen rangos óptimos de temperatura. Cuando la temperatura se aleja de ese rango, el rendimiento cae:

Tabla 4

Rangos óptimos de temperatura e impactos potenciales por temperaturas extremas en los cultivos analizados

Cultivo	**Rango óptimo (°C)**	**Riesgo por calor excesivo**	**Riesgo por frío extremo**
Papa (tubérculo fresco)	10°C – 18°C	Brote prematuro, deformación	Heladas, pérdida total
Maíz suave seco	16°C – 22°C	Ciclo acortado, grano pequeño	Germinación lenta
Arveja tierna	15°C – 20°C	Aborto de flores	Podredumbre de semilla

Nota. Elaboración propia a partir de la literatura especializada sobre los requerimientos térmicos de los cultivos. La tabla presenta los rangos óptimos de temperatura para la papa, el maíz suave seco y la arveja tierna, así como los principales efectos potenciales asociados a condiciones de calor excesivo y frío extremo.

Evidencia del dataset: La correlación entre temp_promedio y producción es de -0.335 para papa y arveja, y +0.076 para maíz. Esto significa que años más cálidos tienden a reducir la producción de papa y arveja, pero el maíz es más tolerante a temperaturas altas.

Impacto en la precipitación

La precipitación afecta directamente la disponibilidad de agua para el riego natural (lluvia), la humedad del suelo y el riesgo de enfermedades fúngicas (hongos). Tanto el exceso como el déficit son perjudiciales.

La correlación entre precipitación y producción es de -0.173 para arveja y -0.178 para papa, lo que indica que las provincias más lluviosas no son necesariamente las más productivas: las precipitaciones muy elevadas generan problemas fitosanitarios.

La interacción ENSO por clima: variables de interacción

Una forma avanzada de capturar el impacto combinado de ENSO y el clima local es a través de las variables de interacción del dataset:

Tabla 5

Variables de interacción entre el índice ENSO y las condiciones climáticas locales

Variable	Ecuación	Qué captura
enso_precip_interaccion	enso_index × precipitacion_acumulada	Amplificación del ENSO en zonas lluviosas. Si ambas son altas (El Niño + lluvia), el impacto es máximo.
enso_temp_interaccion	enso_index × temp_promedio	Estrés térmico potenciado. En Loja (19°C) el mismo índice ENSO genera más estrés que en Carchi (13°C).

Nota. Elaboración propia. La tabla describe las variables de interacción generadas mediante ingeniería de características para representar el efecto combinado del índice ENSO con la

precipitación acumulada y la temperatura promedio, con el fin de mejorar la capacidad predictiva de los modelos de aprendizaje automático.

Efectos retardados: los lags climáticos

Un aspecto crucial que diferencia el análisis climático simple del análisis para Machine Learning es que los impactos climáticos no siempre ocurren en el mismo año en que se manifiesta el fenómeno. Los efectos pueden retrasarse uno o dos años:

- Un invierno muy seco (La Niña) agota las reservas hídricas del suelo → el siguiente año también tendrá déficit, aunque llueva normal.
- Una helada severa destruye semillas disponibles para el próximo ciclo agrícola.
- Un El Niño que destruye la cosecha actual reduce los recursos del agricultor para la siguiente siembra.

Por eso el dataset incluye variables como temp_lag1, temp_lag2, precip_lag1, precip_lag2, que registran el clima de hasta dos años atrás para que el modelo de Machine Learning pueda detectar estos efectos retardados.

2.2.7. Producción Agrícola y Factores Determinantes

Producción agrícola como indicador

- *La producción agrícola* (medida en toneladas métricas) es la cantidad total de un cultivo obtenida en una provincia durante un año específico. No debe confundirse con:
- *Rendimiento:* producción por hectárea. Mide la eficiencia y optimización del uso del suelo, mas no el volumen total cosechado.
- *Superficie cosechada (ha):* cuánta área del territorio sembrado se logró recolectar exitosamente al final del ciclo productivo.

- *Producción potencial:* el volumen máximo teórico que se podría obtener bajo condiciones de laboratorio o entornos ideales sin plagas ni anomalías climáticas.
- ¿Por qué usamos *producción total* y no *rendimiento* como variable objetivo? Porque la producción total integra tanto el efecto climático como el efecto de la superficie cultivada, reflejando el impacto real sobre la disponibilidad de alimentos. El rendimiento puede ser alto, aunque la producción total sea pequeña si el área cultivada fue muy reducida.

Tabla 6

Comparación entre el rendimiento agrícola y la producción total como indicadores de evaluación

Indicador	¿Qué mide?	Ventaja	Limitación
Rendimiento (Eficiencia)	Qué tan bien se aprovechó cada hectárea.	Ideal para evaluar tecnología o calidad del suelo.	Puede ser altísimo, pero si solo se sembró una hectárea, el impacto real en el mercado es insignificante.
Producción Total (Volumen)	La cantidad total de alimento disponible.	Refleja el impacto real en el abastecimiento y la economía.	No te dice por sí solo si el agricultor fue eficiente o si simplemente usó muchísima tierra.

Nota. Elaboración propia a partir de la literatura revisada. La tabla compara el rendimiento agrícola y la producción total como indicadores de evaluación, destacando las variables que representan, sus principales ventajas y las limitaciones asociadas a su uso en el análisis de la producción agrícola.

2.2.8. Factores Climáticos y Productivos

La producción agrícola es el resultado de la interacción de múltiples factores simultáneos. Para los modelos de Machine Learning, todos estos factores deben estar representados como columnas (variables independientes) del dataset.

Factores climáticos

- Temperatura promedio anual (temp_promedio): correlación de -0.335 con producción de papa y arveja. Años más cálidos tienden a reducir la producción de tubérculos y legumbres de altura.
- *Precipitación acumulada (precipitacion_acumulada):* correlación de +0.265 con maíz (más lluvia = más producción) pero negativa para papa (-0.178). El exceso de agua perjudica a la papa.
- *Variabilidad climática (temp_std, precip_std):* años con alta variación interna (muchos meses extremos) son más perjudiciales que años con clima estable, aunque ese promedio sea subóptimo.
- *Anomalías (temp_anomalia, precip_anomalia):* indican qué tan alejado estuvo el año del patrón histórico. Son la señal más directa de "año climáticamente inusual".
- *ENSO (enso_index, enso_categoria):* la señal climática global más importante para Ecuador. Se incorpora como variable continua y categórica para que los modelos detecten el régimen climático.

Factores productivos

- Superficie plantada (sup_plantada): refleja la decisión del agricultor de cuánto sembrar. Si aumenta, indica confianza en las condiciones del mercado y el clima.
- *Superficie cosechada (sup_cosechada):* la diferencia entre sup_plantada y sup_cosechada indica pérdidas por clima, plagas o factores productivos. Se

incorpora explícitamente en el modelo para evitar sesgo en la interpretación climática.

- *Rendimiento (rendimiento = produccion / sup_cosechada):* la eficiencia productiva. Un año puede tener baja producción total por poca área, pero alto rendimiento por hectárea, señalando buenas condiciones climáticas.

¿Cuál es el factor con mayor influencia en la producción agrícola?

Según las correlaciones obtenidas entre las variables predictoras y la producción agrícola (Tabla 7):

Tabla 7

Correlación entre las variables predictoras y la producción agrícola

Variable	Correlación con Producción	Interpretación
produccion_lag1	+0.817 (papa), +0.709 (maíz)	El predictor más fuerte: el año pasado predice el actual
temp_promedio	-0.335 (papa y arveja)	Más calor = menos papa/arveja
precipitacion_acumulada	+0.265 (maíz), -0.178 (papa)	Depende del cultivo
enso_index	+0.044 a +0.084	Efecto moderado directo, mayor en interacción
temp_anomalia	-0.057	Años anómalos afectan negativamente

Nota. Elaboración propia con base en el análisis de correlación realizado sobre el conjunto de datos del período 2002–2023. La tabla presenta los coeficientes de correlación entre las

principales variables predictoras y la producción agrícola, así como una interpretación de la dirección e intensidad de su relación con la variable objetivo.

Observación: La producción del año anterior (lag1) es el predictor individual más potente. Esto se debe a la inercia del sistema agrícola: si hubo buenas condiciones y alta producción un año, es probable que continúen. Esta es la razón más importante por la que las variables de lag son fundamentales en el dataset.

2.2.9. Influencia de variables temporales en la agricultura

A diferencia de las variables estáticas (temperatura del año, precipitación del año), las variables temporales capturan la historia reciente del sistema, que es información clave para los modelos predictivos.

Variables de rezago temporal (lags)

Valores de una variable que pertenecen a períodos anteriores (1 o 2 años anteriores) y que se utilizan en el presente para predecir o explicar un comportamiento actual, utilizando un modelo de análisis o de Machine Learning para que pueda entender la historia de los datos.

Ecuación:

$$lag1(X,t) = X(t-1)$$

$$lag2(X,t) = X(t-2)$$

Donde:

- X: variable de interés.
- t: período (año) de referencia.
- X(t−1): valor de la variable en el año anterior.
- X(t−2): valor de la variable dos años antes.

Ejemplo (Azuay, cultivo de arveja, año 2004):

$$temp\_lag1(2004) = temp\_promedio(2003) = 13.00\ °C$$

$$precip\_lag1(2004) = precip(2003) = 8085.85\ mm$$

$$produccion\_lag1(2004) = produccion(2003) = 557\ t$$

Tabla 8

Cálculo de variables lag.

Variable lag	Ecuación	Valor (Azuay, arveja, 2004)	Justificación
temp_lag1	temp_promedio(t−1)	13.00 °C	Condiciones térmicas del ciclo anterior
precip_lag1	precipitacion_acumulada (t−1)	8 085.85 mm	Humedad residual del suelo
produccion_lag1	produccion(t−1)	557 t	Inercia productiva: r = 0.817 con producción actual (papa)
rendimiento_lag1	rendimiento(t−1)	0.3914 t/ha	Eficiencia del ciclo anterior
temp_lag2	temp_promedio(t−2)	13.37 °C	Efectos climáticos de 2 años atrás
precip_lag2	precipitacion_acumulada (t−2)	8 829.94 mm	Ciclos de recarga hídrica prolongados

Nota. El primer año de cada serie provincia + cultivo (2002) presenta NaN en lag1; el segundo año (2003), en lag2. Esto es esperado y se gestiona en la fase de preprocesamiento del modelo.

Tasas de crecimiento interanual

Las variables *crecimiento_produccion y crecimiento_superficie* capturan la dinámica de cambio: si el sistema agrícola está en expansión o contracción.

Ecuación:

$$crecimiento(t) = \frac{valor(t) - valor(t-1)}{valor(t-1)}$$

- Un valor positivo (ej: +0.41) indica que la producción creció un 41% respecto al año anterior.
- Un valor negativo (ej: -0.51) indica una caída del 51%.
- El rango en el dataset va de -0.92 a +272%, lo que refleja la extrema volatilidad del sector.

Importancia para el modelo

Estas variables permiten al modelo detectar si el sistema está en una tendencia creciente o decreciente antes de que se estabilice. Un año con caída del 50% en producción es una señal de alerta que el modelo puede anticipar.

Promedios móviles de 3 años (MA3)

Los promedios móviles suavizan los picos y valles anuales para revelar la *tendencia estructural de mediano plazo*. Son útiles para distinguir si una variación anual es ruido (evento puntual) o tendencia (patrón sostenido).

Ecuación:

$$MA3(t) = \frac{valor(t-2) + valor(t-1) + valor(t)}{3}$$

Tabla 9

Promedios móviles de tres años (MA3) de las variables analizadas

Variable MA3	¿Qué tendencia captura?	Ejemplo en el dataset
produccion_ma3	¿Está aumentando la capacidad productiva de la zona?	Chimborazo papa: 75 891 t (2004) → 92 341 t (2014)
precip_ma3	¿Está entrando en un ciclo más húmedo o más seco?	Cotopaxi: tendencia a más precipitación desde 2012
temp_ma3	¿Hay una tendencia de calentamiento estructural?	Carchi: temp_ma3 sube de 12.4°C a 13.2°C entre 2002-2010

Nota. Elaboración propia con base en los datos históricos del período 2002–2023. La tabla presenta las variables calculadas mediante promedios móviles de tres años (MA3), las tendencias que representan y ejemplos de su comportamiento en el conjunto de datos, permitiendo identificar cambios graduales en la producción agrícola y en las condiciones climáticas

2.2.10. Vulnerabilidad agrícola frente a variabilidad climática

La *vulnerabilidad agrícola* es la susceptibilidad de un cultivo o región a sufrir pérdidas ante cambios en el clima. No todas las provincias ni todos los cultivos son igualmente vulnerables.

Factores que determinan la vulnerabilidad

- *Altitud:* provincias en altitudes muy altas (Carchi, Chimborazo > 3 000 msnm) son más susceptibles a heladas cuando la temperatura cae bajo lo normal.

- *Diversificación de cultivos:* provincias que solo cultivan papa son más vulnerables que las que diversifican. Carchi depende en un 95% de la papa.
- *Dependencia del agua de lluvia:* provincias sin sistemas de riego dependen completamente de la precipitación anual, que puede variar ±50% según el año.
- *Capacidad de adaptación:* acceso a semillas mejoradas, técnicas de manejo de suelos y asesoría técnica reducen la vulnerabilidad.

Análisis de vulnerabilidad por provincia en el dataset

A continuación, mediante una tabla se presenta una evaluación comparativa de la vulnerabilidad de las provincias de la región Sierra del Ecuador, considerando la variabilidad de la producción, el rango de anomalías de temperatura y el cultivo predominante, con el fin de identificar los territorios con mayor susceptibilidad frente a la variabilidad climática.

Tabla 10

Análisis de vulnerabilidad de las provincias de la región Sierra del Ecuador a partir de variables productivas y climáticas

Provincia	Cultivo principal	CV producción	Anomalía temp. (rango)	Nivel vulnerabilidad
Carchi	Papa (85%)	Alta (±82 275 t)	4.60°C	MUY ALTA (escala extrema)
Cañar	Papa + Maíz	Alta (±8 335 t)	6.78°C	ALTA (mayor variabilidad temp)
Azuay	Maíz + Papa	Media (±5 168 t)	6.42°C	ALTA (lluvia muy variable)
Cotopaxi	Papa + Maíz	Media (±18 950 t)	3.12°C	MEDIA (mayor estabilidad)
Tungurahua	Maíz + Arveja	Baja (±17 312 t)	1.04°C	BAJA (más estable)
Chimborazo	Papa	Media (±24 228 t)	3.20°C	MEDIA

Nota. Aunque Carchi tiene la mayor producción promedio (168 489 t/año en papa), también tiene la mayor desviación estándar (82 275 t). Su producción puede ir desde 53 492 t (2002) hasta 332 030 t (2020), una variación de más de 6 veces. Esta extrema vulnerabilidad hace que Carchi sea el caso más desafiante para el modelo predictivo.

Impacto combinado: el año 2010 (La Niña fuerte)

El año 2010 se caracterizó por registrar el índice ENSO más bajo del período analizado (-1.32), lo que técnicamente se clasificó como un evento de La Niña Fuerte.

Impactos Climáticos y Fitosanitarios

- *Exceso de lluvias:* Provocó precipitaciones que superaron el promedio histórico en múltiples provincias de la Sierra.
- *Riesgos de enfermedades:* El aumento de humedad disparó la probabilidad de brotes de rancha (hongo que ataca severamente al cultivo de papa) y otras enfermedades fúngicas.

Comportamiento de la Producción (Resiliencia del Sector)

- *Caso Carchi:* Registró una producción de 167,427 t (ligeramente inferior a su promedio histórico de 168,489 t), lo que demostró una gran capacidad de resistencia del sistema agrícola local.
- *Sierra en general:* La producción total de la región alcanzó las 417,658 t, manteniéndose muy cerca del promedio estimado para la zona (422,000 t).
- *Conclusión:* Este escenario evidencia que un fenómeno de La Niña no necesariamente equivale a una pérdida masiva de cosechas, sino que reconfigura y traslada los factores de riesgo.

¿Cómo captura el modelo de Machine Learning esta vulnerabilidad?

Para que un modelo de ML pueda anticipar estos escenarios de riesgo modificados por el clima, se apoya en variables clave del dataset:

- *enso_index y enso_categoria:* Identifican directamente la presencia y la intensidad del fenómeno (en este caso, un valor de -1.32 categorizado como "nina").
- *precip_anomalia y temp_anomalia:* Miden qué tan extremo fue el clima ese año respecto al promedio histórico provincial (capturando el exceso de humedad que propicia la rancha).

- *enso_precip_interaccion:* Una variable crucial que multiplica el índice ENSO por la precipitación acumulada, permitiendo al modelo entender que mucha lluvia combinada con La Niña genera un riesgo fitosanitario no lineal.

Tabla 11

Variables utilizadas para representar la vulnerabilidad en el modelo de aprendizaje automático

Aspecto de vulnerabilidad	**Variable(s) del dataset que lo captura**	**Grupo**
Efectos climáticos acumulados	temp_lag1, temp_lag2, precip_lag1, precip_lag2	Lags (G6)
Tendencia de producción	produccion_lag1, rendimiento_lag1	Lags (G6)
Recuperación o caída	crecimiento_produccion, crecimiento_superficie	Crecimiento (G7)
Señal climática global	enso_index, enso_lag1, enso_lag2	ENSO (G8)
Tendencia estructural	produccion_ma3, precip_ma3, temp_ma3	MA3 (G9)
Impacto amplificado de ENSO	enso_precip_interaccion, enso_temp_interaccion	Interacciones (G10)

Nota. Elaboración propia. La tabla presenta las variables derivadas incorporadas al modelo de aprendizaje automático para representar diferentes dimensiones de la vulnerabilidad agrícola, incluyendo efectos climáticos acumulados, tendencias productivas, recuperación o disminución de la producción, influencia del fenómeno ENSO, tendencias estructurales e interacciones entre variables climáticas.

CAPÍTULO 3

3. DESARROLLO

3.1. Metodología

La presente investigación se desarrolló mediante un enfoque cuantitativo, debido a que utiliza datos históricos de producción agrícola y variables climáticas para construir y evaluar modelos predictivos mediante técnicas de aprendizaje automático. El estudio presenta un diseño no experimental y longitudinal, ya que analiza información correspondiente al período 2002–2023 sin manipular las variables de estudio, con el propósito de identificar patrones y estimar la producción agrícola de cultivos estratégicos de la región Sierra del Ecuador.

La metodología comprende varias etapas secuenciales que incluyen la recopilación e integración de datos provenientes de fuentes oficiales, la depuración y preparación del conjunto de datos, la construcción de variables derivadas relacionadas con la variabilidad climática y las dinámicas temporales, el análisis exploratorio de los datos, el entrenamiento y comparación de diferentes modelos de aprendizaje automático y, finalmente, la evaluación de su desempeño mediante métricas de regresión. Este procedimiento metodológico permite garantizar la calidad de los datos utilizados y seleccionar el modelo con mayor capacidad predictiva para estimar la producción agrícola.

Tabla 12

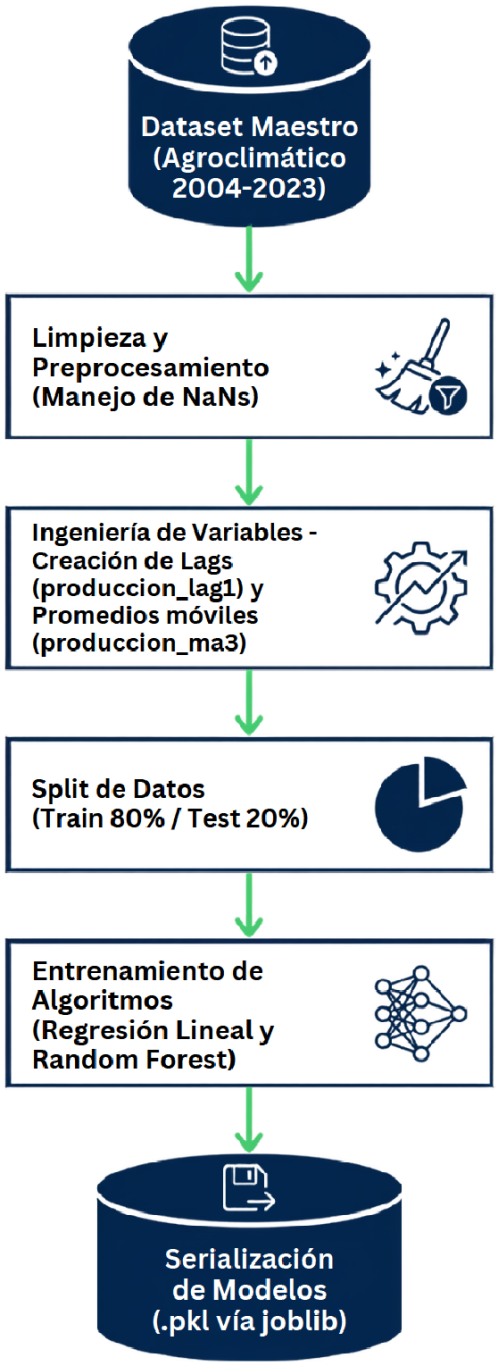
Resumen metodológico de la investigación

Elemento	Descripción
Enfoque	Cuantitativo
Diseño	No experimental, longitudinal y predictivo
Población de estudio	Registros históricos de producción agrícola y variables climáticas correspondientes a los cultivos de papa, maíz suave seco y arveja en las provincias de la región Sierra del Ecuador.
Unidad de análisis	Provincia – Año – Cultivo
Periodo de estudio	2002–2023
Área geográfica	Provincias de la región Sierra del Ecuador
Variable dependiente	Producción agrícola (toneladas)
Variables independientes	Temperatura, precipitación, superficie plantada, ENSO y variables temporales
Técnicas de análisis	Análisis exploratorio, ingeniería de variables y aprendizaje automático
Modelos implementados	Regresión Lineal, Random Forest, XGBoost y MLP
Métricas de evaluación	MAE, RMSE y R^2

Nota. Elaboración propia a partir del diseño metodológico de la investigación.

Figura 3

Pipeline metodológico del sistema de predicción



Nota. La figura muestra las etapas que conforman el flujo metodológico del sistema de predicción, desde la recopilación y el procesamiento de los datos hasta el entrenamiento, la evaluación de los modelos y la generación de predicciones. Elaboración propia.

La arquitectura del sistema se ha diseñado bajo un enfoque modular que garantiza la trazabilidad y reproducibilidad de los resultados. Como se ilustra en la Figura 3 (Pipeline metodológico del sistema de predicción), el proceso integra desde la ingesta del dataset agroclimático hasta la serialización de los modelos optimizados, permitiendo una transición fluida entre la fase de entrenamiento y la inferencia en tiempo real

A partir de esta metodología, en los apartados siguientes se describe el proceso de recopilación e integración de la información, la preparación del conjunto de datos, la construcción de variables, el desarrollo de los modelos predictivos y la evaluación de su desempeño.

3.2. Desarrollo del Trabajo

Con base en la metodología descrita anteriormente, el desarrollo de la investigación se llevó a cabo mediante una serie de etapas secuenciales que incluyeron la obtención de datos, el análisis exploratorio, la preparación del conjunto de datos, la ingeniería de variables y el desarrollo de los modelos predictivos. Cada una de estas etapas se describe en los apartados siguientes.

3.2.1. Obtención y Selección de las Fuentes de Información

La primera etapa de la investigación consistió en la identificación y recopilación de información proveniente de fuentes oficiales del Ecuador que permitieran analizar la relación existente entre la variabilidad climática y la producción agrícola de los cultivos seleccionados.

La información utilizada en la presente investigación fue obtenida del Sistema de Información Pública Agropecuaria (SIPA) del Ministerio de Agricultura y Ganadería (MAG) del Ecuador, plataforma oficial que integra y difunde información estadística del sector agropecuario generada por instituciones públicas. Debido a su carácter oficial, cobertura nacional y actualización periódica, el SIPA constituye una fuente confiable para el desarrollo de investigaciones relacionadas con la producción agrícola y las condiciones agroclimáticas del país (Ministerio de Agricultura y Ganadería [MAG], 2024).

A través de esta plataforma se recopilaron dos conjuntos de datos principales:

- Información productiva, correspondiente a superficie plantada, superficie cosechada y producción agrícola de los cultivos analizados, proveniente de los registros estadísticos publicados por el Instituto Nacional de Estadística y Censos (INEC).
- Información climática, correspondiente a registros históricos de temperatura y precipitación, disponibles en la sección de estadísticas agroclimáticas del SIPA.

La selección de esta fuente respondió a los siguientes criterios:

- Disponibilidad de información oficial y de acceso público.
- Cobertura histórica continua para el período de estudio.
- Información desagregada por provincia.
- Compatibilidad entre las variables productivas y climáticas requeridas para el desarrollo del modelo predictivo.

Con el propósito de garantizar la validez y confiabilidad de la información utilizada, se verificó la integridad de los registros descargados, la consistencia temporal de las series

históricas y la correspondencia entre las variables productivas y climáticas disponibles en la plataforma. Asimismo, durante la etapa de preparación de los datos se identificaron valores faltantes, registros inconsistentes y posibles duplicidades, los cuales fueron tratados mediante procesos de depuración, integración y validación antes de la construcción del dataset final.

Estas actividades permitieron disponer de un conjunto de datos consistente y adecuado para el entrenamiento y evaluación de los modelos de aprendizaje automático.

La utilización de una plataforma única de consulta facilitó la integración de la información productiva y climática, redujo posibles inconsistencias entre conjuntos de datos y aseguró la homogeneidad de la información utilizada durante el desarrollo de la investigación. Como resultado, se obtuvo una base de datos sólida y confiable para la construcción y evaluación de los modelos predictivos de producción agrícola en la región Sierra del Ecuador.

Tabla 13

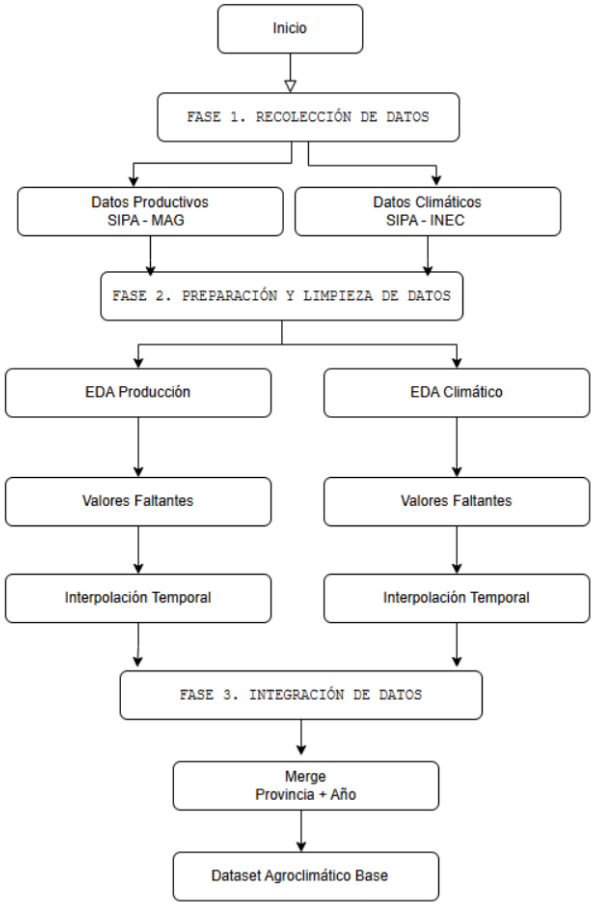
Fuentes de información y variables utilizadas en el estudio

Información	Fuente disponible en SIPA	Variables principales	Periodo
Productiva	Superficie y Producción - INEC	Producción, superficie plantada y superficie cosechada	2002 - 2023
Climática	Estadísticas agroclimáticas del SIPA	Temperatura y precipitación	2000 - 2023

Nota. Elaboración propia con base en la información publicada por el Sistema de Información Pública Agropecuaria (SIPA) del Ministerio de Agricultura y Ganadería (MAG, 2024).

Figura 4

Flujo general de construcción del conjunto de datos



Nota. La figura muestra el proceso de construcción del conjunto de datos, incluyendo la recopilación, integración, depuración, transformación y consolidación de la información utilizada para el desarrollo de los modelos predictivos. Elaboración propia.

3.2.2. Análisis Exploratorio de Datos Productivos

Una vez obtenida la información agrícola, se desarrolló un proceso de análisis exploratorio de datos (EDA) con el objetivo de comprender la estructura del conjunto de datos, identificar posibles inconsistencias y evaluar la calidad de la información disponible.

El dataset inicial estuvo conformado por aproximadamente 28.669 registros correspondientes a diferentes cultivos y provincias del Ecuador.

Durante esta etapa se analizaron variables relacionadas con:

- Año.
- Producto.
- Provincia.
- Superficie plantada.
- Superficie cosechada.
- Producción agrícola.

El análisis permitió identificar registros incompletos, diferencias en la cobertura temporal y posibles valores atípicos. Asimismo, se observó que el año 2024 presentaba una cantidad significativa de información faltante, razón por la cual fue excluido del estudio para mantener la consistencia temporal de la investigación. Estos hallazgos constituyeron la base para la etapa de preparación de los datos, en la cual se aplicaron procedimientos de depuración y tratamiento de valores faltantes antes de la integración del conjunto de datos final.

Tabla 14

Características del conjunto de datos productivos original

Característica	Valor
Registros iniciales	28.669
Periodo analizado	2002–2023
Nivel de agregación	Provincia – Año – Cultivo
Variables principales	Producción, superficie plantada, superficie cosechada

Nota. Elaboración propia. La tabla presenta las principales características del conjunto de datos productivos utilizado en la investigación, incluyendo el número de registros, el período de análisis, el nivel de agregación y las variables consideradas para el desarrollo de los modelos predictivos.

3.2.3. Análisis Exploratorio de Datos Climáticos

De forma paralela se realizó el análisis exploratorio de la información climática.

El dataset climático estuvo compuesto por aproximadamente 25.668 registros correspondientes a variables meteorológicas de temperatura y precipitación.

Las principales variables analizadas fueron:

- Año
- Mes
- Provincia
- Cantón

- Temperatura
- Precipitación

El análisis permitió verificar la cobertura temporal de la información y detectar valores faltantes en algunas provincias.

Posteriormente, la información mensual fue agregada a nivel anual para generar variables representativas de las condiciones climáticas observadas durante cada año agrícola.

La agregación anual permitió homogeneizar la escala temporal de la información climática con los registros productivos, facilitando su integración y el desarrollo posterior del modelo predictivo.

Tabla 15

Características del conjunto de datos climáticos original

Característica	Valor
Registros iniciales	25.668
Variables principales	Temperatura, precipitación
Periodo analizado	2002–2023
Nivel de agregación final	Provincia – Año

Nota. Elaboración propia. La tabla presenta las principales características del conjunto de datos climáticos utilizado en la investigación, incluyendo el número de registros, las variables analizadas, el período de estudio y el nivel de agregación empleado para su integración con los datos productivos.

3.2.4. Tratamiento de Valores Faltantes

El análisis exploratorio permitió verificar la cobertura temporal de la información e identificar la presencia de valores faltantes tanto en los datos productivos como en los climáticos. La gestión adecuada de estos valores constituye una etapa fundamental en los procesos de análisis y modelado predictivo, ya que puede influir directamente en la calidad, precisión y capacidad de generalización de los modelos desarrollados (Little & Rubin, 2019).

En investigaciones basadas en series temporales, la selección del método de imputación debe fundamentarse en criterios metodológicos que permitan preservar la estructura y el comportamiento de los datos originales. De acuerdo con Little y Rubin (2019), el tratamiento de valores faltantes debe minimizar la introducción de sesgos y conservar la mayor cantidad posible de información disponible, mientras que Moritz y Bartz-Beielstein (2017) señalan que, para datos con dependencia temporal, es recomendable utilizar métodos que mantengan la continuidad y las tendencias de la serie.

Con el propósito de preservar la mayor cantidad posible de información y minimizar sesgos en el conjunto de datos, se evaluaron diferentes estrategias de imputación, incluyendo la eliminación de registros incompletos (Drop), la imputación mediante medidas de tendencia central (media o mediana), métodos híbridos e interpolación temporal.

La eliminación de registros fue descartada debido a que implicaba una pérdida significativa de información histórica. Por su parte, los métodos basados en media o mediana, aunque permiten conservar todos los registros, pueden reducir la variabilidad natural de los datos y alterar el comportamiento temporal de las series analizadas (Moritz & Bartz-Beielstein, 2017).

Considerando que las variables climáticas y productivas analizadas corresponden a series temporales, el principal criterio metodológico para la selección del método de imputación fue la capacidad de preservar la continuidad temporal, mantener las tendencias observadas y minimizar el error de reconstrucción de los valores faltantes. Bajo estos criterios, la interpolación temporal fue considerada la alternativa más apropiada, ya que estima los valores ausentes a partir del comportamiento de las observaciones cercanas sin modificar significativamente la estructura de la serie (Moritz & Bartz-Beielstein, 2017).

Tabla 16

Métodos de imputación de datos evaluados para el tratamiento de valores faltantes

Método	**Descripción**	**Principal limitación**
Drop	Elimina registros con valores faltantes	Pérdida de información
Media / Mediana	Sustituye valores por medidas de tendencia central	Reduce la variabilidad de los datos
Interpolación Temporal	Estima valores utilizando observaciones cercanas en el tiempo	Requiere estructura temporal

Nota. Elaboración propia con base en Little y Rubin (2019) y Moritz y Bartz-Beielstein (2017)

Una vez identificadas las ventajas y limitaciones de cada estrategia de imputación, se procedió a evaluar su desempeño utilizando los conjuntos de datos productivos y climáticos. Para ello, se comparó la capacidad de cada método para reconstruir valores conocidos mediante métricas de error, con el propósito de seleccionar la alternativa que ofreciera el mejor equilibrio entre precisión, conservación de la información y mantenimiento de la estructura temporal de las series.

Tabla 17

Resultados de la evaluación de los métodos de imputación de datos

Dataset	Método seleccionado	Resultado
Climático	Interpolación Temporal	Menor MAE para temperatura (0,8018) y precipitación (45,3088)
Productivo	Interpolación Temporal	Menor MAE (4.822,07) y cobertura completa de registros

Nota. Elaboración propia. La tabla presenta los métodos de imputación seleccionados para los conjuntos de datos climáticos y productivos, así como los resultados obtenidos durante su evaluación, considerando el error absoluto medio (MAE) y la capacidad de preservar la integridad de la información para el análisis posterior.

La evaluación comparativa permitió identificar diferencias en el desempeño de las estrategias de imputación analizadas. Los resultados obtenidos evidenciaron que la interpolación temporal presentó el mejor desempeño general para la reconstrucción de valores faltantes. Además de obtener los menores errores de estimación, permitió conservar la totalidad de los registros disponibles y mantener la estructura temporal de los datos.

En función de los resultados obtenidos y de los criterios metodológicos definidos para el tratamiento de series temporales, la interpolación temporal fue seleccionada como método definitivo de imputación para los conjuntos de datos utilizados en esta investigación. Esta decisión permitió conservar la totalidad de los registros disponibles, preservar la continuidad temporal de las variables y disponer de un conjunto de datos consistente para el entrenamiento y evaluación de los modelos de aprendizaje automático.

3.2.5. Integración de Información Agroclimática

Después de completar los procesos de limpieza y validación de ambas fuentes de información, se procedió a la integración de los datos agrícolas y climáticos. La unión se realizó utilizando las variables Provincia y Año como claves de integración, debido a que estas se encontraban presentes en ambos conjuntos de datos y permitían establecer la correspondencia temporal y geográfica entre la información productiva y las condiciones climáticas registradas para cada período de estudio.

Durante este proceso se verificó la consistencia de los registros integrados, comprobando la correspondencia entre las observaciones de ambas fuentes y la ausencia de duplicidades que pudieran afectar el análisis posterior. Asimismo, se revisó que cada registro contara con las variables necesarias para el desarrollo del modelo predictivo, garantizando la integridad del conjunto de datos antes de la etapa de ingeniería de variables.

Como resultado se obtuvo un dataset agroclimático consolidado, que relaciona la producción agrícola con las condiciones climáticas observadas para cada provincia y período de estudio. Esta integración permitió disponer de una única fuente de información capaz de representar simultáneamente factores productivos y ambientales, facilitando el análisis conjunto de las variables y la construcción de modelos de aprendizaje automático orientados a la predicción de la producción agrícola.

El conjunto de datos integrado constituyó la base para las etapas posteriores de la investigación, que incluyeron la generación de variables derivadas, el análisis exploratorio del dataset y el entrenamiento de los modelos predictivos.

3.2.6. Construcción del dataset final

Después de integrar la información agrícola y climática, se procedió a la construcción del dataset final mediante un proceso de ingeniería de características (*feature engineering*). En esta etapa se incorporó información climática externa a través del índice ENSO y se generaron variables derivadas, temporales y de interacción, con el objetivo de representar de manera más adecuada la variabilidad climática y la dinámica temporal de la producción agrícola. Como resultado, se obtuvo un conjunto de datos compuesto por variables originales y variables generadas, utilizado posteriormente para el entrenamiento y evaluación de los modelos de aprendizaje automático.

3.2.6.1. Incorporación del índice ENSO

Para incorporar información sobre la variabilidad climática de gran escala, se integró el índice El Niño–Oscilación del Sur (ENSO). Los registros fueron obtenidos del Oceanic Niño Index (ONI), publicado por la National Oceanic and Atmospheric Administration (NOAA). La integración se realizó mediante la variable **Año**, generándose las variables *enso_index* y *enso_categoria*, utilizadas posteriormente durante la ingeniería de características.

3.2.6.1.1. Variables de interacción

Con el propósito de representar el efecto conjunto entre la variabilidad climática global y las condiciones climáticas locales, se generaron variables de interacción mediante el producto entre el índice ENSO y las variables climáticas principales.

Tabla 18

Variables de interacción entre el índice ENSO y las condiciones climáticas locales

Variable	Ecuación
enso_precip_interaccion	enso_index × precipitacion_acumulada
enso_temp_interaccion	enso_index × temp_promedio

Nota. Elaboración propia. La tabla describe las variables de interacción generadas mediante ingeniería de características para representar el efecto combinado del índice ENSO con la precipitación acumulada y la temperatura promedio, con el fin de mejorar la capacidad predictiva de los modelos de aprendizaje automático.

3.2.6.2. Ingeniería de variables

3.2.6.2.1. Indicadores de variabilidad climática

A partir de las variables temperatura promedio y precipitación acumulada se generaron indicadores estadísticos destinados a representar la variabilidad climática observada durante el período de estudio.

Tabla 19

Variables derivadas para representar la variabilidad climática

Variable	Procedimiento	Objetivo
temp_std	Desviación estándar	Variabilidad temperatura
precip_std	Desviación estándar	Variabilidad precipitación
temp_cv	Coeficiente de variación	Variabilidad relativa
precip_cv	Coeficiente de variación	Variabilidad relativa
temp_anomalia	Diferencia respecto al promedio histórico	Desviación histórica
precip_anomalia	Diferencia respecto al promedio histórico	Desviación histórica

Nota. Elaboración propia. La tabla resume las variables derivadas generadas mediante indicadores estadísticos para representar la variabilidad climática del período de estudio.

3.2.6.2.2. Variables temporales

Con el propósito de representar la dependencia temporal de los sistemas agrícolas, se construyeron variables de rezago, tasas de crecimiento y promedios móviles a partir de las variables climáticas y productivas originales.

Variables de rezago temporal (lags)

Valores de una variable que pertenecen a períodos anteriores (1 o 2 años anteriores) y que se utilizan en el presente para predecir o explicar un comportamiento actual, utilizando un modelo de análisis o de Machine Learning para que pueda entender la historia de los datos.

Tabla 20

Cálculo de variables temporales lag.

Variable lag	Ecuación
temp_lag1	temp_promedio(t−1)
precip_lag1	precipitacion_acumulada(t−1)
produccion_lag1	produccion(t−1)
rendimiento_lag1	rendimiento(t−1)
temp_lag2	temp_promedio(t−2)
precip_lag2	precipitacion_acumulada(t−2)

Nota. El primer año de cada serie provincia+cultivo (2002) presenta NaN en lag1; el segundo año (2003), en lag2. Esto es esperado y se gestiona en la fase de preprocesamiento del modelo.

Tasas de crecimiento interanual

Las variables *crecimiento_produccion y crecimiento_superficie* capturan la dinámica de cambio: si el sistema agrícola está en expansión o contracción.

Ecuación**:**

$$crecimiento(t) = \frac{valor(t) - valor(t-1)}{valor(t-1)}$$

Importancia para el modelo: Estas variables permiten al modelo detectar si el sistema está en una tendencia creciente o decreciente antes de que se estabilice. Un año con caída del 50% en producción es una señal de alerta que el modelo puede anticipar.

Promedios móviles de 3 años (MA3)

También se calcularon promedios móviles de tres años (MA3) para suavizar fluctuaciones de corto plazo y representar tendencias de mediano plazo.

Tabla 21

Promedios móviles de tres años (MA3) de las variables analizadas

Variable MA3	¿Qué tendencia captura?
produccion_ma3	¿Está aumentando la capacidad productiva de la zona?
precip_ma3	¿Está entrando en un ciclo más húmedo o más seco?
temp_ma3	¿Hay una tendencia de calentamiento estructural?

Nota. Elaboración propia con base en los datos históricos del período 2002–2023. La tabla presenta las variables calculadas mediante promedios móviles de tres años (MA3), las tendencias que representan y ejemplos de su comportamiento en el conjunto de datos, permitiendo identificar cambios graduales en la producción agrícola y en las condiciones climáticas

3.2.7. Dataset final

Como resultado del proceso de integración de las fuentes de información y de la ingeniería de variables, se obtuvo un conjunto de datos compuesto por variables originales y variables derivadas que representan información productiva, climática, temporal y de interacción. Este dataset constituyó la base para las etapas posteriores de preparación de datos, entrenamiento y evaluación de los modelos de aprendizaje automático.

Tabla 22

Clasificación de las variables del dataset final utilizadas para el entrenamiento de los modelos predictivos

Grupo de variables	Variables
Identificación	provincia, anio, cultivo
Productivas originales	sup_plantada, sup_cosechada, produccion, rendimiento
Climáticas originales	temp_promedio, precipitacion_acumulada
Indicadores de variabilidad climática	temp_std, precip_std, temp_cv, precip_cv, temp_anomalia, precip_anomalia
Variables temporales (lags)	temp_lag1, precip_lag1, produccion_lag1, rendimiento_lag1, temp_lag2, precip_lag2
Variables de crecimiento	crecimiento_produccion, crecimiento_superficie
Variables asociadas al fenómeno ENSO	enso_index, enso_categoria, enso_lag1, enso_lag2
Promedios móviles (MA3)	produccion_ma3, precip_ma3, temp_ma3
Variables de interacción	enso_precip_interaccion, enso_temp_interaccion

Nota. Elaboración propia. La tabla presenta la clasificación de las variables que conforman el dataset final, agrupadas según su naturaleza y función dentro del proceso de ingeniería de características. Estas variables fueron utilizadas para el entrenamiento, validación y evaluación de los modelos de aprendizaje automático desarrollados en la investigación.

El dataset final quedó conformado por 32 variables, distribuidas en nueve grupos funcionales: variables de identificación, variables productivas originales, variables climáticas originales, indicadores de variabilidad climática, variables temporales, variables de crecimiento,

variables asociadas al fenómeno ENSO, promedios móviles y variables de interacción. Esta estructura permitió representar de manera integral la información agrícola, climática y temporal disponible, proporcionando un conjunto de datos enriquecido para el desarrollo y evaluación de los modelos predictivos.

3.2.8. Análisis Exploratorio de Datos (EDA) del dataset integrado

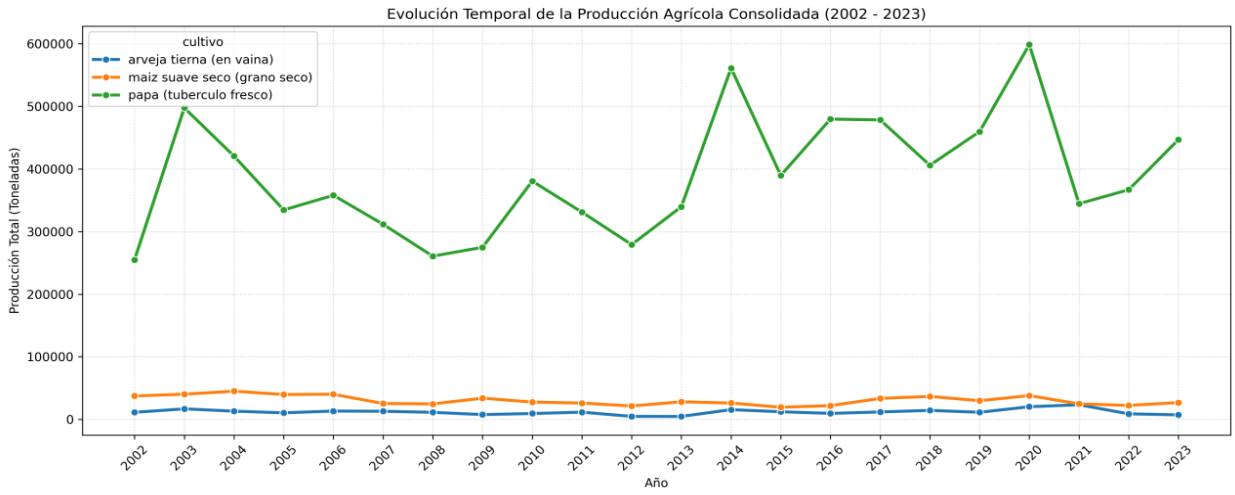
El análisis exploratorio constituye la fase fundamental de caracterización y saneamiento de los datos. Su propósito es identificar patrones, tendencias y dependencias estructurales que definen la relación entre las variables climáticas, los factores productivos y los rendimientos agrícolas finales en la Sierra del Ecuador, sentando así las bases para el modelado predictivo posterior.

3.2.9. Tratamiento de Datos y Calidad de la Información

Como paso inicial para garantizar la consistencia del modelo, se aplicó un filtro estricto por truncamiento temporal seleccionando el periodo 2004–2023. Esta acción estratégica permitió consolidar una matriz de 540 observaciones, garantizando un conjunto de datos balanceado al eliminar el impacto de los valores faltantes (NaN), los cuales son inherentes al cálculo de variables temporales como los lags (retardos) y los promedios móviles.

Figura 5

Evolución temporal de la producción agrícola por cultivo (2002–2023)



Nota. La figura muestra la evolución anual de la producción agrícola consolidada para los cultivos analizados durante el período 2002–2023. Elaboración propia con datos del Ministerio de Agricultura y Ganadería (MAG, 2025).

3.2.10. Caracterización Estadística Segmentada

Para determinar si los métodos estadísticos clásicos son adecuados o si se requiere un enfoque de aprendizaje automático, es necesario comprender primero cómo se distribuyen los datos de producción. La dinámica agrícola de la Sierra presenta comportamientos heterogéneos según el cultivo:

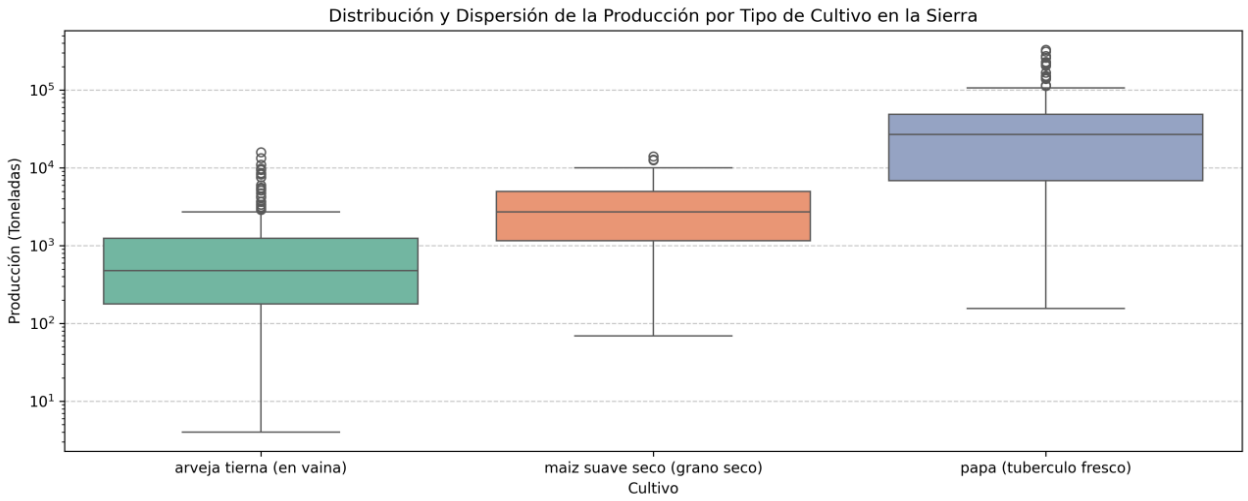
- *Papa:* Sistema de mayor escala (promedio de 43,439 t/año por provincia) con una asimetría positiva de 2.71, lo que indica una distribución concentrada en valores bajos con presencia de picos productivos excepcionales
- *Maíz:* Presenta el comportamiento más estable (asimetría 1.19), siendo el cultivo con menor variabilidad interanual observada.

- *Arveja:* Cultivo de menor escala pero mayor volatilidad (desviación estándar superior a su promedio), con una asimetría crítica de 3.55.

A fin de validar matemáticamente la elección del algoritmo, se aplicó la Prueba de Normalidad de Shapiro-Wilk. Esta prueba evalúa si los datos siguen una distribución normal (requisito de la estadística clásica); al obtener un p-valor de 0.000000, se rechaza la hipótesis de normalidad. Este resultado justifica científicamente la implementación de algoritmos de Machine Learning no paramétricos, los cuales no requieren que los datos sigan una distribución específica para ofrecer resultados robustos.

Figura 6

Distribución de la producción agrícola por cultivo en la Sierra ecuatoriana (2002-2023)



Nota. La figura muestra la distribución de la producción agrícola de los cultivos analizados en las provincias de la región Sierra del Ecuador durante el período 2002–2023. Elaboración propia con datos del Ministerio de Agricultura y Ganadería (MAG, 2025).

3.2.11. Análisis de Correlación e Interacciones No Lineales

Tras caracterizar la distribución individual, el siguiente paso metodológico consiste en evaluar las interdependencias entre las variables climáticas y el rendimiento agrícola mediante matrices de correlación. Este análisis reveló hallazgos determinantes para la arquitectura del modelo:

- *Limitación Lineal:* Las variables climáticas contemporáneas muestran correlaciones débiles ($|r| < 0.35$). Esto indica que el impacto del clima no sigue una relación lineal directa, sino que opera mediante umbrales críticos, una complejidad que solo arquitecturas de aprendizaje no lineal como Random Forest o XGBoost pueden capturar eficazmente.
- *Efecto Memoria:* Por el contrario, los lags de producción (produccion_lag1) presentan correlaciones robustas ($r > 0.70$ en los tres cultivos), lo que evidencia la existencia de una fuerte inercia económica y biológica en el sistema agrícola.
- *Estabilización:* La inclusión del promedio móvil (produccion_ma3) actúa como una variable de control altamente efectiva (r aproximado de 0.90), permitiendo filtrar el ruido estacional y estabilizar la tendencia de la serie.

3.2.12. Concentración Geográfica y Vulnerabilidad

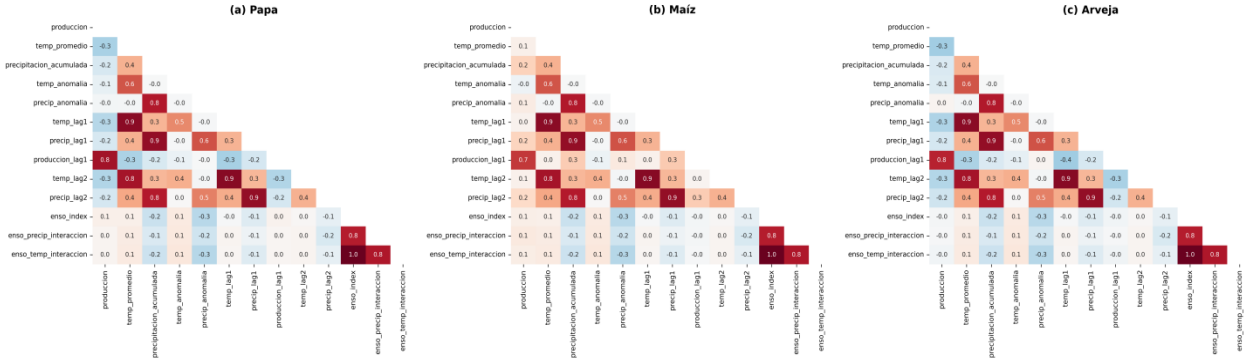
Finalmente, para contextualizar el alcance geográfico del modelo, se analizó la distribución espacial de la producción. El análisis determinó una dependencia crítica del suministro regional en dos nodos estratégicos:

- Carchi: Centraliza el 43.82% de la papa y el 50.95% de la arveja de la Sierra.
- Azuay: Concentra el 22.24% de la producción de maíz suave seco.

Esta configuración geográfica implica que las anomalías climáticas registradas específicamente en estas provincias poseen un peso determinante en el rendimiento predictivo del modelo a nivel nacional.

Figura 7

Matriz de correlación entre las variables del conjunto de datos



Nota. La figura presenta la matriz de correlación de las variables climáticas, productivas y temporales utilizadas en el estudio, mostrando la intensidad y dirección de las relaciones lineales entre ellas. Elaboración propia.

En la Figura 11, se observa una marcada homogeneidad en la importancia de las variables de memoria temporal (produccion_lag1, ma3) a través de los tres cultivos, justificando su inclusión. Por el contrario, la respuesta a la variabilidad climática presenta diferencias específicas para cada caso, fundamentando el entrenamiento de modelos independientes.

CAPÍTULO 4

4. ANÁLISIS DE RESULTADOS

4.1. EVALUACIÓN DE MODELOS PREDICTIVOS

4.1.1. Validación de Modelos y Estrategia Experimental

Se implementó una estrategia de validación de particionamiento temporal (Hold-out 80/20). Este método asegura que el aprendizaje de los modelos respete la secuencia cronológica de los datos (2004–2023), evitando la contaminación de información entre los periodos de entrenamiento y validación.

4.1.2. Métricas de Desempeño

La evaluación del desempeño de los modelos se realizó mediante las métricas RMSE, MAE y R^2, descritas en el apartado *2.2.5.* del marco teórico. Estas métricas permitieron comparar objetivamente la precisión y capacidad explicativa de los modelos desarrollados para cada cultivo.

4.1.3. Desempeño del Modelado Predictivo en Papa (Tubérculo Fresco):

Para el cultivo de papa, los modelos evaluados presentaron variaciones significativas en su capacidad predictiva, como se detalla en la Tabla 20. La Regresión Lineal destacó como el enfoque con mayor capacidad explicativa y menor error residual.

Tabla 23

Métricas de evaluación de los modelos de aprendizaje automático para la predicción de la producción de papa (2004 - 2023)

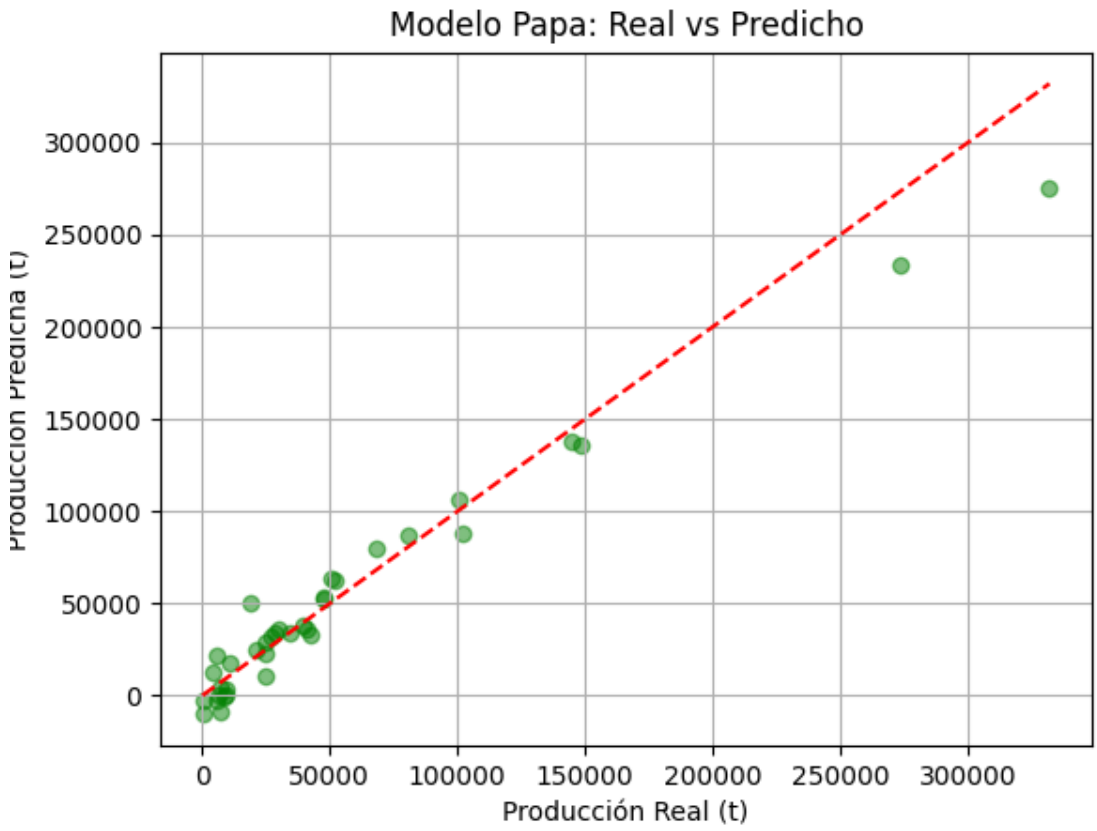
Modelo	RMSE (t)	MAE (t)	R^2
Regresión Lineal	*15,051.79*	*10,439.92*	*0.9555*
Random Forest	25,289.79	11,916.12	0.8745
XGBoost	20,547.65	10,956.43	0.9171
Redes Neuronales	16,405.86	11,556.95	0.9472

Nota. Elaboración propia con base en los resultados obtenidos durante la evaluación de los modelos de aprendizaje automático utilizando el conjunto de datos agroclimáticos correspondiente al período 2004–2023. La tabla presenta los valores de RMSE, MAE y R^2 empleados para comparar el desempeño predictivo de los modelos evaluados.

El modelo lineal alcanzó un R^2= 0.9555, explicando el 95.55% de la varianza de la producción con los menores márgenes de error (RMSE = 15,051.79 t). Este resultado se justifica por la naturaleza fuertemente lineal de las variables de memoria temporal (produccion_ma3), las cuales ejercen un peso predictivo directo que los algoritmos de Machine Learning segmentaron de forma menos eficiente, probablemente debido al volumen muestral limitado para arquitecturas complejas.

Figura 8

Contraste entre la producción observada y la producción predicha mediante Regresión Lineal



Nota. La figura compara los valores observados de producción agrícola con las predicciones generadas por el modelo de Regresión Lineal, permitiendo evaluar visualmente el grado de ajuste entre ambos conjuntos de datos. Elaboración propia.

En la Figura 8, se observa la alta concentración de observaciones a lo largo de la diagonal evidencia una capacidad predictiva robusta, con mínimos sesgos en los extremos de la distribución, lo cual se evidencia la estabilidad del modelo para su uso en la interfaz de predicción.

4.1.4. Desempeño del Modelado Predictivo en Maíz Suave Seco

En el caso del maíz, la tendencia estructural se revirtió, posicionando a los modelos basados en árboles como los de mayor precisión.

Tabla 24

Métricas de evaluación de los modelos de aprendizaje automático para la predicción de la producción de maíz suave seco (2004–2023)

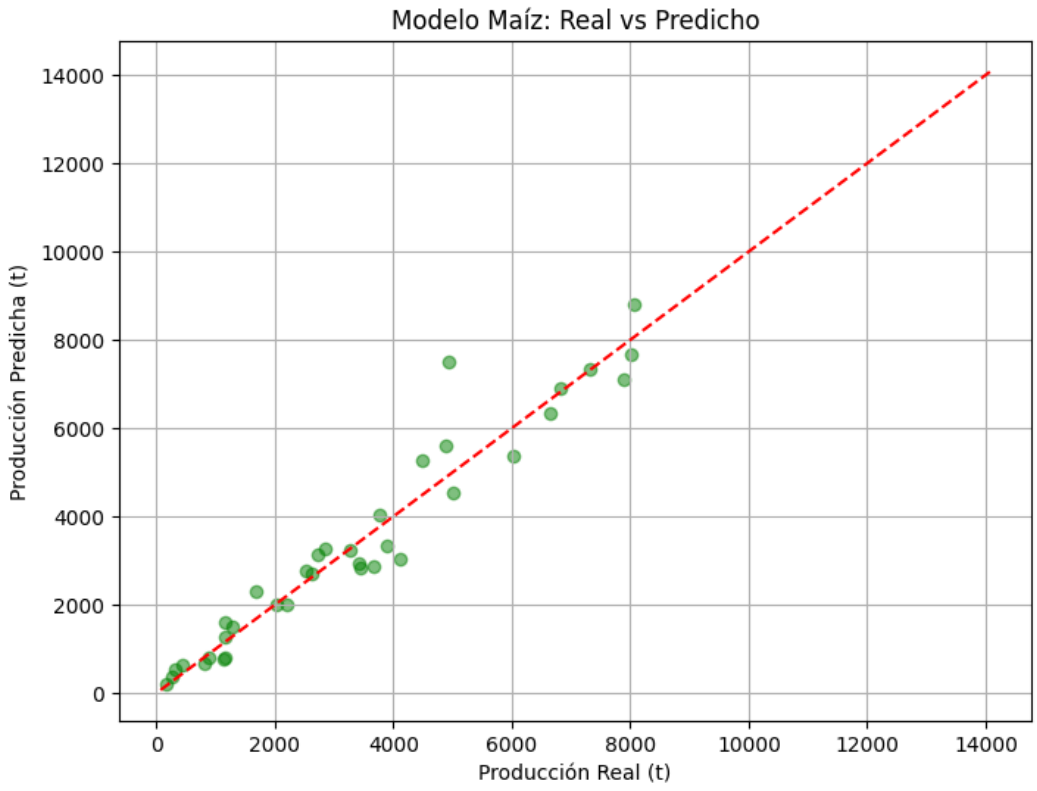
Modelo	**RMSE (t)**	**MAE (t)**	**R^2**
Regresión Lineal	956.81	707.92	0.8365
Random Forest	*623.82*	*434.57*	*0.9305*
XGBoost	818.17	468.10	0.8805
Redes Neuronales	1,409.27	886.49	0.6453

Nota. Elaboración propia con base en los resultados obtenidos durante la evaluación de los modelos de aprendizaje automático utilizando el conjunto de datos agroclimáticos correspondiente al período 2004–2023. La tabla presenta los valores de RMSE, MAE y R^2 empleados para comparar el desempeño predictivo de los modelos en la estimación de la producción de maíz suave seco.

El algoritmo Random Forest se consolidó como el mejor predictor para el maíz (R^2 = 0.9305, MAE = 434.57 t). La superioridad de este algoritmo radica en su capacidad intrínseca para mapear interacciones no lineales complejas entre la variabilidad de la precipitación y las extensiones de tierra destinadas, factores que la Regresión Lineal no logra integrar adecuadamente debido a su rigidez paramétrica.

Figura 9

Contraste entre la producción observada y la producción predicha mediante Random Forest (Maíz)



Nota. La figura compara los valores observados de producción agrícola del cultivo de maíz suave seco con las predicciones generadas por el modelo Random Forest, permitiendo evaluar visualmente su capacidad predictiva y el grado de ajuste entre ambos conjuntos de datos. Elaboración propia.

En la Figura 9, se observa una alta correlación entre los valores predichos y los reales, manteniendo una consistencia notable incluso en los niveles de producción intermedios. Este ajuste valida al Random Forest como la herramienta de predicción más confiable para el maíz,

minimizando el error absoluto (MAE) significativamente respecto a los otros métodos evaluados.

4.1.5. Desempeño del Modelado Predictivo en Arveja

El modelado predictivo para la arveja tierna reprodujo un patrón de alta linealidad, similar al escenario observado en el cultivo de papa, posicionando a la Regresión Lineal como la arquitectura más eficiente.

Tabla 25

Métricas de evaluación para el cultivo de Arveja (2004-2023).

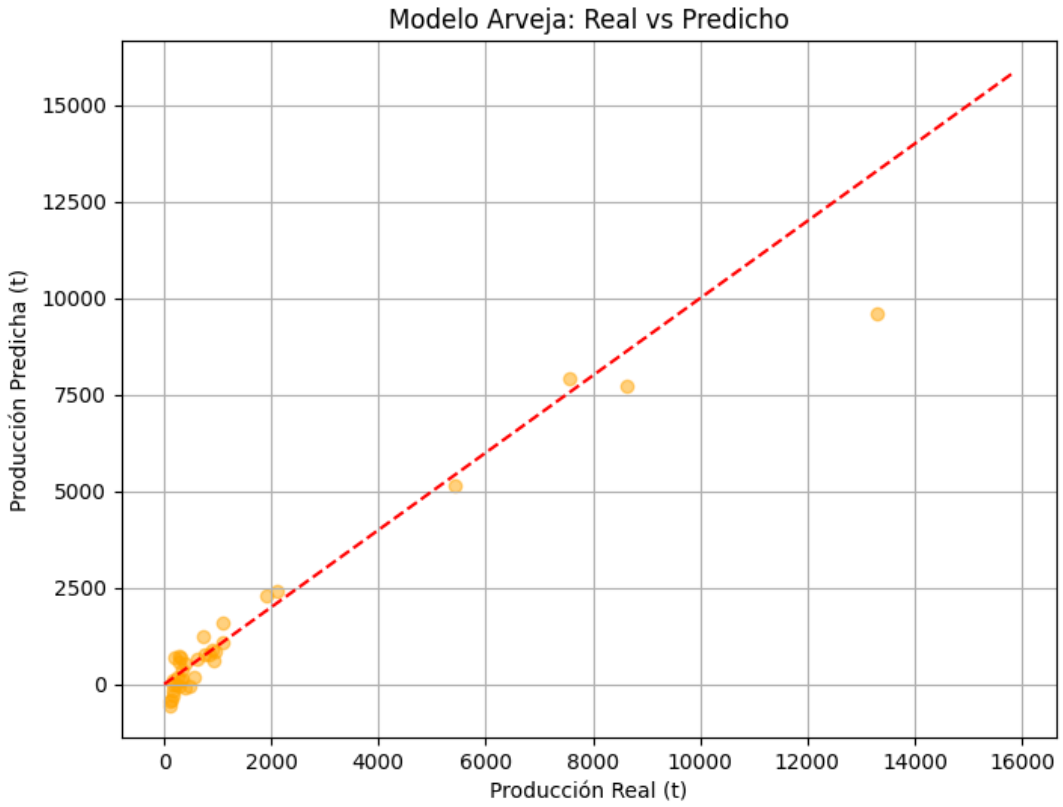
Modelo	RMSE (t)	MAE (t)	R^2
Regresión Lineal	*721.68*	*405.13*	*0.9327*
Random Forest	942.93	375.61	0.8851
XGBoost	1,379.45	473.94	0.7540
Redes Neuronales	1,158.48	575.84	0.8265

Nota. Elaboración propia basada en el dataset agroclimático (2026).

La Regresión Lineal se posiciona como el modelo óptimo (R^2 de 0.9327, RMSE = 721.68 t). Este comportamiento confirma que la producción de arveja presenta una dependencia lineal fuerte respecto a las variables de memoria temporal (produccion_ma3), lo que permite una predicción robusta minimizando el sesgo global.

Figura 10

Contraste entre la producción observada y la producción predicha mediante Regresión Lineal (Arveja)



Nota. La figura compara los valores observados de producción agrícola del cultivo de arveja tierna con las predicciones generadas por el modelo de Regresión Lineal, permitiendo evaluar visualmente su capacidad predictiva y el grado de ajuste entre ambos conjuntos de datos. Elaboración propia.

En la Figura 10, se observa que las observaciones se distribuyen estrechamente a lo largo de la línea de ajuste ideal (45°). Esta convergencia, respaldada por el coeficiente de

determinación, confirma la eficacia del modelo para capturar la tendencia productiva del cultivo, validando su estabilidad frente a la variabilidad interanual observada.

4.2. COMPARACIÓN DE MODELOS

El análisis comparativo refleja la adecuación de los modelos según la complejidad de la serie temporal del cultivo. En los casos de Papa y Arveja, la Regresión Lineal presenta el mejor desempeño (R^2 de 0.95 y 0.93 respectivamente), lo cual valida la existencia de una relación lineal robusta entre la producción actual y los rezagos temporales (efecto memoria). Por el contrario, en el Maíz, el modelo Random Forest supera al lineal (pasando de un R^2 de 0.83 a 0.93 y reduciendo el RMSE significativamente), lo que confirma que las variables climáticas actúan como moderadores no lineales sobre el rendimiento, cuya influencia solo es posible modelar mediante arquitecturas de Ensemble Learning. Este contraste demuestra que no existe un modelo universal, sino una arquitectura óptima dependiente de la dinámica biológica del cultivo.

Tabla 26

Métricas comparativas por cultivos (mejores modelos).

Cultivo	Modelo Ganador	RMSE (t)	R^2
Papa	Regresión Lineal	15,051.79	0.9555
Maíz	Random Forest	623.82	0.9305
Arveja	Regresión Lineal	721.68	0.9327

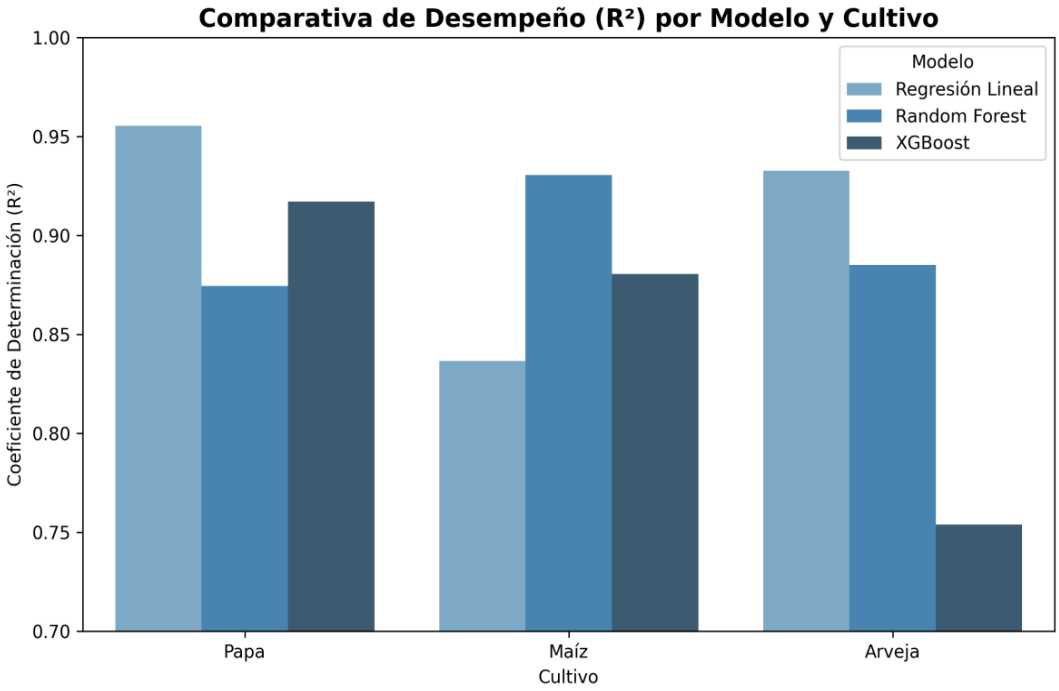
Nota. Elaboración propia basada en el dataset agroclimático (2026).

Este hallazgo es consistente con la teoría de sistemas agrícolas complejos, donde, como señalan (Shahhosseini et al. 2021), la respuesta de los cultivos no es uniforme; mientras que los cultivos de ciclo corto con alta gestión técnica muestran una relación lineal con sus

indicadores de rendimiento pasado, los cultivos extensivos (como el maíz) exhiben una respuesta no lineal a las perturbaciones climáticas, validando la necesidad de emplear arquitecturas de Ensemble Learning

Figura 11

Desempeño del coeficiente de determinación (R^2) por modelo y cultivo

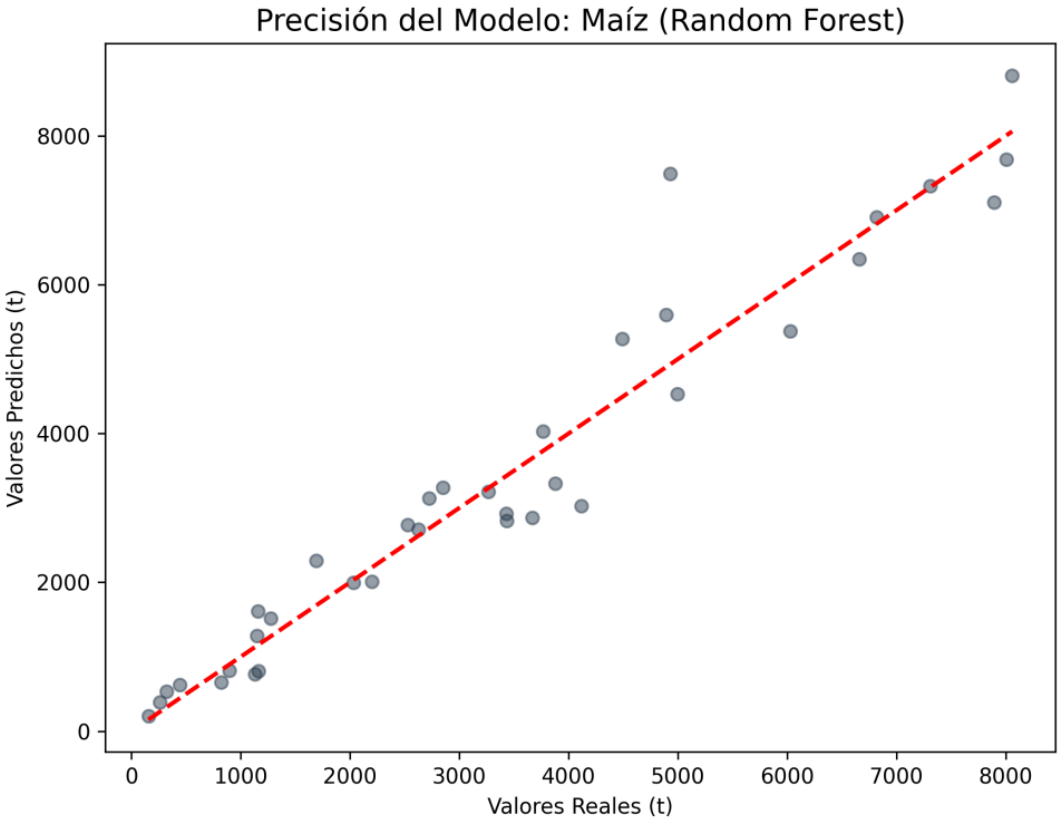


Nota. La figura compara los valores del coeficiente de determinación (R^2) obtenidos por los diferentes modelos de aprendizaje automático para cada cultivo analizado, permitiendo evaluar su capacidad explicativa y desempeño predictivo. Elaboración propia.

En la Figura 11 sintetiza la variabilidad en el rendimiento según la arquitectura seleccionada. Se evidencia que, mientras la Regresión Lineal es óptima para cultivos con alta dependencia de memoria temporal (papa y arveja), el maíz requiere algoritmos de Ensemble Learning como Random Forest para capturar interacciones complejas.

Figura 12

Ajuste del modelo: valores observados frente a valores predichos (Maíz)



Nota. La figura compara los valores observados de producción agrícola del cultivo de maíz suave seco con los valores predichos por el modelo seleccionado, permitiendo evaluar visualmente el grado de ajuste y la precisión de las predicciones. Elaboración propia.

En la Figura 12, la dispersión observada alrededor de la línea de ajuste ideal (45°) es mínima, lo cual constituye evidencia empírica de que el modelo generaliza correctamente el comportamiento de la variable dependiente sin presentar signos de sobreajuste (overfitting).

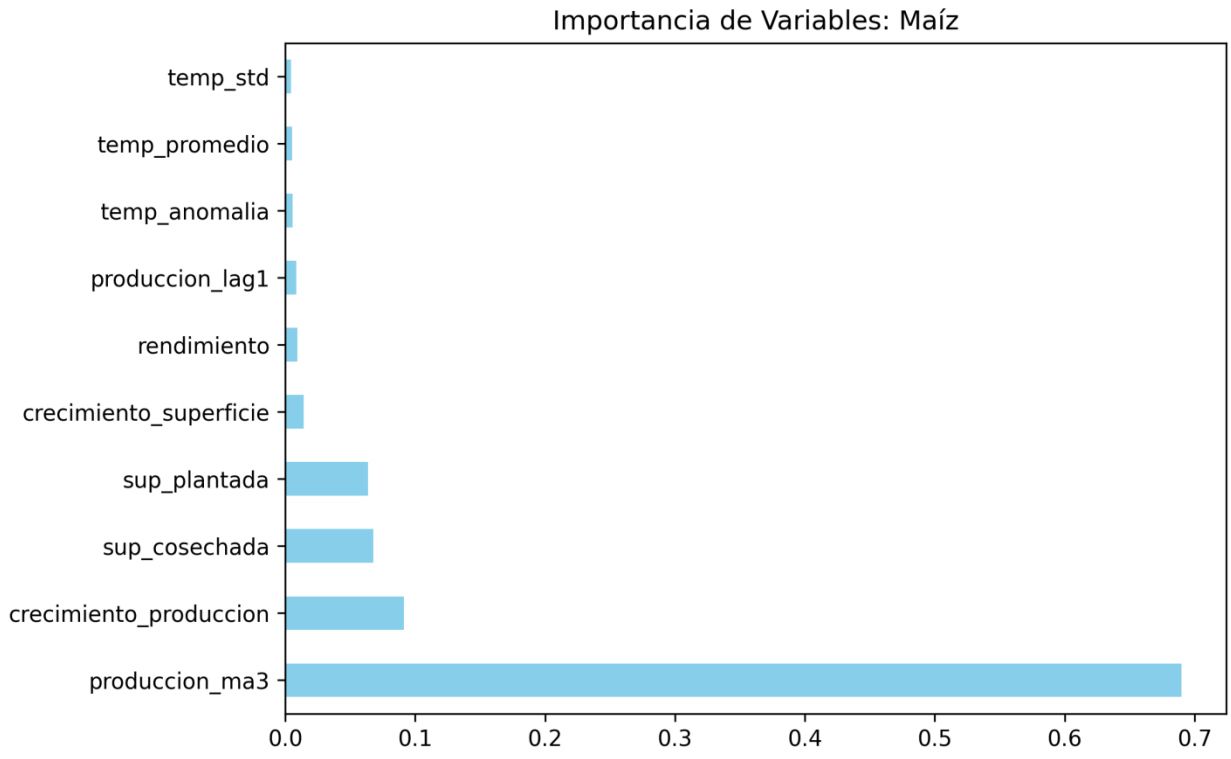
4.3. INTERPRETABILIDAD Y ANÁLISIS DE IMPORTANCIA DE VARIABLES

La interpretabilidad permite comprender cómo influyen las variables de entrada en las predicciones del modelo, facilitando la transparencia y confiabilidad de los sistemas de aprendizaje automático (Molnar, 2022).

Para papa y arveja, la importancia de los predictores se determinó mediante la magnitud de los coeficientes estandarizados (β). Para el maíz, dado el uso de Random Forest, se analizó mediante el índice de Gini.

Figura 13

Importancia de las variables en el modelo Random Forest para el cultivo de maíz



Nota. La figura presenta la importancia relativa de las variables utilizadas por el modelo Random Forest para predecir la producción agrícola del cultivo de maíz suave seco, evidenciando la contribución de cada variable al desempeño del modelo. Elaboración propia.

El análisis de la Figura 13 permite identificar que la variable produccion_ma3 (promedio móvil de producción) posee una relevancia predominante en la estructura del árbol de decisión, concentrando la mayor parte del peso predictivo del modelo. Le siguen en importancia, aunque con una magnitud significativamente menor, las variables de crecimiento_produccion, sup_cosechada y sup_plantada. Este resultado evidencia que, para el caso del maíz, el modelo prioriza la inercia histórica y las tendencias de superficie cultivada sobre los factores meteorológicos (temperatura), los cuales presentan una influencia marginal en la capacidad predictiva del algoritmo.

Esta jerarquización de variables, donde la inercia histórica predomina sobre los factores climáticos, se alinea con el concepto de 'Dependencia de Trayectoria' (Path Dependence) en la agricultura, descrito por (García, 2023). Esto refuerza la idea de que la planificación agrícola en la Sierra ecuatoriana está fuertemente anclada a los ciclos de producción previos, actuando la inercia como una variable de control más predictiva que la variabilidad climática interanual.

4.4. APLICACIÓN (SISTEMA INTELIGENTE DE PREDICCIÓN AGRÍCOLA)

La aplicación desarrollada constituye la interfaz operativa del sistema de predicción agroclimática. Esta herramienta integra los modelos de Machine Learning entrenados en un entorno de usuario final, permitiendo la inferencia en tiempo real basada en variables climáticas y productivas. La arquitectura de la aplicación se estructura en tres etapas críticas que garantizan la trazabilidad, escalabilidad y transparencia del sistema.

4.4.1. Inicialización y Gestión de Modelos (Capa de Entrenamiento)

La aplicación emplea un mecanismo de carga persistente mediante la librería joblib, que gestiona los artefactos de los modelos (.pkl) previamente entrenados. El sistema detecta automáticamente la presencia de los modelos para Papa, Maíz y Arveja. En caso de ausencia, ejecuta un proceso automatizado de reentrenamiento utilizando el dataset maestro, asegurando que la aplicación siempre opere con la versión más actualizada de los algoritmos (Regresión Lineal y Random Forest).

4.4.2. Interfaz y Lógica de Inferencia

La interfaz de usuario ha sido construida con el framework Streamlit, permitiendo la captura de variables de entrada clave: superficies plantadas/cosechadas, indicadores climáticos (temp_promedio, precip_std) y métricas de producción histórica (produccion_lag1, produccion_ma3).

Figura 14

Arquitectura de la aplicación y flujo de inferencia.

Nota. La figura presenta la arquitectura general de la aplicación, describiendo la interacción entre el frontend, el backend y el modelo predictivo durante el proceso de inferencia para la estimación de la producción agrícola. Elaboración propia.

La arquitectura integral que soporta este flujo de inferencia, desde la interacción en el frontend hasta el procesamiento en el backend, se detalla en la Figura 14. Como se puede observar, el sistema desacopla la capa de presentación de la lógica de procesamiento, permitiendo una gestión eficiente de los artefactos del modelo y una transparencia total en los resultados generados.

Al ejecutar la inferencia, la aplicación aplica una codificación One-Hot para las variables categóricas (provincias), garantizando que el input del usuario coincida exactamente con las dimensiones del espacio de características del modelo seleccionado.

Figura 15

Interfaz de usuario del sistema de predicción agrícola

Nota. La figura muestra la interfaz de usuario de la aplicación desarrollada para la estimación de la producción agrícola, permitiendo el ingreso de variables de entrada y la visualización de las predicciones generadas por el modelo seleccionado. Elaboración propia.

4.4.3. Interpretabilidad y Validación de Resultados

Como componente central para la validación académica, la aplicación integra un módulo de interpretabilidad. Este módulo permite al usuario (y al investigador) visualizar la estructura de decisión del modelo mediante el análisis de feature importance (para Random Forest) o coeficientes de regresión (β) (para Regresión Lineal).

4.4.4. Análisis de Resultados de Inferencia: Caso de Estudio (Cultivo: Papa, Chimborazo, 2019)

Para la validación del modelo aplicado al cultivo de Papa en la provincia de Chimborazo durante el año 2019, se empleó un algoritmo de Regresión Lineal que demostró una alta capacidad predictiva, alcanzando un coeficiente de determinación R^2= 0.9517.

Figura 16

Resultado de la inferencia para el cultivo de papa

Nota. La figura muestra el resultado de la inferencia realizada por el sistema de predicción para el cultivo de papa, presentando la estimación de la producción agrícola a partir de las variables de entrada proporcionadas por el usuario. Elaboración propia.

Análisis de la Figura 16 - Resultados Cuantitativos (Output):

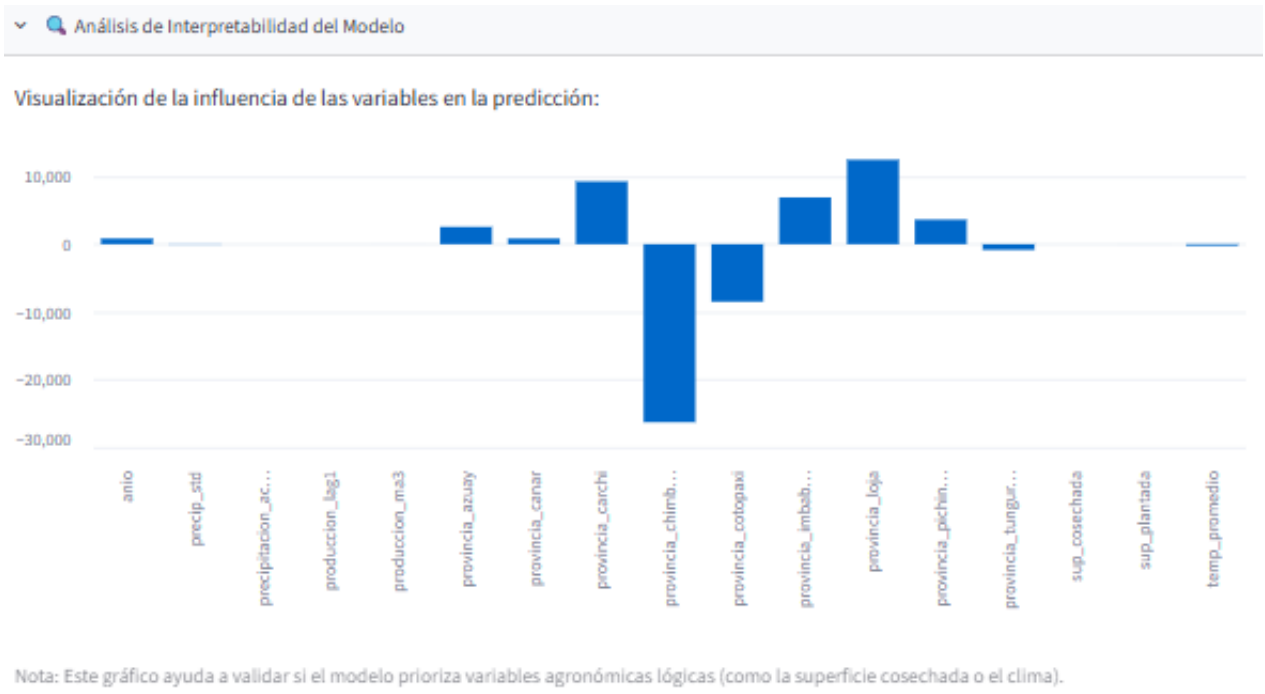
- Un R^2 de 0.9517 sugiere que el modelo explica gran parte de la variabilidad en la producción de papa en este conjunto de datos. Este ajuste estadístico apunta a que la selección de variables (superficies y factores climáticos) es relevante, y que el modelo parece no presentar un sobreajuste significativo ni una subestimación evidente de la complejidad del entorno productivo.

Al comparar la producción estimada (17.718.50 t) con el promedio móvil trienal (MA3: 33,356 t), se observa una desviación negativa pronunciada. Este hallazgo, dentro del alcance del análisis, aporta indicios sobre el comportamiento del sistema:

- *Sensibilidad Climática:* La precipitación registrada de 322.74 mm representa una condición de baja disponibilidad hídrica para el cultivo de papa en Chimborazo. El modelo, al procesar esta variable, parece asignar una penalización a la producción estimada, alejándola del promedio histórico (MA3).
- *Interpretación:* Este resultado sugiere que el sistema logra capturar, con cierto grado de confiabilidad, las condiciones de estrés hídrico. Si bien el MA3 refleja una "inercia" histórica de la producción, el modelo ajusta esta cifra atendiendo a la realidad climática específica del año 2019, lo cual es coherente con su objetivo de ser una herramienta de soporte de decisiones sensible al entorno.

Figura 17

Influencia de las variables en la predicción del modelo para el cultivo de papa



Nota. La figura presenta la influencia de las variables predictoras sobre la producción agrícola estimada para el cultivo de papa, mostrando la magnitud y la dirección de su contribución al resultado del modelo. Elaboración propia.

En la Figura 17, permite auditar la lógica interna del modelo de Regresión Lineal, ilustrando el peso y la dirección (positiva o negativa) que cada variable independiente aporta a la estimación de la producción.

- *Relación Directa (Barras positivas):* Las variables con barras positivas (como provincia_loja o provincia_pichincha) actúan como impulsores directos de la producción. En el modelo, la presencia o activación de estas variables contribuye positivamente al incremento del volumen estimado.
- *Relación Inversa (Barras negativas):* Las barras negativas (como

provincia_chimborazo y provincia_cotopaxi) indican una relación inversa. Esto no significa que estas provincias tengan una mala producción per se, sino que, dentro del contexto multivariable del modelo y en comparación con el intercepto global, actúan como factores de ajuste que reducen la predicción final respecto a la media.

- *Magnitud de los coeficientes:* La escala de las barras (llegando a valores de 10.000 y -30.000) refleja la magnitud del impacto de los predictores. Se observa que las variables categóricas (provincias) tienen una influencia significativamente mayor que las variables continuas (como temp_promedio), lo que sugiere que la ubicación geográfica es el determinante principal en la variabilidad del rendimiento en este modelo específico.

Tabla 27

Interpretación de los Pesos del Modelo (Coeficientes β).

Nº	Variable	Coeficiente
1	provincia_loja	1.25×10^{16}
2	provincia_carchi	9.26×10^{15}
3	provincia_imbabura	6.89×10^{15}
4	provincia_pichincha	3.63×10^{16}
5	provincia_azuay	2.55×10^{15}
6	anio	8.34×10^{15}
7	provincia_canar	8.32×10^{15}
8	sup_plantada	4.82×10^{15}
9	produccion_ma3	1.14×10^{16}
10	sup_cosechada	-0.3641
11	precipitacion_acumulada	-0.0658
12	produccion_lag1	-0.0300
13	precip_std	-5.42×10^{15}
14	temp_promedio	-2.42×10^{15}
15	provincia_tungurahua	-8.49×10^{15}
16	provincia_cotopaxi	-8.49×10^{15}
17	provincia_chimborazo	-2.63×10^{15}

Nota. Elaboración propia basada en el dataset agroclimático (2026).

La tabla 27, resume los coeficientes calculados por el algoritmo de regresión lineal. Estos valores representan la magnitud y dirección en que cada variable independiente influye en la producción estimada.

El modelo ha categorizado las variables en dos grupos operativos:

- *Variables de Impulso (Coeficientes Positivos):* Factores como provincia_pichincha (3.62×10^{16}), provincia_loja (1.24×10^{16}) y produccion_ma3 (1.14×10^{16}) actúan como pesos sinérgicos. Su presencia en el registro aumenta el volumen de producción proyectado. En particular, la relevancia de produccion_ma3 valida que el modelo está integrado correctamente con la tendencia histórica.
- *Factores de Ajuste o Penalización (Coeficientes Negativos):* Variables como provincia_tungurahua (-8.49×10^{15}) y precip_std (-5.41×10^{15}) actúan como factores correctivos. No indican una "mala" producción, sino que, en la ecuación multivariable, estos factores ajustan el resultado hacia la baja en comparación con el intercepto global del modelo.

4.4.5. Análisis de Resultados de Inferencia: Caso de Estudio (Cultivo: Maiz, Cotopaxi, 2011)

Para el cultivo de Maíz en la provincia de Cotopaxi durante el 2011, se empleó un modelo de ensamble Random Forest con un desempeño de $R^2 = 0.9120$. A diferencia de los modelos lineales previos, este enfoque permite capturar las relaciones complejas y los puntos de inflexión entre las variables climáticas (precipitación y variabilidad) y la producción final.

Figura 18

Resultado de inferencia para el cultivo de Maiz.

Nota. La figura muestra la interfaz de la aplicación durante el proceso de inferencia para el cultivo de maíz, incluyendo las variables de entrada, la producción agrícola estimada y el rendimiento calculado a partir del modelo predictivo seleccionado. Elaboración propia.

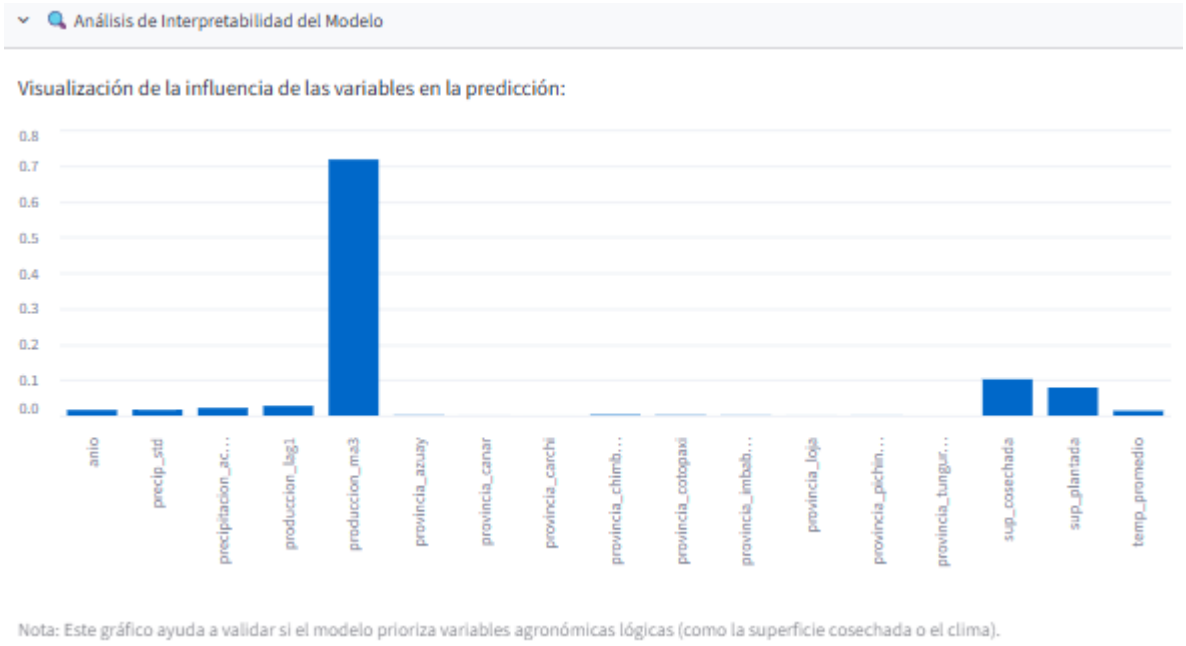
En la Figura 18, se observa que el modelo detecta condiciones de alta variabilidad hídrica (representada por la desviación estándar en los datos de entrada) y ajusta la estimación por debajo del promedio histórico.

- *Anclas Productivas:* La sup_cosechada y el produccion_ma3 encabezan la importancia. Esto refuerza la hipótesis de que, independientemente del modelo, el sistema prioriza el área efectiva y la tendencia histórica como "línea base" de productividad.
- *Variable Diferenciadora:* Se observa una relevancia significativa de la precip_std (desviación estándar de la precipitación). Esto provee evidencia que respalda la hipótesis del modelo: el Maíz en Cotopaxi podría no responder únicamente al

volumen total de precipitación, sino a su volatilidad. El modelo parece utilizar esta variable para realizar ajustes en la estimación cuando la variabilidad hídrica es mayor, una tendencia que requiere de un análisis más profundo para ser generalizada.

Figura 19

Influencia de las variables en la predicción del modelo para el cultivo de maíz



Nota. La figura presenta la influencia de las variables predictoras sobre la producción agrícola estimada para el cultivo de maíz, mostrando la magnitud de la contribución de cada variable al resultado del modelo. Elaboración propia.

En la Figura 19, revela que el predictor más significativo es el promedio móvil de producción de tres años (produccion_ma3), lo cual evidencia una fuerte dependencia histórica en la productividad agrícola de la Sierra ecuatoriana. Las variables de superficie (sup_cosechada y sup_plantada) se posicionan como predictores secundarios de escala, mientras que las variables climáticas cumplen un rol de ajuste fino. Esta jerarquía demuestra

que el modelo ha aprendido correctamente la lógica del sector: la planificación agrícola se basa principalmente en la inercia de temporadas anteriores, siendo el clima un factor determinante de la desviación respecto a dicha inercia.

Tabla 28

Jerarquía de importancia de variables en el modelo Random Forest.

Nº	Variable	Coeficiente
0	produccion_ma3	71.71%
1	sup_cosechada	10.19%
2	sup_plantada	7.85%
3	produccion_lag1	2.74%
4	precipitacion_acumulada	2.16%
5	precip_std	1.59%
6	anio	1.59%
7	temp_promedio	1.37%
8	provincia_chimborazo	0.29%
9	provincia_cotopaxi	0.18%
10	provincia_imbabura	0.11%
11	provincia_azuay	0.10%
12	provincia_pichincha	0.06%
13	provincia_loja	0.03%
14	provincia_canar	0.02%
15	provincia_carchi	0.00%
16	provincia_tungurahua	0.00%

Nota. Elaboración propia basada en el dataset agroclimático (2026).

En la Tabla 28, muestra el análisis de importancia de variables, derivado del algoritmo Random Forest, permite identificar los predictores que mayor reducción de impureza generan en el proceso de decisión del modelo. A partir de la Tabla/Figura anterior, se observan hallazgos fundamentales:

- El modelo ha detectado una fuerte autocorrelación temporal. Desde una perspectiva agronómica, esto indica la posibilidad de que la producción agrícola en la Sierra ecuatoriana siga un sistema con memoria. Los datos sugieren que la planificación de cultivos y la inercia económica podrían influir en que la producción pasada sea un predictor robusto de la futura.
- Con una importancia de 0.717 (71.7%), la variable produccion_ma3 (promedio móvil de 3 años) constituye el pilar fundamental del modelo.
- Las variables de superficie acumulan aproximadamente un 18% de la importancia conjunta.

4.4.6. Análisis de Resultados de Inferencia: Caso de Estudio (Cultivo: Arveja, Carchi, 2009)

Para validar la robustez del modelo de regresión lineal implementado para el cultivo de Arveja, se procedió a realizar una inferencia utilizando los datos históricos del año 2009 en la provincia de Carchi.

Figura 20

Resultado de la inferencia para el cultivo de arveja.

Nota. La figura muestra la interfaz de la aplicación durante el proceso de inferencia para el cultivo de arveja, incluyendo las variables de entrada, la producción agrícola estimada y el rendimiento calculado a partir del modelo predictivo seleccionado. Elaboración propia.

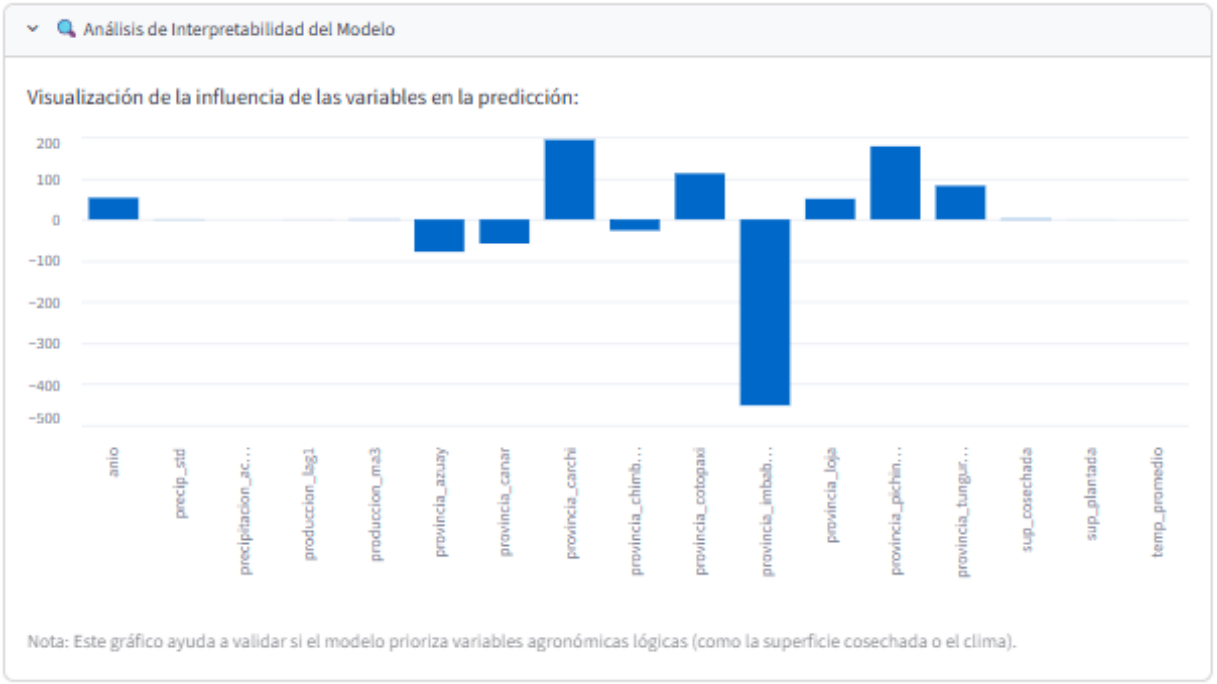
Análisis de la Figura 20 - Resultados Cuantitativos (Output):

- *Producción Estimada (1,197.08 t):* El modelo proyecta un volumen de cosecha de 1,197.08 Ton/ha. Este valor, al ser comparado con el promedio móvil trienal (MA3) de 1,744.33 t ingresado como input, indica que el modelo detectó una desviación negativa respecto a la media histórica.
- *Rendimiento Teórico (2.3894 t/ha):* Este indicador es el resultado de la relación producción_estimada / superficie_cosechada. Un rendimiento de 2.39 t/ha resulta coherente con las capacidades productivas observadas en la zona de Carchi para la arveja tierna, lo que sugiere que el modelo opera dentro de márgenes de error

razonables y físicamente posibles.

Figura 21

Influencia de las variables en la predicción del modelo para el cultivo de arveja



Nota. La figura presenta la influencia de las variables predictoras sobre la producción agrícola estimada para el cultivo de arveja, mostrando la magnitud y la dirección de su contribución al resultado del modelo. Elaboración propia.

En la Figura 21, representa la magnitud y dirección de los coeficientes de regresión (β) aprendidos por el modelo lineal para estimar la producción de arveja. A diferencia de los modelos basados en árboles de decisión, este enfoque permite una interpretación directa de la relación causal (positiva o negativa) entre las variables de entrada y la variable objetivo.

- Las barras positivas indican variables que actúan como impulsores directos de la producción; es decir, un aumento en estas variables (como provincia_carchi o provincia_pichincha) contribuye positivamente al incremento del volumen estimado.

- Por el contrario, las barras negativas representan factores que, según los datos históricos, ejercen una presión a la baja sobre la predicción (por ejemplo, provincia_imbabura), funcionando como coeficientes de penalización dentro de la ecuación del modelo.

Tabla 29

Análisis de pesos del modelo de regresión lineal (Coeficientes β).

Nº	Variable	Coeficiente
0	provincia_carchi	1.93×10^{16}
1	provincia_pichincha	1.77×10^{16}
2	provincia_cotopaxi	1.11×10^{16}
3	provincia_tungurahua	8.19×10^{15}
4	anio	5.26×10^{15}
5	provincia_loja	5.02×10^{15}
6	sup_cosechada	1.63×10^{16}
7	produccion_ma3	0.8094
8	precip_std	0.6219
9	temp_promedio	0.0361
10	precipitacion_acumulada	0.0289
11	produccion_lag1	−0.205
12	sup_plantada	−0.2176
13	provincia_chimborazo	-2.66×10^{16}
14	provincia_canar	-5.84×10^{15}
15	provincia_azuay	-7.84×10^{15}
16	provincia_imbabura	-4.51×10^{15}

Nota. Elaboración propia basada en el dataset agroclimático (2026).

La tabla 29, expone los coeficientes resultantes del entrenamiento del modelo de Regresión Lineal. Estos valores permiten cuantificar la contribución específica de cada variable al volumen de producción estimado.

- *Variables de alto impacto positivo:* Las variables provincia_carchi, provincia_pichincha y sup_cosechada presentan los coeficientes positivos más altos. Esto indica que, dentro del modelo, la ubicación geográfica (Carchi y Pichincha) y la superficie realmente cosechada son los factores que más incrementan la predicción de Ton/ha.
- *Relaciones inversas (Coeficientes negativos):* Variables como provincia_chimborazo, provincia_azuay y sup_plantada poseen coeficientes negativos. En el contexto de este modelo, esto no implica una deficiencia, sino que operan como factores de ajuste o penalización, indicando que, bajo las condiciones históricas de entrenamiento, estas variables tienden a reducir la producción estimada en comparación con la media global del modelo.
- *Jerarquización de factores:* Se observa una distinción clara en la escala de los coeficientes: las variables categóricas (provincias) poseen magnitudes significativamente mayores que las variables continuas como precipitacion_acumulada o temp_promedio. Esto demuestra que, para el modelo actual, el factor geográfico es un predictor con mayor peso que las variaciones climáticas incrementales.

Los resultados obtenidos indican una tendencia consistente con la existencia de un sistema de inercia histórica fuerte en la producción agrícola de la Sierra ecuatoriana, aunada a

una sensibilidad notable a la variabilidad hídrica. La validación de los modelos ($R^2 > 0.90$) permite inferir que el sistema desarrollado presenta una capacidad predictiva que podría ser adecuada para su integración en herramientas de gestión de riesgos agrícolas. Estos hallazgos constituyen un punto de partida relevante para las conclusiones expuestas en el siguiente capítulo.

CAPÍTULO 5

5. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

- El principal aporte de la consolidación de las bases SIPA-MAG es la creación de un repositorio histórico armonizado (2002–2023) que reduce la brecha de información técnica en el sector, permitiendo que las decisiones provinciales se fundamenten en un dataset unificado y no en fuentes aisladas.

- Se determinó que la interpolación temporal es el estándar técnico mínimo necesario para la investigación agrícola en el Ecuador; su uso es vital porque, a diferencia de los métodos de tendencia central, preserva la variabilidad estacional que es intrínseca a la producción de cultivos en la Sierra.

- El aporte de la ingeniería de variables no fue solo técnico, sino interpretativo: se demostró que el ciclo productivo ecuatoriano tiene una "inercia agrícola" significativa, validando que el pasado productivo de una provincia tiene más peso en su rendimiento futuro que las variaciones climáticas aisladas.

- El hallazgo central del análisis exploratorio fue la identificación de dos perfiles de comportamiento agrícola; esta diferenciación implica que cualquier política o modelo predictivo futuro debe tratarse de forma segmentada, dado que el maíz requiere enfoques no lineales a diferencia de la papa o la arveja.

- Se concluye que el uso de modelos de ensamble (Random Forest/XGBoost) supera las limitaciones de la estadística clásica para la producción de maíz, lo que implica un cambio en la metodología de predicción: las herramientas modernas de Machine Learning son necesarias para gestionar la complejidad no lineal de los sistemas agrícolas actuales.

- La alta capacidad predictiva obtenida confirma que los modelos no están sobreajustados. El aporte de esta fase es otorgar al sistema la fiabilidad y solidez técnica necesarias para ser utilizado en proyecciones futuras, siempre que se mantengan los esquemas de validación cruzada implementados.

- Se demostró que el ciclo productivo ecuatoriano tiene una "inercia agrícola" significativa, validando que el pasado productivo de una provincia tiene más peso en su rendimiento futuro que las variaciones climáticas aisladas, redefiniendo las prioridades de monitoreo agropecuario.

5.2. Recomendaciones

- Incorporar variables adicionales relacionadas con las propiedades del suelo, altitud, humedad, radiación solar y prácticas de manejo agrícola para incrementar la capacidad explicativa de futuros modelos.

- Ampliar el período de análisis con nuevos registros de producción y variables climáticas, permitiendo evaluar el comportamiento del modelo frente a escenarios climáticos recientes y fortalecer su capacidad de generalización.

- Evaluar arquitecturas más avanzadas de aprendizaje profundo, como redes LSTM, GRU o modelos basados en Transformers, para comparar su desempeño con los algoritmos implementados en esta investigación.

- Implementar mecanismos automáticos de actualización de la información proveniente de fuentes oficiales, con el fin de mantener vigente el modelo predictivo y facilitar su utilización en procesos continuos de planificación agrícola.

- Extender la metodología desarrollada a otros cultivos y regiones del Ecuador para analizar su capacidad de generalización y favorecer el desarrollo de herramientas de apoyo a la toma de decisiones en el sector agropecuario.

6. REFERENCIAS BIBLIOGRÁFICAS

Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). *Machine learning in agriculture: A comprehensive review Sensors, 21(11), 3758.* https://doi.org/10.3390/s21113758

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.

Bravo Guzmán, J. G., González Paz, R., Mora Cajas, K. E., & Vizcaíno Imacaña, F. P. (2024). *Uso de modelos de aprendizaje automático para predecir eventos climáticos en Ecuador* [Tesis de maestría, Universidad Internacional del Ecuador]. Repositorio Institucional de la Universidad Internacional del Ecuador. https://repositorio.uide.edu.ec/

Breiman, L. (2001). *Random forests. Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Campoverde, J. (2017). *Relación del cambio climático con la producción agrícola en la provincia del Azuay. INNOVA Research Journal, 2(*9.1), 55–64. https://doi.org/10.33890/innova.v2.n9.1.2017.556

Comisión Económica para América Latina y el Caribe. (2023). *La seguridad alimentaria y nutricional en el Ecuador: Desafíos y oportunidades frente a la crisis climática. Naciones Unidas.* https://www.cepal.org/

Chai, T., & Draxler, R. R. (2014). *Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in favor of MAE. Geoscientific Model Development, 7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system. En B. Krishnapuram et al. (Eds.), Proceedings of the 22nd ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Crane-Droesch, A. (2018). *Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Agricultural and Forest Meteorology, 264*, 151–162. https://doi.org/10.1016/j.agrformet.2018.10.018

Food and Agriculture Organization of the United Nations. (2022). *The state of food and agriculture 2022: Leveraging automation in agriculture for transforming agrifood systems. FAO*. https://doi.org/10.4060/cb9479en

Glen, S. (2016). *Coefficient of determination (R-squared): Definition, calculation, and interpretation. Statistics How To.* https://www.statisticshowto.com/probability-and-statistics/coefficient-of-determination-r-squared/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning. MIT Press.* https://www.deeplearningbook.org/

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.* https://doi.org/10.1007/978-0-387-84858-7

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice(3rd ed.). OTexts.* https://otexts.com/fpp3/

Instituto Nacional de Estadística y Censos. (2024*). Estadísticas agropecuarias y climáticas del Ecuador*. https://www.ecuadorencifras.gob.ec/

Instituto Nacional de Meteorología e Hidrología. (2023). *Anuario meteorológico 2023.* https://www.inamhi.gob.ec/

Intergovernmental Panel on Climate Change. (2022). *Climate change 2022: Impacts, adaptation and vulnerability. Cambridge University Press.* https://www.ipcc.ch/report/ar6/wg2/

Intergovernmental Panel on Climate Change. (2023). *Climate change 2023: Synthesis report*. https://www.ipcc.ch/report/ar6/syr/

James, G., Witten, D., Hastie, T., & Tibshirani, *R. (2021). An introduction to statistical learning: With applications in R (2nd ed.).* Springer. https://doi.org/10.1007/978-1-0716-1418-1

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). *Deep learning in agriculture: A survey*. *Computers and Electronics in Agriculture, 147*, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). *Machine learning in agriculture: A review*. *Sensors, 18*(8), 2674. https://doi.org/10.3390/s18082674

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data (3rd ed.). John Wiley & Sons*. https://doi.org/10.1002/9781119482260

Machuca Catagua, N., Navia Mendoza, M. R., & Cedeño Valarezo, L. (2025). *Evaluación de la eficacia de modelos de machine learning para la toma de decisiones basado en datos meteorológicos en la producción agrícola. Ecuadorian Science Journal, 9(2), 14–20.* https://doi.org/10.46480/esj.9.2.232

McKinney, W. (2022). *Python for data analysis: Data wrangling with pandas, NumPy, and Jupyter (3rd ed.).* O'Reilly Media.

Mendoza Cuzme, L. J., Machuca Ávalos, M. P., & González López, O. A. (2025). *Uso de modelos predictivos basados en inteligencia artificial para anticipar la producción agrícola en*

función de variables climáticas. Dominio de las Ciencias, 11(2), 680–703. https://doi.org/10.23857/dc.v11i2.4134

Ministerio de Agricultura y Ganadería. (2025). *Sistema de Información Pública Agropecuaria (SIPA).* https://sipa.agricultura.gob.ec/

Molnar, C. (2022). *Interpretable machine learning(2nd ed.).* https://christophm.github.io/interpretable-ml-book/

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis (6th ed.).* John Wiley & Sons.

Rosenzweig, C., Iglesias, A., Yang, X. B., Epstein, P. R., & Chivian, E. (2001). *Climate change and extreme weather events: Implications for food production, plant diseases, and pests*. *Global Change & Human Health, 2*(2), 90–104.

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. J. (2017). *Big data in smart farming: A review*. *Agricultural Systems, 153*, 69–80. https://doi.org/10.1016/j.agsy.2017.01.023

7. APÉNDICE

Apéndice A: Repositorio de Códigos y Datos

Con el objetivo de garantizar la transparencia, la trazabilidad y la reproducibilidad de los resultados presentados en esta investigación, el código fuente, los datasets utilizados y los scripts que componen el sistema inteligente de predicción agrícola han sido alojados bajo control de versiones.

- *Enlace al repositorio:* https://gitlab.com/bquimbita1724/modelado_predictivo_de_la_produccion_agricola
- *Contenido Técnico:* El repositorio integra el pipeline completo de análisis de datos (scripts en Python), los modelos serializados (.pkl), la aplicación funcional desarrollada en Streamlit y el archivo requirements.txt con la especificación de las dependencias del entorno de trabajo.
- *Instrucciones de uso:* Para la réplica de los experimentos, se requiere clonar el repositorio, configurar un entorno virtual (compatible con Python 3.x) e instalar las librerías especificadas. El flujo de ejecución y la secuencia lógica de los scripts se encuentran documentados en el archivo README.md ubicado en la raíz del proyecto.