MODELO PREDICTIVO DEL POTENCIAL DE RENDIMIENTO DEL CULTIVO DE MAÍZ A TRAVÉS DE TÉCNICAS DE APRENDIZAJE ESTADÍSTICO Y AUTOMÁTICO.

Sebastian Alejandro Ortiz Florez, Juan Sebastián Gutiérrez Orjuela y Laura Rojas Prado

Proyecto aplicado para optar al título de Magíster en Ciencia de Datos

Director

David Arango Londoño

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO DE 2026

FICHA RESUMEN
PROYECTO DE TRABAJO DE GRADO

1. **TÍTULO:** Modelo predictivo del potencial de rendimiento del cultivo de maíz a través de técnicas de aprendizaje estadístico y automático.
2. **ÁREA DE TRABAJO:** Agricultura.
3. **TIPO DE PROYECTO:** Aplicado.
4. **ESTUDIANTE(S):** Sebastian Alejandro Ortiz Florez, Juan Sebastián Gutiérrez Orjuela y Laura Rojas Prado
5. **CORREO ELECTRÓNICO:** sebastianortizf@javerianacali.edu.co, rojaslp@javerianacali.edu.co y juanorjuela@javerianacali.edu.co
6. **DIRECCIÓN Y TELÉFONO:** Universidad Javeriana de Cali.
7. **DIRECTOR:** David Arango Londoño.
8. **VINCULACIÓN DEL DIRECTOR:** Profesor de planta.
9. **CORREO ELECTRÓNICO DEL DIRECTOR:** david.arango@javerianacali.edu.co
10. **CO-DIRECTOR (Si aplica):** N/A
11. **GRUPO O EMPRESA QUE LO AVALA (Si aplica):** N/A
12. **OTROS GRUPOS O EMPRESAS:** N/A
13. **PALABRAS CLAVE:** Aprendizaje automático, Rendimiento potencial, Maíz, Datos agroclimáticos, Modelos predictivos, Agricultura 4.0, Colombia, modelo estocástico.
14. **FECHA DE INICIO:** 2025-1.
15. **DURACIÓN ESTIMADA:** 8 - 12 Meses.

RESUMEN

La producción de maíz en Colombia es un componente estratégico para la seguridad alimentaria y la economía agropecuaria nacional; sin embargo, su desempeño presenta elevada variabilidad espacio - temporal asociada a factores agroclimáticos, tecnológicos y estructurales que dificultan la planificación territorial y la gestión del riesgo agrícola. En este contexto, el proyecto propone un marco predictivo para estimar el rendimiento potencial del cultivo de maíz a escala municipal, integrando información agroproductiva y climática histórica de fuentes oficiales de acceso abierto: de la Unidad de Planificación Rural Agropecuaria (UPRA) y el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). El rendimiento potencial se operacionaliza mediante el cuantil $\tau = 0{,}95$ del rendimiento observado, criterio sustentado en la teoría de fronteras de producción y eficiencia técnica, pues representa un umbral productivo empíricamente alcanzable bajo condiciones eficientes y con menor sensibilidad a valores atípicos.

A partir del marco predictivo propuesto, el estudio estimará la frontera productiva municipal y cuantificará las brechas de rendimiento como la diferencia entre el rendimiento observado y dicha frontera, lo que permitirá aproximar escenarios de pérdidas de productividad e ingresos asociados a limitaciones técnicas, climáticas o de manejo agronómico. Se espera que los resultados constituyan una herramienta analítica reproducible y escalable, útil para productores, entidades gubernamentales, programas de extensión rural, aseguradoras y actores financieros del sector, en la focalización de asistencia técnica, la priorización de inversiones y la formulación de estrategias de adaptación agroclimática y planificación agrícola basada en evidencia cuantitativa en Colombia.

Contenido

Lista de figuras

Lista de tablas

INTRODUCCIÓN

El cultivo de maíz constituye un pilar de la agricultura colombiana y desempeña un papel central en la seguridad alimentaria, la generación de ingresos rurales y el abastecimiento de las cadenas agroindustriales del país. No obstante, su desempeño productivo se caracteriza por una marcada heterogeneidad espacial asociada a la diversidad de pisos térmicos en los que se desarrolla: las zonas templadas de la Sabana, los valles cálidos interandinos y los pisos cálido - húmedos presentan diferencias sustanciales en duración del ciclo del cultivo, potencial productivo, disponibilidad hídrica y prácticas de manejo agronómico. Estas diferencias explican por qué departamentos como Tolima, Meta y Córdoba registran respuestas productivas notoriamente distintas aun cuando se emplean variedades similares, lo que evidencia la necesidad de aproximaciones analíticas que reconozcan explícitamente la variabilidad agroclimática del territorio nacional [1][2].

El problema es la falta de modelos predictivos que integren explícitamente las variables climatológicas por municipio y la heterogeneidad de prácticas agrícolas. Sin esos modelos la planificación local y el uso eficiente de insumos se dificultan, y se incrementa la vulnerabilidad de sistemas productivos en pisos térmicos específicos. Los datos e inventarios nacionales muestran diferencias sistemáticas de rendimiento entre pisos térmicos y una cobertura limitada de modelos con resolución municipal [2][1], mientras que la bibliografía científica nacional presenta relativamente pocas aplicaciones de modelado de rendimiento con validación espacio - temporal a escala regional, así como pocos trabajos metodológicos aplicados a contextos productivos concretos como el de Córdoba; en conjunto, estos antecedentes confirman la existencia de una brecha metodológica vinculada a la ausencia de modelos predictivos que integren de manera explícita las variables climatológicas por municipio y la heterogeneidad de las prácticas agrícolas [3].

Frente a este panorama, el presente proyecto propone desarrollar un modelo predictivo estocástico y escalable para estimar el rendimiento potencial del maíz a nivel municipal, integrando datos agroclimáticos abiertos y series históricas productivas, y articulando enfoques de fronteras estocásticas de producción, regresión cuantílica y aprendizaje automático cuantílico. La contribución específica del estudio respecto a la literatura citada radica en tres aspectos: primero, la operacionalización del rendimiento potencial mediante un cuantil alto del rendimiento observado, lo que permite un referente empíricamente alcanzable y robusto frente a valores atípicos; segundo, la construcción de una base municipio - semestre que combina información de la UPRA y del IDEAM bajo un esquema reproducible de depuración e integración; y tercero, la cuantificación de brechas de eficiencia técnica con resolución municipal, dimensión poco explorada en los trabajos previos de alcance regional [1] [2] [3]. Se espera que los hallazgos aporten evidencia cuantitativa reproducible sobre el rendimiento potencial y las brechas de eficiencia técnica

municipales, constituyendo un referente analítico útil para investigadores, entidades gubernamentales y actores del sector agropecuario en la formulación de estrategias basadas en evidencia para la gestión del riesgo y la planificación agrícola en Colombia.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

El maíz es uno de los cultivos agrícolas de mayor importancia económica a nivel mundial, tanto por su valor como alimento humano al ser uno de los granos más antiguos y ampliamente consumidos como por su rol en la alimentación animal y su versatilidad como insumo en numerosas industrias. Su capacidad de adaptación a distintos entornos geográficos y climáticos ha permitido su cultivo en una amplia gama de condiciones, desde regiones ecuatoriales hasta zonas templadas y desde niveles cercanos al mar hasta altitudes medias y altas [4].

En Colombia, el maíz representa un componente clave en la seguridad alimentaria y la economía rural, con presencia en casi todos los departamentos. Sin embargo, a pesar de su relevancia, la productividad del maíz a nivel municipal presenta una alta variabilidad interanual e interregional [5]. Esta variabilidad se debe a factores como las diferencias en las prácticas agronómicas, el uso de tecnologías, la calidad de los suelos y, de manera destacada, las condiciones climáticas locales, las cuales influyen directamente en el rendimiento de las cosechas.

Aunque el país cuenta con fuentes de datos agroclimáticos abiertos, como las Evaluaciones Agropecuarias Municipales (EVA) generadas por la Unidad de Planificación Rural Agropecuaria (UPRA) y los registros climáticos históricos del IDEAM, estos datos rara vez se integran o analizan de forma sistemática para generar conocimiento útil. En consecuencia, las decisiones suelen basarse en promedios históricos; este proyecto usará esos datos como referencia para estimar la brecha frente al rendimiento potencial. El presente estudio propone aprovechar estas fuentes de datos mediante su integración y análisis sistemático, con el fin de estimar el rendimiento potencial municipal del maíz y cuantificar las brechas respecto al rendimiento observado.

Esta desconexión entre la disponibilidad de datos y su uso analítico tiene consecuencias visibles:

- Incapacidad para identificar brechas de rendimiento entre lo observado y lo potencial.
- Incertidumbre en la planeación agrícola.
- Falta de estrategias diferenciadas por territorio.
- Escasa anticipación a eventos climáticos adversos.

Desde la ciencia de datos, el sector agrícola colombiano tiene una oportunidad estratégica para superar sus desafíos mediante la integración y análisis de datos agroclimáticos. La adopción de la

Agricultura 4.0, que incluye el uso de Big Data, Inteligencia Artificial y técnicas avanzadas como el aprendizaje automático y el modelado predictivo, puede transformar la forma en que se gestionan los cultivos, proporcionando estimaciones precisas de rendimientos potenciales por municipios [6]. Esta integración de datos no solo optimiza la toma de decisiones, sino que también mejora el uso de recursos y aumenta la resiliencia del sector frente a los cambios climáticos. No obstante, la falta de infraestructura tecnológica en zonas rurales y la baja adopción de herramientas digitales por parte de los actores del sector agrario siguen siendo barreras clave que deben abordarse para maximizar los beneficios de estas tecnologías.

1.2. FORMULACIÓN DEL PROBLEMA

¿Cómo puede un modelo predictivo basado en técnicas de aprendizaje automático y estadístico alimentado con datos abiertos agroclimáticos históricos de Colombia, contribuir a estimar el potencial de rendimiento del cultivo de maíz a escala municipal y apoyar la toma de decisiones estratégicas en el sector agropecuario?

Preguntas de sistematización

1. ¿Cómo integrar y preprocesar datos abiertos provenientes de distintas fuentes para construir una base de datos unificada y funcional?

2. ¿Qué variables agroclimáticas públicas disponibles son más relevantes para predecir el rendimiento potencial del maíz a nivel municipal?

3. ¿Cómo desarrollar un modelo de aprendizaje automático supervisado y estocástico que ayude a predecir el rendimiento potencial del cultivo de maíz a nivel municipal empleando las variables seleccionadas?

4. ¿Cuáles modelos de aprendizaje estadístico y automático son los ideales para predecir el rendimiento potencial del cultivo de maíz a nivel municipal?

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

Predecir el rendimiento potencial del cultivo de maíz a escala municipal en Colombia mediante el desarrollo de un modelo predictivo, utilizando técnicas de aprendizaje automático y datos abiertos agroclimáticos históricos.

2.2. OBJETIVOS ESPECÍFICOS

- Integrar y preprocesar los datos agroclimáticos abiertos provenientes de distintas fuentes, con el fin de construir una base de datos unificada y funcional para su análisis.
- Identificar las variables agroproductivas y climáticas más influyentes sobre el rendimiento potencial del cultivo de maíz, mediante análisis exploratorios y técnicas de selección de características.
- Desarrollar modelos estocásticos y de aprendizaje automático para la predicción del rendimiento potencial del cultivo de maíz a nivel municipal, utilizando datos agroclimáticos históricos.
- Evaluar modelos de aprendizaje estadístico y automático mediante validación cruzada y métricas adecuadas, para predecir el rendimiento potencial del cultivo de maíz a nivel municipal.

2.3. RESULTADOS ESPERADOS

- Una base de datos unificada limpia y preprocesada que integre información climatológica y agroproductiva a nivel municipal en Colombia.
- Un conjunto de variables agroclimáticas identificadas como las más influyentes sobre el rendimiento potencial del cultivo de maíz, sustentado en análisis exploratorios y técnicas de selección de características.
- Un conjunto de modelos estocásticos y de aprendizaje automático desarrollados y entrenados para la estimación del rendimiento potencial del maíz a nivel municipal, utilizando datos agroclimáticos históricos.
- Una evaluación comparativa de los modelos desarrollados mediante validación cruzada temporal y métricas de precisión puntual y calibración probabilística, que permita identificar la especificación con mejor desempeño predictivo y mayor robustez frente a la heterogeneidad espaciotemporal.

3. ALCANCE

1. Integración y preprocesamiento de datos

Se construirá una base de datos agroclimática unificada, compuesta por:

- Datos agropecuarios históricos de la EVA - UPRA.
- Datos climáticos históricos del IDEAM.

Los datos serán sometidos a procesos de limpieza, estandarización, interpolación de valores faltantes y alineación espacio y temporal. Se excluirán municipios cuya información no cumpla con los criterios mínimos de calidad y completitud requeridos para el análisis.

2. Análisis de variables relevantes

Se realizará un análisis exploratorio y estadístico con el fin de identificar las variables agroproductivas y climatológicas que más influyen en el rendimiento potencial del maíz. Este análisis se desarrollará a nivel municipal y combinará métodos tradicionales de exploración de datos con técnicas modernas de selección de características, tales como algoritmos de importancia de variables en modelos de aprendizaje automático.

3. Desarrollo y validación de modelos predictivos

Se entrenarán y evaluarán modelos estadísticos y de aprendizaje automático. La elección del modelo final se basará en su capacidad predictiva medida a través de métricas estadísticas específicas, como el error medio absoluto (MAE) y el coeficiente de determinación (R^2), entre otros. El enfoque del modelado será estocástico, permitiendo la estimación de la ineficiencia en las predicciones.

El modelo seleccionado se validará mediante un esquema de validación cruzada con estructura temporal (rolling window) y su desempeño se contrastará sobre un conjunto de prueba independiente, asegurando la coherencia con la naturaleza secuencial de los datos y evitando fuga de información.

4. Evaluación del modelo final

Para medir la precisión puntual se utilizarán el MAE, el RMSE y el R^2, calculados sobre la variable en escala logarítmica. Dado el enfoque cuantílico del problema, también se calcularán métricas de evaluación probabilística. En particular, se empleará la Pinball Loss y Pinball normalizada como medida principal de ajuste de los cuantiles estimados, y el Interval Score (IS) para evaluar la calidad de los intervalos de predicción, considerando de forma conjunta su amplitud y su capacidad de cobertura. Además, se analizará la calibración del modelo a través del Error de Calibración (EC), definido como la diferencia entre la cobertura empírica observada (Coverage) y el nivel nominal

esperado (τ), con el fin de verificar la consistencia entre la incertidumbre estimada y la observada en los datos.

5. Limitaciones del proyecto

El alcance del proyecto se limita expresamente a la construcción y validación del modelo predictivo, sin incluir el desarrollo de una plataforma digital funcional ni el despliegue de la solución en tiempo real. No se implementará un sistema de acceso abierto ni se realizará integración automatizada con servicios externos o flujos de datos en tiempo real. Estas funcionalidades podrán considerarse en fases futuras del proyecto.

6. Delimitación geográfica y temporal

- **Geográfica**: El análisis se restringirá al territorio colombiano y se enfocará únicamente en los municipios para los que se cuente con datos disponibles de las fuentes mencionadas.
- **Temporal**: El análisis se basará exclusivamente en series históricas disponibles. No se incorporarán datos proyectados ni actualizaciones dinámicas durante el desarrollo del proyecto.

4. JUSTIFICACIÓN

El desarrollo de un modelo predictivo para estimar el rendimiento potencial de cultivos en Colombia mediante técnicas de aprendizaje automático y estadístico representa una respuesta innovadora y necesaria ante los desafíos estructurales que enfrenta el sector agropecuario. En particular, este proyecto es útil porque permite aprovechar grandes volúmenes de datos históricos climatológicos que, aunque disponibles públicamente, no se integran de forma sistemática ni se traducen en herramientas accesibles para la toma de decisiones productivas. Esta desconexión entre la disponibilidad de datos y su aprovechamiento efectivo limita la capacidad de respuesta de los productores frente a fenómenos como el cambio climático, la degradación de suelos y las ineficiencias productivas persistentes [7].

Desde el punto de vista de la viabilidad, el proyecto es realizable técnica y metodológicamente. Se cuenta con fuentes de datos abiertos y consolidadas como las Evaluaciones Agropecuarias Municipales (EVA) de la UPRA y los registros climáticos del IDEAM, los cuales permiten construir un conjunto de entrenamiento suficiente para el desarrollo de modelos robustos.

Las variables disponibles en la base de datos de la EVA son:

- Departamento
- Municipio
- Desagregación cultivo: Nombre genérico del cultivo
- Cultivo: Nombre del cultivo
- Año: Año de producción
- Periodo: Periodo de cultivo
- Área sembrada (ha): unidad de medida en hectáreas
- Área cosechada (ha): unidad de medida en hectáreas
- Producción (t): Tiempo de producción
- Rendimiento (t/ha): Rendimiento de la cosecha

Ahora, las variables climatológicas que se tomarán de la página del IDEAM son:

- **Temperatura húmeda y seca:** es el promedio aritmético de las temperaturas húmedas y secas diarias registradas durante todos los días de un mes determinado, expresado en grados Celsius (°C).
- **Humedad relativa:** representa el porcentaje de saturación del vapor de agua presente en el aire y registrado cada hora.
- **Precipitación:** Es la cantidad total de lluvia (o cualquier forma de precipitación como llovizna,

nieve, granizo, etc.) que ha caído en un mes completo, expresada en milímetros.

- **Velocidad del viento:** representa el valor promedio de la velocidad del viento registrado durante todo el día (24 horas), medido en metros por segundo (m/s).
- **Dirección del viento:** representa el punto cardinal de donde proviene, medida en grados sexagesimales (0° - 360°) en el sentido de las manecillas del reloj desde el norte geográfico.

Por otro lado, el maíz fue seleccionado debido a su importancia estratégica como tercer cultivo con mayor área sembrada y alta demanda nacional, que supera ampliamente la producción local, generando una gran dependencia de importaciones. Además, cuenta con una amplia disponibilidad de datos agroclimáticos y productivos que facilitan el desarrollo de modelos de aprendizaje automático capaces de capturar la compleja interacción entre variables climáticas y sistemas productivos (tecnificados y tradicionales). Esto permite construir modelos robustos que contribuyan a anticipar el rendimiento, apoyando la planificación agrícola y alineándose con el plan estratégico "Maíz para Colombia", orientado a incrementar la productividad sostenible y la seguridad alimentaria en el país [8].

El proyecto tiene un impacto positivo proyectado en múltiples niveles, al integrar datos históricos de producción y variables climatológicas para predecir el rendimiento potencial de los cultivos. A nivel territorial e institucional, el modelo de predicción puede llegar a orientar en la generación de indicadores basados en evidencia y en la toma de decisiones estratégicas por parte de entidades públicas, cooperativas y empresas agroindustriales, posibilitando la focalización de inversiones, asistencia técnica y políticas agrícolas adaptadas a las condiciones locales. En conjunto, esta iniciativa promueve un uso estratégico de los datos abiertos, aportando al desarrollo rural sostenible en Colombia [7]

Además, se encuentra alineado con varios Objetivos de Desarrollo Sostenible (ODS), establecidos por la Organización de las Naciones Unidas (ONU) en el marco de la Agenda 2030 [7], como una hoja de ruta global para erradicar la pobreza, proteger el planeta y promover el desarrollo sostenible. En particular:

- **ODS 1: Fin de la Pobreza**, al generar herramientas basadas en datos que pueden ser utilizadas por pequeños y medianos productores para mejorar su capacidad de decisión y reducir riesgos productivos.

- **ODS 2: Hambre Cero**, al proponer modelo de aprendizaje automático basado en datos que mejoran la eficiencia productiva y la planificación agrícola basada en evidencia.

- **ODS 9: Industria, Innovación e Infraestructura**, al aplicar tecnologías como el aprendizaje automático y la ciencia de datos en el contexto agropecuario.

- **ODS 12: Producción y Consumo Responsables**, al promover un uso más racional y sostenible de los recursos agrícolas mediante análisis predictivo.

- **ODS 13: Acción por el Clima**, al integrar datos de variables climáticas en modelos predictivos que permiten la anticipación de escenarios y adaptar la producción a condiciones cambiantes.

- **ODS 17: Alianzas para lograr los objetivos**, al utilizar y articular información de datos abiertos provenientes de entidades públicas como el IDEAM y la UPRA, fomentando la interoperabilidad institucional y la colaboración intersectorial basada en evidencia.

5. MARCO DE REFERENCIA

5.1. MARCO TEÓRICO

El marco teórico se organiza en nueve secciones que cubren los fundamentos matemáticos de los modelos empleados, las métricas de evaluación y los conceptos agroclimáticos que sustentan la especificación de variables. Las secciones 5.1.1 a 5.1.4 presentan los modelos de estimación de frontera de producción; 5.1.5 define el sistema de métricas; 5.1.6 a 5.1.8 desarrollan los fundamentos de las variables de entrada y el tratamiento econométrico del panel.

Frontera de Producción y Operacionalización del Rendimiento Potencial

El rendimiento potencial se define como el rendimiento máximo de un cultivo bajo las condiciones climáticas de un sitio dado, supuesto que los factores limitantes bióticos y abióticos están controlados [9]. En datos observacionales de panel municipal, ese techo biofísico no es directamente observable: los registros disponibles provienen de explotaciones con distintos niveles de insumos, tecnología y gestión. La econometría de frontera de producción aborda este problema descomponiendo el rendimiento observado en un componente de frontera y un componente de ineficiencia [10].

Esta investigación adopta el cuantil $\tau = 0{,}95$ del rendimiento observado como proxy operacional del potencial. La elección se apoya en tres argumentos. Primero, la literatura de yield gap analysis emplea percentiles altos de la distribución empírica para definir rendimientos alcanzables (attainable yields) dentro de zonas climáticas homogéneas, dado que el cuantil 95 corresponde a los municipios que operan con mínima ineficiencia bajo la tecnología disponible [11][12]. Segundo, en el marco SFA, la estimación de $\tau = 0{,}95$ equivale a filtrar el componente de ineficiencia u hasta dejarlo cercano a cero, obteniendo una frontera no paramétrica consistente con la formulada por Aigner, Lovell y Schmidt [10]. Tercero, la función de pérdida pinball para $\tau = 0{,}95$ asigna un peso 19 veces mayor a los residuos positivos que a los negativos ($\tau/(1-\tau) = 19$), garantizando que el estimador se posicione en la cola superior de la distribución condicional [13].

Implicaciones de la elección de $\tau = 0{,}95$

A diferencia del máximo absoluto ($\tau = 1{,}00$), el cuantil 0,95 es resistente a errores de medición y a rendimientos excepcionales irrepetibles en las fuentes primarias. Al excluir el 5% superior, el potencial estimado corresponde a la mejor práctica estadísticamente reproducible, no a un caso singular. La diferencia entre el rendimiento observado de un municipio y el cuantil 0,95 estimado se interpreta como brecha de eficiencia técnica accionable mediante mejoras de gestión agronómica, en analogía con la eficiencia media estimada por SFA en la literatura aplicada a países en desarrollo [14].

5.1.1. Rendimiento Potencial Municipal como Objetivo de Modelación

El rendimiento potencial de un cultivo corresponde al rendimiento máximo alcanzable bajo las condiciones climáticas locales cuando los factores limitantes son gestionados de forma óptima [9]. A escala municipal, se operacionaliza como el cuantil τ = 0,95 de la distribución empírica de rendimientos: el nivel que solo el 5% de los municipios supera en condiciones de gestión excelente. Este cuantil es estimable directamente con datos observacionales, sin condiciones experimentales controladas.

El panel municipio-semestre del cultivo de maíz en Colombia (2006 – 2024, n ≈ 31.514 observaciones) permite estimar ese cuantil superior con enfoques paramétricos SFA y regresión cuantílica que imponen estructura económica al problema, y con métodos no paramétricos QRF y XGBoost que capturan interacciones no lineales entre variables agroclimáticas.

5.1.2. Regresión Cuantílica para Estimación de Rendimiento Potencial

La regresión cuantílica (Koenker & Bassett, 1978) [13] extiende el modelo de regresión de la media condicional E[Y|X] al cuantil condicional Q_τ[Y|X], definido como el valor q que satisface P(Y ≤ q | X = x) = τ. Para τ = 0.95, el modelo estima la frontera superior de la distribución de rendimientos condicionada a las covariables agroclimáticas. No requiere supuestos sobre la distribución del término de error, por lo que es robusto a heterocedasticidad y asimetría, características comunes en datos agronómicos de panel [13].

Incorporar efectos fijos α_i por municipio controla la heterogeneidad geográfica permanente (altitud, tipo de suelo, infraestructura) que no varía en el tiempo:

$$Q_\tau[y_{it}|x_{it}] = \alpha_i + x_{it}^T\beta_\tau \quad (1)$$

donde: Q_τ = cuantil condicional de nivel τ; y_{it} = rendimiento del municipio i en semestre t; x_{it} = vector de covariables agroclimáticas; α_i = efecto fijo municipal; β_τ = coeficientes estimados.

Canay (2011) [15] demuestra que, bajo condiciones de panel largo, el estimador de efectos fijos para regresión cuantílica es consistente.

5.1.3. Métodos de Aprendizaje de Máquina para Predicción Cuantílica

Los métodos de aprendizaje de máquina para predicción cuantílica no asumen forma funcional entre covariables y rendimiento, lo que permite capturar interacciones no lineales y efectos de umbral que los modelos paramétricos no representan. Esta flexibilidad es particularmente relevante cuando las relaciones clima-rendimiento son no monótonas o presentan umbrales abruptos.

5.1.3.1. Quantile Random Forest (QRF)

El Quantile Random Forest (Meinshausen, 2006) [16] extiende los bosques aleatorios reteniendo, en cada hoja de cada árbol, la distribución empírica completa de observaciones de entrenamiento no solo su media. Para un punto de predicción x, el cuantil τ estimado es el valor q que satisface:

$$\hat{Q}_\tau(Y|X=x): \Sigma_i w_i(x) \cdot 1y_i \leq q \geq \tau \quad (2)$$

donde: $\hat{Q}_\tau$ = cuantil estimado; $w_i(x)$ = peso de co-asignación en hojas del bosque; y_i = rendimiento observado en entrenamiento; q = valor del cuantil; τ = nivel de cuantil.

El QRF no impone linealidad en la relación clima-rendimiento y estima intervalos de incertidumbre sin supuestos distribucionales [16].

5.1.3.2. XGBoost Cuantílico

XGBoost (Chen & Guestrin, 2016) [17] construye secuencialmente un ensamble de T árboles débiles $\{f_1, \ldots, f_T\}$, donde cada f_t corrige los residuos del ensamble previo:

$$\hat{y}_i(T) = \Sigma_{t=1}^{T} f_t(x_i) \quad (3)$$

donde: $\hat{y}_i(T)$ = predicción acumulada final; T = número total de árboles; f_t = árbol de la iteración t; x_i = covariables de la observación i.

La extensión cuantílica reemplaza la pérdida cuadrática por la Pinball Loss ρ_τ, con regularización explícita:

$$L^{(t)} = \Sigma_i \rho_\tau \left(y_i - \hat{y}_i^{(t-1)} - f_t(x_i) \right) + \Omega(f_t) \quad (4)$$

donde: ρ_τ = función Pinball Loss; y_i = rendimiento observado; $\hat{y}_i^{(t-1)}$ = predicción acumulada hasta t−1; f_t = árbol de la iteración t; $\Omega(f_t)$ = término de regularización del árbol [17].

Los valores SHAP (Lundberg & Lee, 2017) [18] permiten descomponer cada predicción en contribuciones atribuibles a cada variable climática, facilitando la interpretación agronómica a nivel municipal.

5.1.3.3. Ensemble XGBoost + QRF

XGBoost supera a QRF en precisión puntual (MAE = 0,130 vs 0,242 en $\log_{10}$) pero presenta menor error de calibración (EC = -0,039 vs +0,048). Un ensamble que use XGBoost como predictor puntual y QRF para construir los intervalos de predicción captura ambas fortalezas, combinando la precisión del boosting con la calibración distribucional del bosque.

5.1.4. Frontera Estocástica de Producción (SFA) – Especificación Cobb-Douglas

El modelo de frontera estocástica (Aigner, Lovell & Schmidt, 1977; Meeusen & van den Broeck, 1977) [10] [19] descompone el rendimiento observado en frontera técnica e ineficiencia:

$$\ln(y_{it}) = \beta_0 + \Sigma_k \beta_k \ln(x_{itk}) + v_{it} - u_{it} \quad (5)$$

donde $v_{it} \sim N(0, \sigma^2_v)$ es ruido aleatorio simétrico y $u_{it} \sim N^+(\mu, \sigma^2_u)$ es el término de ineficiencia técnica unilateral ($u_{it} \geq 0$). La proporción de varianza total atribuible a ineficiencia es:

$$\gamma = \sigma_u^2/(\sigma_u^2 + \sigma_v^2) \quad (6)$$

donde: γ = proporción de varianza explicada por ineficiencia técnica; σ^2_u = varianza del componente de ineficiencia; σ^2_v = varianza del ruido aleatorio.

γ → 0 indica que la variación es principalmente ruido no controlable; γ → 1, que es ineficiencia técnica. La especificación Cobb-Douglas estima γ = 0,807 en este estudio, coherente con la literatura de SFA aplicada a maíz en países en desarrollo, donde γ típicamente se encuentra entre 0,70 y 0,90 [14]. La eficiencia técnica de cada municipio se obtiene como la esperanza condicional del componente de eficiencia dado el error compuesto observado (Battese & Coelli, 1988) [20]:

$$TE_{it} = E[\exp(-u_{it}) \,|\varepsilon_{it}] = \exp(-\hat{u}_{it}) \quad (7)$$

donde: TE_{it} = eficiencia técnica del municipio i en t; u_{it} = ineficiencia técnica; ε_{it} = error compuesto observado; $\hat{u}_{it}$ = estimación de la ineficiencia.

TE = 1 corresponde a operar sobre la frontera; TE = 0,73 indica que el municipio opera al 73% de su potencial, con una brecha del 27% accionable sin modificar las condiciones climáticas.

5.1.5. Métricas de Evaluación del Modelo

La evaluación de modelos cuantílicos requiere métricas que cubran dos dimensiones independientes: precisión puntual (proximidad de la predicción al valor observado) y calidad probabilística (calibración y amplitud de los intervalos). Las ocho métricas empleadas se organizan en tres grupos.

Grupo 1: Precisión Puntual

5.1.5.1. Error Absoluto Medio (MAE)

El MAE usa la norma L_1 y es resistente a observaciones atípicas:

$$MAE = (1/n)\Sigma_{i=1}{}^{n}|y_i - \hat{y}_i| \quad (8)$$

donde: n = número de observaciones; y_i = rendimiento observado; $\hat{y}_i$ = cuantil τ = 0,95 estimado.

$\hat{y}_i$ es el cuantil τ = 0,95 estimado. En escala $\log_{10}$, un MAE = 0,130 equivale a un factor multiplicativo de 10^0,130 ≈ 1,35 en escala natural.

5.1.5.2. Raíz del Error Cuadrático Medio (RMSE)

El RMSE usa la norma L_2 y penaliza errores grandes de forma cuadrática:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (9)$$

donde: n = número de observaciones; y_i = rendimiento observado; $\hat{y}_i$ = cuantil estimado.

La relación RMSE/MAE cuantifica el peso de la cola de la distribución de errores: bajo normalidad se espera RMSE/MAE ≈ √(π/2) ≈ 1,25; valores superiores indican colas pesadas.

5.1.5.3. Pseudo R^2 Cuantílico

El Pseudo R^2 (Koenker & Machado, 1999) [21] extiende el coeficiente de determinación al marco cuantílico:

$$R_\tau^2 = 1 - \left[\Sigma_i \rho_\tau\left(y_i - \hat{q}_\tau(x_i)\right)\right] / \left[\Sigma_i \rho_\tau(y_i - \hat{Q}_\tau)\right] \quad (10)$$

donde: ρ_τ = función Pinball Loss; q̂_τ(x_i) = cuantil predicho con covariables; Q̂_τ = cuantil incondicional.

q̂_τ(x_i) es el cuantil predicho con covariables; Q̂_τ, el cuantil incondicional (modelo nulo). R^2_τ = 1 indica ajuste perfecto; R^2_τ < 0 indica que las covariables perjudican la predicción respecto al cuantil muestral. La identidad R^2_τ = 1 − PL_norm conecta esta métrica con el Pinball Loss Normalizado.

Grupo 2: Calidad Probabilística

5.1.5.4. Pinball Loss (Pérdida Cuantílica)

La Pinball Loss evalúa la calidad de las predicciones cuantílicas mediante la misma función usada en la estimación:

$$PL_\tau = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(y_i - \hat{q_\tau}(x_i)\right) \quad (11)$$

donde: n = número de observaciones; ρ_τ = función Pinball Loss; y_i = rendimiento observado; q̂_τ(x_i) = cuantil estimado dado x_i.

La Pinball Loss es una regla de puntuación propia (proper scoring rule): su valor esperado se minimiza únicamente cuando q̂_τ coincide con el cuantil τ verdadero de la distribución condicional de Y dado X (Gneiting & Raftery, 2007) [22]. Ningún modelo puede mejorar artificialmente esta métrica reportando cuantiles sesgados.

5.1.5.5. Pinball Loss Normalizado

Para comparar modelos en distintas escalas o submuestras, la Pinball Loss se normaliza respecto al modelo nulo:

$$PL_{norm} = \frac{PL_\tau(\text{modelo})}{PL_\tau(\text{nulo})} \quad (12)$$

donde: PL_τ(modelo) = pérdida Pinball del modelo con covariables; PL_τ(nulo) = pérdida Pinball del cuantil incondicional.

PL_τ(nulo) = (1/n) Σᵢ ρ_τ(yᵢ − Q̂_τ). Un valor PL_norm > 1 implica que las covariables añaden ruido: condición suficiente para descartar el modelo. PL_norm = 0 corresponde a predicción perfecta. La equivalencia con el Pseudo R^2 es PL_norm = 1 − R^2_τ.

5.1.5.6. Interval Score (Winkler Score)

El Interval Score (Winkler, 1972; Gneiting & Raftery, 2007) [22] [23]evalúa simultáneamente amplitud y cobertura del intervalo de predicción [l̂ᵢ, ûᵢ] al nivel (1−α):

$$IS_\alpha(\hat{l}, \hat{u}, y) = (\hat{u} - \hat{l}) + \frac{2}{\alpha}(\hat{l} - y) \cdot 1_{y<\hat{l}} + \frac{2}{\alpha}(y - \hat{u}) \cdot 1_{y>\hat{u}} \quad (13)$$

donde: Î = límite inferior del intervalo de predicción; û = límite superior; y = valor observado; α = nivel de significancia.

El término (û − Î) penaliza la amplitud; los términos condicionales penalizan la subcobertura con factor 2/α. Para τ = 0,95 (α = 0,05), cada unidad fuera del intervalo penaliza 40 unidades adicionales de amplitud. El IS también es una regla de puntuación propia: su mínimo en esperanza corresponde al intervalo correcto al nivel nominal [22].

Grupo 3: Calibración Probabilística

5.1.5.7. Cobertura Empírica (Coverage)

La cobertura empírica mide la frecuencia con que las observaciones caen dentro del intervalo [Îᵢ, ûᵢ]:

$$Cov = \frac{1}{n}\sum_{i=1}^{n} 1_{y_i \in [\hat{l}_i, \hat{u}_i]} \quad (14)$$

donde: n = número de observaciones; y_i = rendimiento observado; Îᵢ = límite inferior del intervalo; ûᵢ = límite superior del intervalo.

Para un intervalo al nivel nominal (1−α) = 0,90, la cobertura ideal es Cov = 0,90. Subcobertura (Cov < nominal) indica que el modelo subestima su propia incertidumbre; sobrecobertura (Cov > nominal), que los intervalos son innecesariamente anchos.

5.1.5.8. Error de Calibración

El Error de Calibración cuantifica la desviación entre la cobertura empírica y el nivel nominal:

$$EC = Cov_{emp} - \tau \quad (15)$$

donde: EC = error de calibración; Cov_{emp} = cobertura empírica observada; τ = nivel de cuantil nominal.

Para τ = 0,95, el Error de Calibración se define como la diferencia entre la cobertura empírica observada y el nivel nominal esperado, es decir, EC = Cov_emp − 0,95. Valores positivos indican sobrecobertura y valores negativos indican subcobertura; sin embargo, la interpretación debe considerar principalmente la magnitud absoluta de la desviación. En este estudio, el XGBoost cuantílico presentó un EC de -0,039, equivalente a una subcobertura moderada de 3,9 puntos porcentuales.

5.1.6. Temperatura de Bulbo Húmedo y Estrés Térmico-Hídrico en Zea mays

La temperatura de bulbo húmedo (T_wb) es la temperatura de equilibrio que alcanza una superficie de agua en contacto con el aire bajo evaporación adiabática a presión constante. A diferencia de la temperatura de bulbo seco, integra temperatura y contenido de vapor de agua, constituyendo un indicador del estrés térmico-hídrico simultáneo. Stull (2011) [24] propone la siguiente aproximación empírica a partir de T (°C) y HR (%):

$$T_w b = T \cdot \arctan\left[0{,}151977 \cdot (HR + 8{,}313659)^{1/2}\right] + \arctan(T + HR) - \arctan(HR - 1.676331) + 0{,}00391838 \cdot HR^{3/2} \cdot \arctan(0{,}023101 \cdot HR) - 4{,}686035 \quad (16)$$

donde: T_wb = temperatura de bulbo húmedo (°C); T = temperatura de bulbo seco (°C); HR = humedad relativa (%).

Esta variable compuesta muestra mayor poder predictivo sobre el rendimiento del maíz que T y HR por separado, dado que captura conjuntamente el déficit de presión de vapor y la temperatura del dosel.

5.1.6.1. Sensibilidad Térmica del Maíz por Etapa Fenológica

Floración (R1–R2): T_wb > 25°C reduce la viabilidad del polen. Lobell et al. (2011) [25] estiman una pérdida de rendimiento de 1% por cada día adicional con T_wb > 29°C durante este período.

Llenado de grano (R4–R6): temperaturas nocturnas T_mín > 22°C aceleran la senescencia foliar y reducen el peso de grano en 6–8% (Hatfield & Prueger, 2015) [26].

Variabilidad ENSO: en la región Andina-Caribe colombiana, el fenómeno El Niño reduce precipitaciones en 20–40% e incrementa temperaturas en 1–2°C; La Niña produce el patrón opuesto. T_wb captura esta dualidad, lo que explica fluctuaciones de rendimiento de ±35% respecto

al promedio histórico (Poveda, Álvarez & Rueda, 2011) [27].

5.1.7. Autocorrelación Espacial en Paneles Municipales

Los rendimientos de municipios adyacentes están correlacionados por condiciones climáticas compartidas, prácticas agronómicas regionales y redes de difusión tecnológica. Ignorar esta dependencia produce errores estándar subestimados e inferencia inválida en los modelos paramétricos (Anselin, 1988) [28].

5.1.7.1. Índice de Moran y Modelos SAR/SEM

El índice de Moran I detecta autocorrelación espacial global:

$$I = (n/S_0) \cdot (z^{\mathrm{T}} W z / z^{\mathrm{T}} z) \quad (17)$$

donde $z = y - \bar{y}$, $S_o = \Sigma_i \Sigma_j w_{ij}$ y W es la matriz de pesos espaciales. Para paneles municipales agrícolas colombianos, la literatura reporta $I \in [0{,}25, 0{,}55]$ en paneles agrícolas de economías emergentes. El modelo SAR ($y = \rho W y + X\beta + \varepsilon$) captura la dependencia directa entre vecinos; el modelo SEM ($\varepsilon = \lambda W \varepsilon + u$) captura la dependencia inducida por variables omitidas. Los errores Conley HAC (Conley, 1999) [29] permiten corrección de inferencia sin re-especificar el modelo, permitiendo correlación dentro de un radio d_o de 100–300 km.

5.1.8. Multicolinealidad entre Variables Agroclimáticas

Precipitación y humedad relativa presentan correlaciones de Pearson superiores a r = 0,75 a escala mensual en condiciones tropicales, por su origen común en el ciclo hidrológico. Esta correlación induce coeficientes inestables con signos agronómicamente incoherentes la elasticidad negativa de precipitación observada en las especificaciones paramétricas es el síntoma más frecuente de VIF elevado.

El Factor de Inflación de Varianza cuantifica el grado de multicolinealidad:

$$VIF_j = 1/\left(1 - R_j^2\right) \quad (18)$$

donde R^2_j es el coeficiente de determinación de la regresión de la variable j sobre las demás covariables. VIF > 10 indica multicolinealidad problemática; VIF > 30 hace los coeficientes esencialmente no identificables. Las estrategias de mitigación aplicables son: (1) PCA sobre las variables climáticas para obtener componentes ortogonales; (2) índices compuestos como el SPI (Standardized Precipitation Index) o el PDSI (Palmer Drought Severity Index); (3) regularización Ridge cuantílica. La temperatura de bulbo húmedo actúa implícitamente como un índice compuesto temperatura-humedad, reduciendo la inestabilidad de coeficientes en los modelos paramétricos.

5.2. ANTECEDENTES

- **Medición de la eficiencia técnica del maíz amarillo duro en los distritos de Chongoyape, Lagunas-Mocupe y Nueva Arica: un análisis con frontera estocástica**

El estudio desarrollado por Moreno Ruiz y Mori Julca (2017) se centra en medir la eficiencia técnica del cultivo de maíz amarillo duro en tres distritos del Perú: Chongoyape, Lagunas-Mocupe y Nueva Arica. A través del uso de la metodología de frontera estocástica de producción, las autoras identifican las principales variables que explican la ineficiencia productiva, entre las que destacan la edad y el nivel educativo del agricultor. La investigación es de carácter descriptivo y utiliza datos primarios obtenidos mediante encuestas aplicadas a 106 agricultores. La eficiencia se evalúa considerando variables físicas de producción como hectáreas cultivadas, uso de fertilizantes, mano de obra y tecnología, así como aspectos socioeconómicos del productor. El estudio demuestra que, incluso sin introducir nuevas tecnologías, es posible mejorar el rendimiento si se optimiza el uso de los recursos disponibles [30].

En contraste, el presente proyecto se orienta hacia la estimación del rendimiento potencial del cultivo de maíz a nivel municipal en Colombia mediante técnicas de aprendizaje automático y estadístico, utilizando fuentes de datos abiertos y secundarios como EVA y los registros climáticos históricos del IDEAM. A diferencia del enfoque microeconómico y localizado del estudio de Moreno Ruiz y Mori Julca, este trabajo adopta una perspectiva macroterritorial y predictiva, que busca explotar el valor de los datos históricos integrados con herramientas de la ciencia de datos para anticipar y optimizar el manejo agrícola. Además, el proyecto actual no solo pretende identificar brechas de eficiencia, sino también brindar una herramienta analítica que permita apoyar la toma de decisiones estratégicas en un contexto de alta variabilidad climática, apoyando así a una agricultura más resiliente e informada.

- **Comparison Between Machine Learning Models for Yield Forecast in Cocoa Crops in Santander, Colombia**

En el estudio de Lamos-Díaz, Puentes-Garzón y Zarate-Caicedo (2020) el aprendizaje automático (machine learning) ha emergido como una herramienta clave en la predicción del rendimiento agrícola, especialmente en cultivos complejos como el cacao (Theobroma cacao L.). En este contexto, diversos modelos como Redes Neuronales Artificiales (ANN), Máquinas de Vectores de Soporte (SVM), Árboles de Decisión (DT), Random Forest (RF), k-Vecinos más Cercanos (k-NN), Regresión Lineal, Gradient Boosting y regresión LASSO han sido empleados para identificar patrones y relaciones no lineales entre múltiples factores agroclimáticos, como la radiación fotosintética

activa, la temperatura, la humedad del suelo, la precipitación acumulada y el manejo agrícola. Estos modelos permiten capturar la interacción de variables complejas que no son fácilmente detectables con enfoques tradicionales, lo que los hace especialmente útiles en sistemas agrícolas tropicales con alta variabilidad. Los modelos supervisados, como las Redes Neuronales y las Máquinas de Vectores de Soporte, se destacan por su capacidad de aprender de datos históricos y predecir el rendimiento futuro del cultivo, basándose en diversos factores. Además, técnicas como la validación cruzada y la evaluación de modelos juegan un papel fundamental en la mejora de la precisión y generalización de las predicciones. En particular, la segmentación por condiciones de exposición solar (cultivo a sol o a sombra) resalta la importancia del manejo diferencial en los sistemas agroforestales, sugiriendo enfoques para una agricultura más precisa y sostenible sin comprometer la productividad [31].

El trabajo de Lamos-Díaz et al. (2020) aporta una base metodológica sólida para el uso de técnicas de aprendizaje automático en la predicción del rendimiento de cultivos, subrayando el valor de integrar variables climáticas y de manejo agrícola en modelos predictivos. No obstante, el presente proyecto se diferencia tanto en el enfoque como en el cultivo de estudio. Mientras el artículo mencionado se centra en el cacao bajo condiciones controladas y en un solo sitio experimental, el presente trabajo aborda el rendimiento potencial del cultivo de maíz de importancia crítica para la seguridad alimentaria global a partir de datos espaciales y climáticos multifuente, integrando información de las Evaluaciones Agropecuarias Municipales (EVA) y registros del IDEAM. Además, se pretende incorporar una capa de recomendación productiva específica por municipio, lo cual representa un avance respecto al enfoque más general de predicción de rendimiento del artículo base. Así, el presente estudio no solo busca predecir, sino también optimizar la toma de decisiones agrícolas, con un enfoque territorial más amplio y orientado a cultivos priorizados en Colombia

- **Influencia de la temperatura, precipitación y radiación solar en el rendimiento de maíz en el valle de toluca, méxico**

El estudio de Morales-Ruiz y Díaz-López (2020) analiza la influencia de variables climáticas temperatura, precipitación y radiación solar sobre el rendimiento de seis cultivares de maíz sembrados en tres ciclos agrícolas (2008, 2009 y 2010) en el Valle de Toluca, México. El trabajo se basa en datos experimentales recolectados en campo, bajo un diseño de parcelas divididas y análisis de varianza (ANDEVA), con el objetivo de identificar si las variaciones ambientales influyen significativamente en el rendimiento del cultivo. Los resultados muestran que la precipitación fue el factor determinante, influyendo significativamente en el rendimiento, mientras que la temperatura y la radiación no presentaron efectos estadísticamente significativos. El mayor rendimiento se alcanzó en el año con la mejor distribución de lluvias (2008), lo que resalta la importancia de una

sincronización entre los eventos pluviométricos y las fases fenológicas del cultivo, especialmente la floración y el llenado de grano. El estudio concluye que la precipitación heterogénea limita el rendimiento del maíz, subrayando la necesidad de seleccionar cultivares con mayor plasticidad fenotípica para afrontar esta variabilidad [32].

A diferencia del estudio realizado en Toluca, el presente proyecto se enfoca en desarrollar un modelo predictivo del rendimiento potencial del maíz utilizando técnicas de aprendizaje automático, integrando datos meteorológicos y agropecuarios históricos provenientes de fuentes como el IDEAM y las Evaluaciones Agropecuarias Municipales (EVA) de la UPRA. Mientras el trabajo de Morales-Ruiz y Díaz-López se basa en un experimento de campo específico en una localidad determinada y con datos limitados a tres años, el proyecto propuesto busca generalizar y escalar el análisis para cubrir diversas regiones y condiciones agroclimáticas, aprovechando técnicas modernas de modelado para la predicción y recomendación adaptativa del rendimiento. Esta diferencia metodológica permite generar herramientas de apoyo a la toma de decisiones agrícolas más robustas, dinámicas y aplicables a nivel territorial amplio, favoreciendo así a la sostenibilidad productiva en escenarios de cambio climático.

Crop Yield Time-Series Data Prediction Based on Multiple Hybrid Machine Learning Models

El estudio de Yan et al. (2025) investiga la predicción de rendimientos agrícolas utilizando un conjunto de datos multivariado y temporal (1990-2013) que incluye factores climáticos y el uso de pesticidas. Entre los modelos evaluados destaca el XGBoost, el cual es identificado como una versión avanzada de los modelos de gradiente que incorpora técnicas de regularización para evitar el sobreajuste y permite el procesamiento en paralelo, lo que resulta ideal para grandes volúmenes de datos. En sus pruebas, el modelo XGBoost alcanzó una precisión de 0,973, demostrando una capacidad sobresaliente para capturar relaciones no lineales complejas en el cultivo de maíz. El estudio concluye que estos modelos de ensamble son herramientas robustas para la planificación agrícola y la toma de decisiones dinámicas basadas en datos [33].

Mientras que Yan et al. se enfocan en una predicción general del rendimiento basada en promedios históricos en diversas regiones del mundo, el presente proyecto utiliza el XGBoost cuantílico específicamente para estimar la frontera superior o rendimiento potencial (τ=0,95) en el contexto colombiano. A diferencia del enfoque de Yan et al., que busca la precisión en el valor esperado, este trabajo prioriza la identificación del "techo" productivo municipal, integrando variables climáticas de alta resolución (como la temperatura de bulbo húmedo) para entender no solo cuánto se produce, sino cuánto es posible producir bajo condiciones óptimas en cada territorio.

Conformalized Quantile Regression

La investigación de Romano et al. (2019) aborda el problema de la cuantificación de la incertidumbre en modelos de regresión, proponiendo el método de Regresión Cuantílica (QR) conformalizada. Este enfoque combina la eficiencia de algoritmos como los Quantile Random Forests (QRF) y las redes neuronales cuantílicas con garantías estadísticas de cobertura en muestras finitas. El estudio enfatiza que la QR es fundamental cuando los datos presentan heterocedasticidad (variabilidad desigual), permitiendo que la longitud de los intervalos de predicción se adapte a la variabilidad local de los datos. Los autores demuestran que estimar funciones de cuantiles condicionales (como los niveles de 5% y 95%) proporciona una visión mucho más informativa que la simple estimación de la media condicional [34].

El presente proyecto traslada la teoría de la Regresión Cuantílica (τ=0,95) y los QRF del ámbito estrictamente estadístico de Romano et al. hacia una aplicación agronómica estratégica en Colombia. Mientras el estudio original se centra en la validez técnica de los intervalos de confianza, este trabajo utiliza dichas herramientas para operacionalizar el concepto de rendimiento potencial. Así, la capacidad de los modelos cuantílicos para manejar la heterocedasticidad se aprovecha aquí para modelar cómo el riesgo climático afecta de manera diferenciada el potencial productivo del maíz en los diversos pisos térmicos de los municipios colombianos.

6. METODOLOGÍA

6.1. Enfoque metodológico general

El proyecto se enmarcó en un enfoque cuantitativo, analítico y predictivo, orientado a la estimación del rendimiento potencial del cultivo de maíz a escala municipal en Colombia mediante técnicas de aprendizaje automático y modelado estocástico.

La metodología adoptada se basó en el modelo CRISP-DM (Cross Industry Standard Process for Data Mining), ampliamente reconocido por su aplicabilidad en proyectos de ciencia de datos y minería de información al proporcionar una metodología estandarizada y estructurada [35]. Además, se seleccionó por su capacidad para manejar datos multifuente y heterogéneos (clima, suelo, productividad histórica), característicos de estudios agrícolas a escala municipal; por su estructura iterativa, que facilitó ajustar y refinar modelos conforme se comprendieron mejor los patrones agroclimáticos; y por su orientación hacia la predicción robusta, que superó las limitaciones de enfoques deterministas clásicos incapaces de capturar relaciones no lineales, la variabilidad climática y la heterogeneidad espacial. Este enfoque proporcionó una estructura sistemática y flexible para el desarrollo de modelos predictivos, permitiendo abordar de manera ordenada las etapas de comprensión del problema, procesamiento de datos, modelado, evaluación y comunicación de resultados [36].

Dado el carácter agroclimático y espacial del estudio, el modelo CRISP-DM se complementó con los principios del ciclo de vida de aprendizaje automático (Machine Learning Lifecycle), que aseguraron la reproducibilidad y trazabilidad de los procesos [37], y con un enfoque estocástico que incorporó la estimación de la incertidumbre y la variabilidad de las predicciones [38].

Así, el marco metodológico final adoptado integró tres componentes: CRISP-DM, como estructura base para el proceso analítico; el modelado predictivo estocástico, para la incorporación de la incertidumbre en las estimaciones; y la ciencia de datos reproducible, para garantizar la transparencia y la trazabilidad en todas las fases.

6.2. Descripción de las fases metodológicas

6.2.1. Fase 1: Comprensión del problema agroproductivo

Esta fase correspondió al componente "Comprensión del contexto" del modelo CRISP-DM y tuvo por objetivo contextualizar el problema agroproductivo y definir con precisión el objetivo analítico del estudio. El propósito central del proyecto fue predecir el rendimiento potencial del cultivo de maíz a nivel municipal en Colombia mediante la integración y el análisis de series históricas agroproductivas y climatológicas.

Análisis contextual

Se examinó el contexto nacional del maíz, su importancia estratégica para la seguridad alimentaria y su relación con políticas sectoriales (por ejemplo, el programa “Maíz para Colombia”), así como su vinculación con los Objetivos de Desarrollo Sostenible, en particular el ODS 2 (Hambre Cero) y el ODS 13 (Acción por el Clima) [7]. Este diagnóstico justifica la necesidad de herramientas predictivas que mejoren la planificación territorial y reduzcan la incertidumbre en la toma de decisiones agrícolas.

Se examinó la literatura dirigida a:

- caracterizar la importancia del maíz en los sistemas productivos nacionales;
- identificar la demanda de herramientas analíticas para anticipar escenarios de rendimiento;
- evaluar el papel del aprendizaje automático y del modelado estocástico en la agricultura moderna;
- revisar experiencias recientes de modelación predictiva aplicadas a cultivos y contextos comparables.

De la revisión se derivaron evidencias que sustentaron la relevancia científica, técnica y socioeconómica del proyecto, y que pusieron de manifiesto vacíos concretos en la práctica actual (por ejemplo, integración limitada de EVA e IDEAM; cobertura espacio-temporal incompleto de estaciones climáticas) que el estudio debió abordar.

Se identificaron los principales actores que pudieron beneficiarse de los resultados: productores y asociaciones rurales; entidades públicas (UPRA, IDEAM, autoridades territoriales); centros de investigación y empresas del sector. Se estableció la expectativa de que los productos del proyecto (bases depuradas y modelos) se utilizaran para focalizar asistencia técnica, orientar políticas y priorizar inversiones territoriales.

Se reconocieron riesgos tempranos, como datos climáticos incompletos, heterogeneidad territorial y baja adopción tecnológica en zonas rurales. Para cada riesgo se plantearon mitigaciones preliminares (por ejemplo, solicitudes formales de datos a IDEAM).

El proyecto se basó en datos abiertos y públicos, por lo que no manejó información personal ni confidencial. Se priorizaron principios de transparencia, reproducibilidad y trazabilidad: todas las transformaciones se documentaron (scripts, versiones, diccionarios de datos).

6.2.2. Fase 2: Comprensión e integración de los datos

En esta etapa se realizó la recolección, la inspección inicial y el análisis exploratorio de los datos procedentes de fuentes abiertas: las Evaluaciones Agropecuarias Municipales (EVA) de la UPRA, con información sobre rendimiento, área sembrada y producción; y los registros climáticos históricos del IDEAM, que incluían temperatura del aire, precipitación, humedad relativa, velocidad y dirección del viento. Se verificó la cobertura espacio-temporal de los registros y la coherencia entre fuentes. A continuación, se ejecutó un análisis exploratorio inicial de datos (EDA), orientado a verificar la estructura, cobertura, consistencia y calidad de las fuentes disponibles. Este análisis incluyó la revisión de dimensiones de las tablas, tipos de datos, valores faltantes, duplicados, rangos de las variables, distribución general de los registros y coherencia entre las unidades espaciales y temporales. El propósito de esta fase fue asegurar que las fuentes EVA e IDEAM fueran integrables y que contaran con condiciones mínimas de calidad para avanzar hacia la preparación, selección de variables y modelado posterior.

La obtención de datos se desarrolló mediante un proceso sistemático de búsqueda, verificación, solicitudes formales y validación de cobertura espacio-temporal. Las actividades principales fueron:

1. **Búsqueda de datos agroproductivos en plataformas oficiales**
   Se consultaron los repositorios institucionales de la UPRA y la plataforma de Datos Abiertos del Gobierno de Colombia (datos.gov.co), identificando y descargando las series históricas del rendimiento de maíz a nivel municipal (EVA).
2. **Identificación de vacíos en la información climática pública**
   Se detectó que, aunque parte de la información del IDEAM estaba disponible públicamente, variables clave para el modelado (temperatura, precipitación, humedad relativa, velocidad y dirección del viento) presentaban cobertura incompleta, heterogénea o discontinuidades en los portales abiertos.
3. **Solicitudes formales al IDEAM (PQR) y documentación para trazabilidad**
   Se radicó una Petición, Queja o Reclamo (PQR) ante el IDEAM solicitando series diarias y otros datos faltantes para las estaciones con cobertura en los municipios priorizados. Todas

las PQR y sus respuestas se documentaron en un registro interno que incluyó: número de PQR, fecha de radicación, solicitante, alcance de la petición, fecha de entrega y responsable de recepción. Estos registros se archivaron junto con los metadatos recibidos (formato, cobertura, responsable de la serie) para garantizar trazabilidad.

4. **Recepción, descarga y consolidación de la carpeta entregada por el IDEAM**
   En respuesta a la PQR, el IDEAM compartió una carpeta con información climática organizada por estación y por variable. El equipo descargó dicha carpeta y procedió a unir los numerosos archivos por estación y por variable. El proceso consistió en consolidar las series individuales en repositorios por variable (por ejemplo: uno para temperatura, otro para precipitación), normalizando nombres de columnas, unidades y formatos, y georreferenciando las estaciones. Este trabajo de unión y estandarización generó tablas uniformes por variable, preparadas para el preprocesamiento descrito en la Fase 3.

El EDA se realizó sobre las tablas consolidadas por variable y el dataset EVA integrable. Para garantizar reproducibilidad se emplearon herramientas y librerías estándar en Python: pandas (manipulación y limpieza de datos), numpy (cálculos numéricos), matplotlib y seaborn (visualización), geopandas (georreferenciación y operaciones espaciales) y scikit-learn (técnicas preliminares de imputación y búsqueda de vecinos). Todos los notebooks y scripts del EDA se versionaron y documentaron en el repositorio del proyecto.

Los repositorios por variable, el registro de metadatos y los logs de integración constituyeron los insumos principales para la Fase 3 (Preparación y preprocesamiento de los datos), donde se detallaron los criterios de imputación, agregación temporal y validación de consistencia.

6.2.3. Fase 3: Preparación y preprocesamiento de los datos

Esta fase correspondió a la preparación de datos (Data Preparation) en el modelo CRISP-DM. Se unificaron los conjuntos mediante la estandarización de nombres de columnas, unidades de medida y escalas temporales.

Se eliminaron registros duplicados y se corrigieron inconsistencias en variables y formatos. Dado que los registros meteorológicos del IDEAM estaban originalmente a escala mensual y la información productiva de EVA se encontraba reportada a escala semestral, fue necesario armonizar ambas fuentes en una unidad temporal común municipio-semestre (periodo A y B). Para ello, las variables continuas como temperatura, humedad, velocidad y dirección del viento fueron agregadas mediante promedios semestrales, mientras que la precipitación, por su naturaleza acumulativa, fue sumada dentro del semestre correspondiente. Esta decisión permitió integrar de manera consistente las fuentes disponibles y construir una base compatible con la estructura

temporal de EVA. No obstante, se reconoce que esta agregación reduce la capacidad del modelo para capturar eventos climáticos de corta duración, extremos intraestacionales o variaciones coincidentes con etapas fenológicas específicas del maíz, como floración y llenado de grano. Por tanto, los resultados deben interpretarse como estimaciones basadas en condiciones climáticas semestrales promedio o acumuladas, no como una representación fenológica diaria o mensual del cultivo. El resultado fue un repositorio climático centralizado por variable, coherente y listo para integrarse con EVA.Se georreferenciaron las estaciones y se asociaron a códigos DIVIPOLA municipales (código departamento, código municipio y coordenadas). Cuando un municipio contó con más de una estación, se generó un promedio municipal por periodo para obtener un único registro por municipio/semestre.

Se aplicó ciencia de datos reproducible: todas las transformaciones se documentaron (scripts, notebooks), se elaboró un diccionario de datos y se versionaron los archivos de respaldo.

Imputación de datos faltantes (esquema y justificación)

Se cuantificaron los nulos por variable y periodo. La imputación combinó información espacio-temporal mediante un enfoque basado en Vecinos Cercanos (NearestNeighbors, scikit-learn) con ponderación inversa por distancia. Procedimiento:

- Transformación temporal: se codificó el semestre como número continuo (ej., 2020A → 2020.5; 2020B → 2020.6) para medir distancias temporales.
- Búsqueda de vecinos: se usaron distancias combinadas geográficas (latitud-longitud) y temporales para identificar estaciones válidas cercanas.
- Radio de vecindad e imputación espacial: se adoptó un radio máximo de 50 km para la búsqueda de estaciones vecinas con información válida, utilizando ponderación inversa por distancia. Esta decisión no se interpretó como una garantía de homogeneidad climática absoluta, sino como un compromiso metodológico entre proximidad espacial y recuperación efectiva de datos faltantes. Dada la heterogeneidad topográfica de Colombia, especialmente en zonas andinas donde pueden existir cambios abruptos de altitud y condiciones climáticas en distancias cortas, se reconoce que la distancia geográfica por sí sola no captura completamente la similitud climática entre estaciones. Por esta razón, la imputación fue entendida como una aproximación espacio-temporal razonable ante las restricciones de cobertura de estaciones disponibles, y no como una reconstrucción climática de altresolución.

  Se exploró la posibilidad de incorporar la altitud como criterio adicional de selección de vecinos, con el propósito de restringir la imputación a estaciones ubicadas en condiciones

topográficas más comparables. Sin embargo, al aplicar filtros simultáneos de distancia geográfica, coincidencia temporal y similitud altitudinal, el número de vecinos disponibles se redujo sustancialmente, generando una recuperación muy limitada de datos faltantes, especialmente en municipios con baja densidad de estaciones climáticas. Por tanto, se decidió mantener el esquema de imputación por proximidad espacio-temporal dentro de un radio de 50 km, documentando esta decisión como una limitación metodológica y proponiendo su fortalecimiento en trabajos futuros mediante modelos de interpolación que integren altitud, pisos térmicos, información satelital o superficies climáticas continuas.

Herramientas, librerías y entorno computacional empleado: se utilizaron pandas y NumPy para la manipulación de datos y cálculos numéricos; GeoPandas para operaciones espaciales y georreferenciación; scikit-learn para procesos de vecinos cercanos, particiones y métricas de evaluación; XGBoost para el entrenamiento del modelo cuantílico; Matplotlib y Seaborn para visualización; SHAP para interpretación del modelo; y Git/notebooks para el control de versiones y trazabilidad del flujo analítico. Con el fin de fortalecer la reproducibilidad de los resultados, se documentaron además las versiones de las principales librerías y componentes utilizados durante el desarrollo del proyecto.

Tabla 1 *Versiones de librerías y entorno computacional del proyecto*

Componente	Versión utilizada
Sistema operativo	Windows 10, versión 10.0.26200-SP0
Python	3.11.9
pandas	2.3.1
NumPy	2.4.3
scikit-learn	1.7.1
XGBoost	3.0.5
Matplotlib	3.10.5
Seaborn	0.13.2
SHAP	0.51.0
GeoPandas	1.1.1

Decisiones de validación temporal y justificación (k = 8 y rolling window)

- Se adoptó un split 80/20 para entrenamiento/prueba. Con datos desde 2006 hasta 2024 y agregación semestral, la división dejó 8 periodos para prueba. Por coherencia temporal y para evaluar desempeño sobre bloques temporales completos, se diseñó una validación que consideró k = 8 pliegues temporales (K-fold temporal), de modo que cada pliegue representara un bloque temporal de similar tamaño. Esta elección permitió evaluar el modelo en tantas particiones temporales como periodos disponibles para test, garantizando que el desempeño reportado reflejara la variabilidad interanual/semestral y evitando mezcla temporal entre entrenamiento y prueba.

- Se utilizó un esquema de rolling window porque preserva la dependencia temporal y emula la condición real de pronóstico (entrenar con datos pasados y predecir el futuro inmediato). Además, permitió:
    - Evaluar la estabilidad y degradación del modelo a lo largo del tiempo (sin asumir estacionariedad completa).
    - Simular despliegues reales en los que el modelo se reentrena periódicamente con datos más recientes.
    - Evitar la fuga de información que ocurre cuando se mezclan aleatoriamente muestras temporales en validación cruzada.

Ejemplo operativo: se entrenó con los semestres 2006.5–2014.5, se validó en 2015.5; luego se desplazó la ventana y se repitió hasta cubrir los pliegues definidos (k = 8). Esto proporcionó estimaciones realistas de desempeño y distribuyó los periodos de prueba a lo largo de la serie.

Adicionalmente, con el fin de fortalecer la reproducibilidad e interpretación de la evaluación fuera de muestra, se planteó un análisis complementario sobre el conjunto de prueba 2021–2024. Este consistió en comparar el desempeño del modelo entre semestres regulares y semestres asociados a condiciones climáticas atípicas, utilizando las predicciones ya generadas por el modelo final. Este procedimiento no modificó el entrenamiento ni los hiperparámetros del XGBoost cuantílico, sino que permitió evaluar la estabilidad de sus resultados frente a diferentes contextos climáticos observados durante el periodo de prueba.

En síntesis, mediante la estandarización temporal (mensual → semestral), la consolidación de múltiples archivos en un repositorio único, la incorporación de la codificación DIVIPOLA y la georreferenciación precisa de estaciones, se obtuvo una base de datos agroclimáticos integrada,

coherente y adaptada a la escala analítica requerida por EVA, constituyendo uno de los pilares fundamentales para el modelado predictivo posterior.

Tabla 2 *Diccionario de datos*

Nombre del campo	Descripción	Tipo de dato	Unidad / Formato
Llave	Identificador único del registro integrado entre EVA e IDEAM	Texto / Numérico	Código único
Código DANE departamento	Código oficial asignado por DANE al departamento	Numérico	2 dígitos
Departamento	Nombre del departamento	Texto	—
Código DANE municipio	Código oficial asignado por DANE al municipio	Numérico	5 dígitos
Municipio	Nombre del municipio	Texto	—
Grupo cultivo	Clasificación general del cultivo	Texto	—
Subgrupo	Subcategoría dentro del grupo de cultivo	Texto	—
Cultivo	Nombre del cultivo reportado	Texto	—
Desagregación cultivo	Nivel de detalle adicional del cultivo (variedad, tipo, etc.)	Texto	—
Año	Año del registro agrícola y meteorológico	Numérico	AAAA
Periodo_eva	Periodo de referencia del EVA (mensual, trimestral, etc.)	Texto	—
Área sembrada (ha)	Área sembrada del cultivo	Numérico	Hectáreas
Área cosechada (ha)	Área cosechada del cultivo	Numérico	Hectáreas
Producción (t)	Cantidad producida	Numérico	Toneladas
Rendimiento (t/ha)	Producción por unidad de área cosechada	Numérico	Toneladas/hectárea
Estado físico del cultivo	Condición fenológica o estado productivo del cultivo	Texto	—

Nombre del campo	Descripción	Tipo de dato	Unidad / Formato
Nombre científico del cultivo	Nombre botánico del cultivo	Texto	—
Valor_Humedad	Humedad relativa promedio mensual	Numérico	%
Valor_Precipitacion_Mensual	Precipitación total mensual	Numérico	mm
Valor_Direccion_Viento_Mensual	Dirección promedio del viento	Texto / Numérico	Grados (°)
Valor_Velocidad_Viento_Mensual	Velocidad promedio del viento	Numérico	m/s o km/h
Valor_Temperatura_Humeda_Mensual	Temperatura húmeda promedio mensual	Numérico	°C
Valor_Temperatura_Seca_Mensual	Temperatura seca promedio mensual	Numérico	°C
Longitud	Coordenada geográfica de longitud	Numérico (decimal)	Grados decimales
Latitud	Coordenada geográfica de latitud	Numérico (decimal)	Grados decimales

6.2.4. Fase 4: Modelados predictivos

El objetivo de esta fase fue desarrollar, comparar y documentar modelos capaces de estimar el rendimiento potencial del cultivo de maíz a nivel municipal, así como cuantificar la ineficiencia asociada a dichas estimaciones mediante intervalos y medidas probabilísticas que facilitaran la toma de decisiones territoriales.

La selección del cuantil $\tau = 0{,}95$ respondió al objetivo de estimar una frontera superior asociada al rendimiento potencial del cultivo de maíz, en lugar de modelar el comportamiento promedio de la producción. El cuantil 0,95 constituyó un referente robusto del rendimiento alcanzable, ya que representó el 5 % superior de la distribución condicional, correspondiente a escenarios productivos técnicamente exigentes, pero estadísticamente reproducibles. Este enfoque permitió aproximar el denominado “techo productivo” sin depender de valores máximos absolutos, los cuales pueden verse influenciados por observaciones atípicas, errores de medición o condiciones excepcionales

difíciles de replicar en condiciones reales de producción.

Primero, se implementaron modelos de Frontera Estocástica de Producción (SFA), ampliamente usados en economía agrícola para medir eficiencia técnica. Se estimaron especificaciones Cobb–Douglas y Translog: una versión con la ineficiencia modelada en función de variables climáticas y una formulación general. Estos modelos permitieron separar el error aleatorio de la ineficiencia productiva y estimar una frontera productiva, esto es, el nivel máximo alcanzable condicionado a los insumos y a los scores de eficiencia técnica.

Segundo, se estimaron modelos de regresión cuantílica: una especificación estándar para el cuantil $\tau = 0{,}95$, una versión con efectos fijos municipales y una versión bayesiana. Estas metodologías resultaron apropiadas cuando el interés se centró en la cola superior de la distribución condicional, equivalente aquí a aproximar niveles altos de rendimiento observables bajo condiciones agroclimáticas favorables.

Tercero, se incorporaron modelos de aprendizaje automático cuantílico, en particular Quantile Random Forest (QRF) y XGBoost cuantílico. Estas técnicas permitieron capturar relaciones no lineales, interacciones complejas y la heterogeneidad espacio-temporal en los sistemas productivos, además de estimar cuantiles condicionales y construir intervalos de predicción.

Cuarto, se realizó un análisis de asociación entre variables mediante matrices de correlación, con el propósito de apoyar la selección y fundamentación de las variables explicativas. Este análisis permitió identificar relaciones fuertes, redundancias y posibles riesgos de multicolinealidad entre variables agroproductivas y climáticas. En particular, se utilizó el coeficiente de correlación de Spearman debido a la presencia de relaciones potencialmente no lineales, distribuciones no necesariamente normales y heterogeneidad territorial en el panel municipio-semestre. Los resultados de este análisis orientaron decisiones como conservar el área sembrada frente a variables productivas altamente redundantes, excluir variables directamente derivadas del rendimiento para evitar fuga de información y seleccionar la temperatura húmeda frente a la temperatura seca por su mayor valor agroclimático y alta correlación entre ambas.

Continuando, la variable respuesta fue el rendimiento (Y), trabajada en escala logarítmica para estabilizar la varianza y transformar relaciones multiplicativas en aditivas. Las variables predictoras (X) se mantuvieron en su escala original; las variables empleadas para entrenar los modelos fueron: Valor_Temperatura_Humeda_Mensual, Valor_Precipitacion_Mensual, Valor_Humedad, Valor_Velocidad_Viento_Mensual, Valor_Direccion_Viento_Mensual y la variable agroproductiva de EVA (área sembrada).

Se implementó un esquema de validación cruzada temporal tipo rolling window con k = 8 periodos,

combinando esta validación temporal con la partición cronológica 80/20. En la práctica, los últimos ocho semestres de la base de datos se reservaron como conjunto de prueba (test) y el bloque histórico previo se utilizó para entrenamiento y para generar los pliegues temporales de validación (rolling windows) durante la selección de hiperparámetros y la evaluación interna.

Para cada especificación se obtuvieron y documentaron las siguientes salidas, que se emplearon en la interpretación y comparación de modelos:

- Summary del modelo: coeficientes estimados, errores estándar y p-values, que permitieron evaluar identificación y significancia de las variables.
- Elasticidades: calculadas en escala natural para las especificaciones interpretables (SFA y regresiones cuantílicas).
- Métricas de desempeño y de incertidumbre: MAE, RMSE, Pinball Loss, Interval Score y coverage del cuantil $\tau = 0{,}95$.

Estas métricas permitieron comparar y evaluar la precisión de las predicciones respecto a los valores reales y tomar decisiones sobre el modelo más adecuado.

6.2.5. Fase 5: Evaluación y validación del modelo

En esta fase se validó de forma sistemática la capacidad predictiva, la calibración cuantílica y la robustez espacio-temporal de las especificaciones consideradas, con el propósito de identificar modelos aptos para uso operativo, apoyo a la política pública y análisis interpretativo de brechas de eficiencia.

El marco de evaluación combinó métricas de predicción puntual, criterios orientados a colas e incertidumbre, así como indicadores compuestos y elementos de interpretación econométrica. Las métricas de predicción puntual incluyeron MAE (error medio absoluto) y RMSE (raíz del error cuadrático medio), utilizadas para evaluar la distancia promedio y la sensibilidad a errores extremos entre observaciones y predicciones. Para caracterizar la capacidad de los modelos de estimar fronteras superiores se emplearon la Pinball Loss y su versión normalizada, que midieron la pérdida asociada a la estimación de un cuantil específico y permitieron comparar desempeño cuantílico entre metodologías heterogéneas. La evaluación de incertidumbre se basó en Interval Score y en la cobertura empírica del cuantil objetivo (por ejemplo, $\tau = 0{,}95$), indicadores que informaron sobre la calidad y la calibración de los intervalos predictivos producidos por cada especificación. Complementariamente, se calculó el error de calibración como la desviación entre la cobertura observada y la cobertura teórica, y se estimó un Pseudo R^2 para cuantificar la capacidad explicativa

relativa de los modelos en el espacio del cuantil estimado.

Adicionalmente, la fase incluyó análisis de interpretación econométrica cuando la especificación lo permitió: estimación de elasticidades y, en modelos lineales o semilogarítmicos, interpretación de coeficientes y evaluación de significancia estadística. Para garantizar comparabilidad entre enfoques y estabilidad en la evaluación de intervalos y pérdidas cuantílicas, la mayoría de las métricas principales se calcularon sobre la variable objetivo-transformada en escala logarítmica base 10; específicamente: Interval Score (Log10), Coverage (Log10), Pinball Loss (Log10), Pinball Loss normalizado (Log10), MAE (Test, Log10), RMSE (Test, Log10) y la eficiencia técnica promedio en Log10. Las métricas de error de calibración y Pseudo R^2 se informaron en la escala natural de la variable para mantener su interpretabilidad práctica.

El proceso de selección y priorización de modelos combinó resultados cuantitativos (ranking por métricas individuales y score global compuesto) con criterios cualitativos de robustez fuera de muestra, estabilidad temporal y plausibilidad agronómica, priorizando especificaciones que mostraran equilibrio entre precisión, calibración y capacidad explicativa por encima de ventajas aisladas en una sola métrica

6.2.6. Fase 6: Interpretación, estimación del rendimiento potencial y comunicación de resultados.

Se documentó, de forma narrativa y reproducible, el procedimiento seguido para transformar las salidas del proceso de modelación en insumos útiles para la toma de decisiones territoriales. Primero se explicó el criterio de selección del modelo final derivado del ejercicio comparativo (balance entre precisión predictiva, calibración cuantílica, capacidad explicativa y robustez fuera de muestra) y se describió cómo se aplicó dicho modelo para generar predicciones condicionadas a escenarios agroclimáticos definidos por percentiles superiores de las variables de interés.

Seguidamente se especificó la metodología para cuantificar e informar la incertidumbre asociada a las estimaciones: generación y reporte de intervalos cuantílicos y bandas de incertidumbre derivadas del modelo seleccionado, y criterios formales para identificar y señalizar predicciones extrapolativas cuando las covariables del escenario cayeran fuera del rango observado en la muestra de entrenamiento.

La interpretación agronómica se presentó como un apartado metodológico que justificó el uso de herramientas de explicación del modelo para identificar determinantes asociados a la variación del cuantil estimado, y que estableció la manera en que esa información se tradujo a recomendaciones técnico-operativas. En este marco se precisaron las limitaciones de alcance de los análisis (variables

no incluidas, supuestos sobre replicabilidad de prácticas, riesgo de extrapolación temporal o espacial) y se propuso la integración de un protocolo de verificación empírica mediante muestreos de campo para validar y, en su caso, recalibrar las estimaciones.

Finalmente, se consignaron las condiciones de reproducibilidad del flujo de trabajo (código, versiones de paquetes, rutas de datos y metadatos) y las recomendaciones para un plan piloto de verificación en campo que permitiera contrastar las predicciones con evidencia empírica antes de la implementación a mayor escala.

El modelo entrenado se aplicó para simular el rendimiento teórico en dichas condiciones, proporcionando una estimación del potencial productivo máximo alcanzable. Los resultados se comunicaron mediante representaciones visuales (mapas, gráficos y tablas comparativas) y análisis interpretativos que facilitaron su uso por entidades del sector agropecuario, cooperativas y organismos de planificación territorial.

Estimación de brechas de rendimiento y pérdidas económicas potenciales

A partir de las predicciones del modelo seleccionado, se calculó la brecha de rendimiento municipal como la diferencia entre el rendimiento potencial estimado y el rendimiento efectivamente observado en cada periodo y territorio:

$$Gap_i = Y_{potencial,i} - Y_{observado,i} \quad (19)$$

Esta aproximación permitió dimensionar de manera preliminar los costos de oportunidad asociados a ineficiencias productivas y ofreció insumos cuantitativos para la priorización territorial de asistencia técnica, inversión pública y estrategias de adaptación agroclimática.

6.3. Consideraciones éticas y de limitación

El estudio se basa exclusivamente en datos abiertos y públicos, sin involucrar información personal o confidencial. Se garantizan principios de transparencia, trazabilidad y reproducibilidad científica. El alcance del proyecto se limita a la construcción, validación y análisis del modelo predictivo, sin incluir el desarrollo de una aplicación o plataforma funcional.

Una limitación adicional del estudio corresponde a la no incorporación explícita de indicadores climáticos extremos o de variabilidad interanual, como fases El Niño, La Niña o índices ENSO. Aunque parte de sus efectos puede estar reflejada indirectamente en las variables observadas de

temperatura, humedad y precipitación agregadas a escala semestral, el modelo no incluye una variable específica que identifique la fase, intensidad o duración de estos fenómenos. En consecuencia, la capacidad del modelo para capturar impactos climáticos extremos sobre etapas fenológicas específicas del maíz es limitada. Esta restricción debe considerarse al interpretar las predicciones de rendimiento potencial, especialmente en años con anomalías climáticas marcadas.

7. MODELOS

Se presentaron los modelos econométricos y de aprendizaje automático evaluados para estimar el rendimiento potencial del cultivo de maíz en Colombia. La selección metodológica respondió a la necesidad de comparar enfoques con diferentes fundamentos teóricos y capacidades predictivas, considerando la naturaleza del problema: datos de panel desbalanceado, relaciones no lineales entre variables agroclimáticas y productivas, heterogeneza territorial y la necesidad de estimar una frontera superior de productividad más que un valor promedio de rendimiento.

La comparación entre modelos no se limitó a una única métrica. Se evaluaron criterios de precisión predictiva, calibración cuantílica, capacidad explicativa y robustez general mediante indicadores como Interval Score, Coverage, error de calibración, Pinball Loss, MAE, RMSE, Pseudo R^2 y score global.

El objetivo final consistió en identificar el modelo con mejor equilibrio entre interpretación estadística, capacidad predictiva y utilidad práctica para estimar el rendimiento potencial del maíz a nivel territorial. Los resultados mostraron que XGBoost cuantílico fue seleccionado como modelo preferente por presentar el mejor balance frente a la jerarquía de criterios definida para el objetivo aplicado del estudio.

7.1. Análisis comparativo de desempeño

Interval Score (escala Log_{10})

En la **Tabla 3** de resultados se evidenció que el XGBoost cuantílico obtuvo el mejor desempeño en Interval Score, con un valor de 0,124733, inferior al registrado por los demás modelos comparados. Este resultado indicó que la metodología logró el mejor equilibrio entre dos componentes fundamentales: una frontera productiva suficientemente ajustada y una penalización reducida por observaciones fuera del nivel esperado.

En el contexto de esta investigación, dicho comportamiento fue especialmente relevante porque la estimación del rendimiento potencial requirió intervalos informativos y técnicamente plausibles, no fronteras excesivamente amplias que perdieran capacidad discriminante, ni límites demasiado estrechos que subestimaran el potencial real del cultivo. El valor alcanzado por XGBoost sugirió que el modelo construyó una frontera superior consistente y cercana al comportamiento observado en los mejores desempeños productivos.

Comparativamente, otros modelos presentaron intervalos menos eficientes, ya fuera por amplitudes mayores o por menor capacidad de cobertura. Esto redujo su utilidad práctica al buscar

identificar metas productivas realistas a nivel municipal o regional.

Tabla 3 *Resultados interval score para cada modelo*

Familia	Modelo	IS Log_{10}	Ranking
Machine Learning	XGBoost cuantílico	0,125	1
Regresión Cuantílica	QR Efectos Fijos	0,560	2
Frontera Estocástica	SFA General	0,573	3
Regresión Cuantílica	QR τ=0.95	0,968	4
Machine Learning	QRF	1,399	5
Frontera Estocástica	SFA Inef. Climática	2,463	6
Frontera Estocástica	SFA Translog	2,497	7
Frontera Estocástica	SFA Cobb–Douglas	4,480	8
Regresión Cuantílica	QR Bayesiana	4,639	9

Coverage

La **Tabla 4** comparativa evidencia diferencias importantes en la capacidad de calibración de los modelos. Considerando un cuantil objetivo de 0,95, los modelos con valores de Coverage más cercanos a ese nivel fueron QR Bayesiana, SFA clima y QR con efectos fijos, mientras otras metodologías presentaron subcobertura más marcada.

En el caso del XGBoost cuantílico, la cobertura reportada fue de 0,910827, valor relativamente próximo al objetivo teórico y consistente con una estimación robusta de frontera superior. Esto implica que cerca del 91% de las observaciones reales quedaron por debajo del rendimiento potencial estimado por el modelo, comportamiento coherente con una metodología orientada a capturar niveles altos de productividad y no valores promedio.

Aunque algunos modelos alcanzaron coberturas más cercanas al 95%, ello no necesariamente se tradujo en mejor desempeño global, ya que en varios casos dichos resultados estuvieron acompañados de mayores errores o menor capacidad explicativa. Por tanto, una cobertura alta aislada no garantiza superioridad metodológica.

Tabla 4 *Resultados coverage para cada modelo*

Familia	Modelo	Coverage Log_{10}	Desv. vs τ=0,95	Ranking
Regresión Cuantílica	QR Bayesiana	0,987	0,037	1
Frontera Estocástica	SFA Inef. Climática	0,934	-0,016	2
Regresión Cuantílica	QR Efectos Fijos	0,930	-0,021	3
Frontera Estocástica	SFA Translog	0,927	-0,023	4
Regresión Cuantílica	QR τ=0.95	0,918	-0,032	5
Machine Learning	XGBoost cuantílico	0,911	-0,039	6
Machine Learning	QRF	0,812	-0,138	7
Frontera Estocástica	SFA Cobb–Douglas	0,757	-0,193	8
Frontera Estocástica	SFA General	0,179	-0,771	9

Error de Calibración

La **Tabla 5** de resultados permite complementar la lectura anterior mediante el Error de Calibración, indicador que mide la diferencia entre la cobertura observada y la cobertura objetivo, es decir, $EC = Coverage - \tau$. En modelos orientados al cuantil 0,95, valores cercanos a cero reflejan mejor consistencia entre la frontera estimada y el nivel probabilístico esperado. El ranking del error de calibración se ordena según la desviación absoluta respecto al nivel nominal del cuantil objetivo. Los valores negativos indican subcalibración, es decir, una cobertura empírica inferior al nivel esperado; los valores positivos indican sobrecalibración, es decir, una cobertura empírica superior al nivel esperado.

Para el XGBoost cuantílico, el error de calibración fue reducido en magnitud, lo que confirma que la cobertura observada se mantuvo relativamente próxima al objetivo teórico. Esto sugiere que, aunque el modelo no reprodujo exactamente el 95% esperado, la desviación fue moderada y compatible con un buen desempeño aplicado.

En términos prácticos, una calibración adecuada es fundamental cuando se pretende interpretar la salida del modelo como rendimiento potencial alcanzable. Si la desviación fuese elevada, la frontera podría resultar demasiado conservadora o excesivamente optimista. En este caso, los resultados muestran que el XGBoost mantuvo una estimación equilibrada entre realismo productivo y exigencia técnica.

Tabla 5 *Resultados error de calibración para cada modelo*

Familia	Modelo	Error de calibración	Ranking
Regresión Cuantílica	QR τ=0.95	0,028	1
Regresión Cuantílica	QR Bayesiana	0,037	2
Machine Learning	XGBoost cuantílico	-0,039	3
Frontera Estocástica	SFA Translog	0,041	4
Frontera Estocástica	SFA Inef. Climática	0,043	5
Machine Learning	QRF	0,048	6
Regresión Cuantílica	QR Efectos Fijos	-0,057	7
Frontera Estocástica	SFA General	0,082	8
Frontera Estocástica	SFA Cobb–Douglas	0,193	9

Para el modelo XGBoost cuantílico, con Coverage = 0,910827 y $\tau = 0{,}95$, el error de calibración corresponde a -0,039173, lo que indica una subcobertura moderada de aproximadamente 3,9 puntos porcentuales.

Pinball Loss

Los resultados obtenidos muestran que el XGBoost cuantílico presentó el mejor desempeño en términos de pérdida cuantílica absoluta. El modelo registró una Pinball Loss de 0,009118, valor considerablemente inferior frente a los demás enfoques evaluados. Este resultado indica que las predicciones asociadas al cuantil objetivo presentaron menores desviaciones respecto a la frontera superior estimada del rendimiento.

El comportamiento observado sugiere que la arquitectura basada en boosting secuencial logró capturar de manera más eficiente las relaciones no lineales existentes entre variables climáticas, productivas y territoriales. En comparación con modelos estadísticos tradicionales, el XGBoost cuantílico mostró una capacidad superior para aproximar escenarios de alto rendimiento agrícola, reduciendo tanto errores de subestimación como desviaciones extremas en la estimación del cuantil τ=0,95.

De manera comparativa, modelos como la QR τ=0,95, QR Efectos Fijos y QRF presentaron desempeños competitivos, aunque con pérdidas cuantílicas superiores. Por otro lado, enfoques

como SFA General y QR Bayesiana evidenciaron mayores niveles de error, reflejando dificultades para aproximar adecuadamente la frontera potencial del rendimiento.

En términos aplicados, estos resultados indican que el XGBoost cuantílico fue el modelo con mejor capacidad para representar el techo productivo del cultivo de maíz, permitiendo una estimación más precisa del rendimiento potencial bajo condiciones heterogéneas de producción.

Tabla 6 *Resultados pinball loss para cada modelo*

Familia	Modelo	Pinball Loss Log_{10}	Ranking
Machine Learning	XGBoost cuantílico	0,009	1
Regresión Cuantílica	QR τ=0.95	0,029	2
Regresión Cuantílica	QR Efectos Fijos	0,036	3
Machine Learning	QRF	0,039	4
Regresión Cuantílica	QR Bayesiana	0,103	5
Frontera Estocástica	SFA Inef. Climática	0,142	6
Frontera Estocástica	SFA Translog	0,142	7
Frontera Estocástica	SFA Cobb–Douglas	0,266	8
Frontera Estocástica	SFA General	1,357	9

Pinball Normalizado

En la versión normalizada de la métrica, el XGBoost cuantílico mantuvo el mejor desempeño entre los modelos evaluados, registrando un valor de 0,008604. Este comportamiento confirma que la precisión observada en la pérdida cuantílica no dependió únicamente de la magnitud de la variable respuesta, sino que se mantuvo estable incluso después de ajustar las diferencias de escala entre observaciones.

La estabilidad del Pinball Loss normalizado adquiere especial relevancia considerando la estructura territorial del conjunto de datos, caracterizada por diferencias importantes entre municipios en términos de productividad agrícola, condiciones climáticas y tamaño del área sembrada. En este contexto, el modelo logró conservar un comportamiento robusto tanto en zonas de alto rendimiento como en territorios con menor capacidad productiva.

Los modelos QR Efectos Fijos y QR τ=0,95 también presentaron resultados relativamente favorables en esta métrica, indicando un desempeño competitivo en términos relativos. Sin embargo, modelos como la QR Bayesiana y SFA Cobb–Douglas mostraron pérdidas normalizadas considerablemente superiores, reflejando menor estabilidad predictiva frente a cambios de escala y heterogeneidad estructural de los datos.

En conjunto, los resultados del Pinball Loss normalizado evidencian que el XGBoost cuantílico no solo obtuvo alta precisión cuantílica absoluta, sino también una capacidad consistente de generalización bajo diferentes contextos productivos y territoriales.

Tabla 7 *Resultados pinball normalizado para cada modelo*

Familia	Modelo	Pinball normalizado Log_{10}	Ranking
Machine Learning	XGBoost cuantílico	0,009	1
Regresión Cuantílica	QR Efectos Fijos	0,034	2
Regresión Cuantílica	QR τ=0.95	0,106	3
Machine Learning	QRF	0,135	4
Frontera Estocástica	SFA General	0,202	5
Frontera Estocástica	SFA Inef. Climática	0,214	6
Frontera Estocástica	SFA Translog	0,215	7
Frontera Estocástica	SFA Cobb–Douglas	0,919	8
Regresión Cuantílica	QR Bayesiana	1,615	9

MAE (Mean Absolute Error)

Los resultados obtenidos en la métrica MAE evidencian que el XGBoost cuantílico presentó el menor nivel de error absoluto promedio entre los modelos evaluados, registrando un valor de 0,130497 en escala Log_{10}. Este comportamiento indica que, en términos generales, las predicciones generadas por el modelo se mantuvieron más próximas a los valores observados del rendimiento, mostrando una menor desviación promedio frente a las demás metodologías analizadas.

La diferencia observada respecto a modelos estadísticos tradicionales y enfoques cuantílicos convencionales sugiere que la arquitectura basada en boosting secuencial logró capturar con mayor

precisión las relaciones complejas entre variables climáticas, productivas y territoriales. En particular, la ventaja obtenida frente a modelos como QR Efectos Fijos, QR Bayesiana y los enfoques SFA evidencia una mayor capacidad de ajuste general sobre la estructura del sistema agrícola analizado.

El segundo mejor desempeño correspondió al QRF, con un MAE de 0,241984, aunque todavía distante del nivel alcanzado por el XGBoost cuantílico. Por otro lado, modelos como SFA General presentaron errores promedio considerablemente superiores, reflejando mayores diferencias entre las predicciones estimadas y el comportamiento observado del cultivo.

Desde una perspectiva aplicada, estos resultados indican que el XGBoost cuantílico logró representar de manera más cercana el comportamiento real del rendimiento agrícola, manteniendo precisión promedio incluso bajo condiciones heterogéneas entre municipios y periodos productivos.

Tabla 8 *Resultados MAE para cada modelo*

Familia	Modelo	MAE (Test) Log_{10}	Ranking
Machine Learning	XGBoost cuantílico	0,130	1
Machine Learning	QRF	0,242	2
Regresión Cuantílica	QR Efectos Fijos	0,492	3
Frontera Estocástica	SFA Translog	0,531	4
Frontera Estocástica	SFA Cobb–Douglas	0,532	5
Frontera Estocástica	SFA Inef. Climática	0,538	6
Regresión Cuantílica	QR τ=0.95	0,563	7
Regresión Cuantílica	QR Bayesiana	0,578	8
Frontera Estocástica	SFA General	0,640	9

RMSE (Root Mean Squared Error)

En la evaluación mediante RMSE, el XGBoost cuantílico volvió a registrar el mejor desempeño del conjunto comparado, obteniendo un valor de 0,182354 en escala Log_{10}. Dado que esta métrica penaliza con mayor intensidad los errores de gran magnitud, el resultado obtenido sugiere que el modelo no concentró desviaciones extremas en municipios o periodos específicos, manteniendo estabilidad en la estimación de la frontera productiva.

La diferencia entre el XGBoost cuantílico y los demás modelos evaluados fue considerable. Mientras que el segundo mejor desempeño correspondió nuevamente al QRF con un RMSE de 0,310769, los modelos restantes presentaron errores notablemente superiores, especialmente en metodologías como la QR Bayesiana y SFA General. Este comportamiento indica que dichos enfoques fueron más sensibles a observaciones extremas o escenarios de alta variabilidad productiva.

En términos metodológicos, el bajo RMSE obtenido por el XGBoost cuantílico refleja una adecuada capacidad de generalización y control del sobreajuste, incluso bajo una estructura de datos caracterizada por heterogeneidad territorial y variabilidad climática. La estabilidad observada en esta métrica fortalece la evidencia de que el modelo logró aproximar la frontera de rendimiento potencial sin generar errores severos en determinados subconjuntos de datos.

En conjunto, los resultados del RMSE confirman que el XGBoost cuantílico no solo presentó alta precisión promedio, sino también una mayor robustez frente a desviaciones extremas en la estimación del rendimiento potencial del cultivo de maíz.

Tabla 9 *Resultados RMSE para cada modelo*

Familia	Modelo	RMSE (Test) Log_{10}	Ranking
Machine Learning	XGBoost cuantílico	0,182	1
Machine Learning	QRF	0,311	2
Regresión Cuantílica	QR τ=0.95	0,626	3
Frontera Estocástica	SFA Translog	0,640	4
Frontera Estocástica	SFA Inef. Climática	0,646	5
Frontera Estocástica	SFA Cobb–Douglas	0,661	6
Regresión Cuantílica	QR Efectos Fijos	0,728	7

Familia	Modelo	RMSE (Test) Log_{10}	Ranking
Regresión Cuantílica	QR Bayesiana	0,733	8
Frontera Estocástica	SFA General	0,797	9

Pseudo R^2

Los resultados del Pseudo R^2 evidenciaron diferencias importantes en la capacidad de ajuste de los modelos evaluados para representar el percentil 95 del rendimiento del cultivo de maíz. El XGBoost cuantílico presentó el mejor desempeño del conjunto comparado, alcanzando un valor de 0,96094.

Dado que valores más cercanos a 1 indican mayor capacidad de ajuste, el resultado obtenido evidencia que el modelo logró representar con alta precisión la estructura observada en los niveles superiores de productividad agrícola. En términos prácticos, esto significa que las predicciones generadas por el modelo conservaron una elevada correspondencia con el comportamiento observado de los rendimientos asociados al límite superior de la distribución.

Este comportamiento sugiere que las variables agroclimáticas, productivas y territoriales incorporadas en la estimación aportaron información relevante para explicar el rendimiento potencial del cultivo. Asimismo, los resultados indican que las relaciones entre dichas variables y la productividad agrícola no siguieron patrones estrictamente lineales, siendo capturadas con mayor eficiencia por la estructura flexible del XGBoost cuantílico.

Frente a los demás modelos evaluados, la diferencia fue considerable. El QR Efectos Fijos alcanzó un Pseudo R^2 de 0,55020, mostrando una capacidad de ajuste moderada, aunque distante del nivel obtenido por el XGBoost cuantílico. Por su parte, modelos como la QR Bayesiana y SFA General registraron valores negativos, evidenciando limitaciones importantes para representar adecuadamente la dinámica observada en los niveles altos de rendimiento.

Tabla 10 *Resultados Pseudo R^2 para cada modelo*

Familia	Modelo	Pseudo R^2	Ranking
Machine Learning	XGBoost cuantílico	0,961	1
Regresión Cuantílica	QR Efectos Fijos	0,550	2
Frontera Estocástica	SFA Cobb–Douglas	0,217	3
Frontera Estocástica	SFA Translog	0,051	4
Regresión Cuantílica	QR τ=0.95	0,026	5
Frontera Estocástica	SFA Inef. Climática	-0,057	6
Machine Learning	QRF	-0,175	7
Regresión Cuantílica	QR Bayesiana	-2,557	8
Frontera Estocástica	SFA General	-5,211	9

Ranking Global de Modelos

En la **Tabla 11** de consolidación general permitió sintetizar el comportamiento conjunto de todas las métricas evaluadas. En ella, el XGBoost cuantílico ocupó la primera posición con un score global de 0,851424, superando al resto de modelos considerados.

Tabla 11 *Desempeño de los modelos*

Ranking	Familia metodológica	Modelo	Fortaleza principal en la estimación del rendimiento potencial de maíz	Debilidad principal en la estimación del rendimiento potencial de maíz
1	Machine Learning	XGBoost cuantílico	Presentó el mejor desempeño global para aproximar el percentil 95 del rendimiento del cultivo de maíz. Logró la menor pérdida cuantílica, menores errores promedio y el mayor nivel de ajuste, permitiendo representar de forma robusta la frontera	Presentó una ligera subcalibración del cuantil superior frente al nivel nominal de 0,95, por lo que sus predicciones deben interpretarse como una frontera técnica de referencia y no como una garantía probabilística exacta.

Ranking	Familia metodológica	Modelo	Fortaleza principal en la estimación del rendimiento potencial de maíz	Debilidad principal en la estimación del rendimiento potencial de maíz
			superior de productividad bajo condiciones agroclimáticas heterogéneas.	
2	Regresión Cuantílica	QR Efectos Fijos	Mostró buen desempeño relativo en cobertura y estabilidad cuantílica, incorporando además heterogeneidad territorial mediante efectos asociados a municipios.	Su capacidad predictiva y precisión global fueron considerablemente inferiores frente al XGBoost cuantílico, especialmente en métricas de error absoluto y ajuste general.
3	Machine Learning	QRF	Capturó adecuadamente relaciones no lineales entre variables climáticas y productivas, manteniendo errores relativamente bajos en escenarios con variabilidad territorial.	Presentó menor capacidad explicativa y menor estabilidad en calibración frente al modelo seleccionado.
4	Regresión Cuantílica	QR τ=0,95	Logró una calibración cercana al cuantil objetivo, mostrando coherencia probabilística en la estimación del límite superior del rendimiento agrícola.	Su desempeño en precisión predictiva y ajuste global fue limitado frente a metodologías basadas en árboles de decisión.
5	Frontera Estocástica	SFA Translog	Permitió modelar relaciones funcionales más flexibles entre insumos productivos y rendimiento agrícola.	Mostró baja capacidad explicativa y mayores errores en la representación de la frontera potencial del cultivo de maíz.
6	Frontera Estocástica	SFA Ineficiencia Climática	Incorporó componentes de ineficiencia asociados a variables climáticas,	Presentó precisión limitada y menor capacidad de generalización bajo

Ranking	Familia metodológica	Modelo	Fortaleza principal en la estimación del rendimiento potencial de maíz	Debilidad principal en la estimación del rendimiento potencial de maíz
			mostrando cobertura relativamente cercana al objetivo teórico.	escenarios territoriales heterogéneos.
7	Frontera Estocástica	SFA Cobb–Douglas	Ofreció una interpretación econométrica sencilla de la relación entre variables productivas y rendimiento agrícola.	La rigidez de la estructura funcional limitó la representación de interacciones no lineales presentes en el sistema agrícola analizado.
8	Regresión Cuantílica	QR Bayesiana	Alcanzó una cobertura cercana al percentil objetivo, manteniendo coherencia probabilística en algunos escenarios.	Presentó bajos niveles de ajuste y elevados errores predictivos, dificultando la representación estable del rendimiento potencial.
9	Frontera Estocástica	SFA General	Constituyó una referencia econométrica tradicional para el análisis de frontera productiva agrícola.	Mostró el desempeño más bajo del conjunto evaluado, con limitada capacidad para aproximar la frontera superior del rendimiento potencial del cultivo de maíz.

Este resultado es particularmente relevante porque algunos enfoques mostraron fortalezas aisladas en métricas específicas; sin embargo, no lograron mantener simultáneamente niveles altos de precisión, calibración y capacidad explicativa. El XGBoost, por el contrario, presentó un desempeño equilibrado y consistente en todas las dimensiones críticas del análisis.

En términos metodológicos, ello respalda su selección como modelo final de la investigación. En términos sustantivos, implica contar con una herramienta analítica con mayor desempeño relativo dentro del conjunto de modelos evaluados para estimar el rendimiento potencial del cultivo de maíz, comparar municipios, identificar brechas productivas y orientar estrategias de mejora agrícola con base cuantitativo.

7.2. Análisis de Robustez y Validez Estadística

La robustez de un modelo orientado a estimar rendimiento potencial no debe evaluarse únicamente por su capacidad predictiva puntual, sino también por su comportamiento fuera de muestra, estabilidad temporal, calibración cuantílica y consistencia frente a contextos productivos heterogéneos. En esta investigación, el periodo de prueba 2021–2024 permitió evaluar el modelo sobre semestres posteriores al entrenamiento y bajo condiciones recientes del sistema productivo. De manera complementaria, se realizó una evaluación estratificada entre semestres regulares y semestres asociados a condiciones climáticas atípicas, con el propósito de identificar si el desempeño del modelo se mantenía estable o se deterioraba en dichos contextos. Esta revisión no constituye una modelación explícita de fenómenos ENSO, pero sí aporta evidencia adicional sobre la estabilidad predictiva del modelo frente a variaciones climáticas observadas dentro del periodo de prueba.

El XGBoost cuantílico presentó el mejor desempeño global, con Pseudo R^2 de 0,953, MAE de 0,130 y RMSE de 0,182 en test. Estos resultados indican alta capacidad de generalización temporal y fuerte precisión en la estimación de la frontera productiva. No obstante, la cobertura observada de 0,9108 frente al objetivo τ=0,95 presenta una ligera subcalibración del cuantil superior. En consecuencia, más que evidencia de sobreajuste clásico, los resultados muestran una priorización relativa de precisión predictiva sobre calibración exacta del intervalo cuantílico.

El Quantile Random Forest obtuvo un desempeño competitivo en algunas métricas de error frente a los modelos estadísticos tradicionales, aunque inferior al XGBoost cuantílico. En particular, presentó el segundo mejor MAE y RMSE entre los modelos evaluados, pero mostró menor capacidad explicativa según el Pseudo R^2 y una cobertura más alejada del nivel nominal del cuantil 0,95. Por tanto, QRF se mantuvo como una alternativa metodológicamente relevante, aunque no superó al XGBoost en el balance global definido para este estudio.

El SFA General presentó resultados claramente inferiores, con Coverage de 0,179, MAE de 0,640 y Pseudo R^2 negativo, evidenciando subajuste severo. Esto indica que la especificación no logró capturar adecuadamente la dinámica productiva del cultivo ni representar una frontera técnicamente plausible.

Por su parte, la QR Bayesiana alcanzó cobertura elevada (0.987), pero con bajo ajuste y errores superiores. Este patrón sugiere intervalos excesivamente amplios: el modelo logró contener observaciones reales, aunque sacrificando precisión y utilidad operativa para estimar rendimiento

potencial.

La selección del XGBoost cuantílico como modelo preferente respondió a una jerarquía de criterios definida de acuerdo con el objetivo central del estudio: estimar una frontera superior de rendimiento potencial y derivar brechas productivas municipales. En este sentido, se priorizaron, en primer lugar, las métricas directamente asociadas a la calidad de la estimación cuantílica y a la precisión fuera de muestra, como Pinball Loss, MAE, RMSE y Pseudo R^2; en segundo lugar, se consideró la estabilidad temporal del desempeño bajo validación tipo rolling window; y, en tercer lugar, se evaluó la calibración probabilística mediante Coverage y Error de Calibración. Bajo esta jerarquía, el XGBoost cuantílico fue seleccionado porque presentó el mejor desempeño global en las métricas más vinculadas con el propósito aplicado del proyecto: aproximar el rendimiento potencial y cuantificar brechas territoriales. Si bien otros modelos, como Quantile Random Forest, mostraron un comportamiento competitivo en calibración probabilística, sus errores de predicción fueron superiores y su capacidad para aproximar con precisión la frontera superior resultó menor. Por tanto, la elección del XGBoost no implica desconocer la importancia de la calibración, sino priorizar la precisión de la frontera productiva cuando el uso principal del modelo es la identificación de brechas y la priorización territorial.

7.3. Robustez del modelo frente a periodos climáticos atípicos

Como complemento a la evaluación temporal del modelo, se realizó un análisis estratificado del desempeño del XGBoost cuantílico durante el periodo de prueba 2021–2024. Este análisis tuvo como propósito verificar si el modelo mantenía una capacidad de generalización aceptable tanto en semestres regulares como en semestres asociados a condiciones climáticas atípicas. Para ello, se utilizaron las predicciones generadas sobre el conjunto de prueba, sin reentrenar el algoritmo ni modificar los hiperparámetros previamente seleccionados. Los semestres fueron clasificados en dos grupos: periodos con anomalía climática [39], correspondientes a 2021-B, 2022-A y 2022-B, y periodos regulares, correspondientes a 2021-A, 2023-A, 2023-B, 2024-A y 2024-B. Posteriormente, se calcularon las métricas MAE, RMSE, Pinball Loss, Coverage y Error de Calibración para cada grupo, reportando los resultados en escala natural para facilitar su interpretación en términos del rendimiento del cultivo.

Tabla 12 *Desempeño del modelo XGBoost cuantílico frente a periodos regulares y periodos con anomalía climática, 2021-2024*

Condición climática	Observaciones	MAE	RMSE	Pinball Loss	Coverage	Error de calibración
Periodo con anomalía climática	2.515	0,4327	0,5619	0,0242	0,9666	0,0166
Periodo regular	4.302	0,4474	0,5805	0,0247	0,9702	0,0202
Total prueba 2021–2024	6.817	0,442	0,5737	0,0246	0,9689	0,0189

Desde el punto de vista de calibración, la Coverage total fue de 0,9689 frente al cuantil objetivo de 0,95, lo que representa un error de calibración de 0,0189. Esto indica una ligera sobrecobertura del cuantil superior estimado. Al comparar por grupos, la Coverage fue de 0,9666 en periodos con anomalía climática y de 0,9702 en periodos regulares, con errores de calibración de 0,0166 y 0,0202, respectivamente. La cercanía entre estos valores sugiere que la calibración del modelo se mantuvo relativamente estable entre condiciones climáticas diferenciadas.

El análisis estratificado no muestra evidencia de una degradación significativa del desempeño del XGBoost cuantílico en los periodos asociados a anomalía climática dentro del conjunto de prueba. Por el contrario, las métricas de error, pérdida cuantílica y calibración fueron similares entre los periodos regulares y atípicos. No obstante, dado que el modelo no incorporó explícitamente indicadores ENSO ni variables climáticas fenológicas, estos resultados deben interpretarse como una evaluación complementaria de robustez predictiva y no como una demostración causal de generalización frente a eventos climáticos extremos. En futuras extensiones, se recomienda incorporar variables explícitas de variabilidad climática, tales como fases ENSO, anomalías de precipitación, anomalías de temperatura o indicadores construidos por etapa fenológica del cultivo.

7.4. Modelo XGBoost cuantílico

A partir de la comparación entre alternativas econométricas y de aprendizaje automático, el modelo que mostró mejor desempeño predictivo fue el **XGBoost cuantílico**, estimado para capturar el rendimiento potencial del cultivo en el cuantil superior de la distribución. Su formulación permitió modelar relaciones no lineales, interacciones de alto orden y heterogeneidad entre municipios, sin imponer una estructura funcional rígida. En este sentido, el modelo resultó especialmente útil para

una base de datos de panel desbalanceado, donde la disponibilidad de observaciones varía entre municipios y periodos.

7.4.1. Selección y fundamentación de variables explicativas

La selección de variables explicativas se realizó mediante una combinación de criterios agronómicos, estadísticos y metodológicos. Como etapa inicial, se efectuó un análisis de dependencia entre variables utilizando el coeficiente de correlación de Spearman.

La elección de la correlación de Spearman respondió a las características estructurales del conjunto de datos. Dado que la base consolidada presentó heterogeneidad territorial, comportamiento no lineal y posibles distribuciones no normales en varias variables agroclimáticas y productivas, se optó por una medida no paramétrica de asociación

El análisis de correlación permitió identificar variables altamente redundantes dentro del componente productivo. En particular, el área cosechada presentó una correlación de 0,933 con el área sembrada, mientras que la producción total mostró correlaciones de 0,870 con área sembrada y 0,940 con área cosechada. Estos resultados evidenciaron una fuerte dependencia estructural entre dichas variables, indicando que describían dimensiones similares del sistema productivo. Por esta razón, se decidió conservar únicamente Area_sembrada_ha como variable representativa de escala agrícola, evitando incorporar información duplicada dentro del modelo y reduciendo posibles problemas de redundancia predictiva.

La variable Rendimiento_t_ha fue definida como variable objetivo del estudio, dado que representa directamente la productividad agrícola expresada en toneladas por hectárea. En consecuencia, esta variable no fue incorporada como predictor, sino como respuesta a estimar dentro de la estructura de aprendizaje supervisado.

En relación con las variables climáticas, la selección se realizó considerando tanto su relevancia agronómica como su comportamiento estadístico dentro de la matriz de correlación. La precipitación fue incluida por su relación directa con la disponibilidad hídrica del cultivo y el desarrollo fenológico del maíz. La humedad relativa y la temperatura húmeda fueron seleccionadas debido a su capacidad para representar condiciones termo-higrométricas asociadas al estrés fisiológico, evapotranspiración y balance hídrico del sistema agrícola.

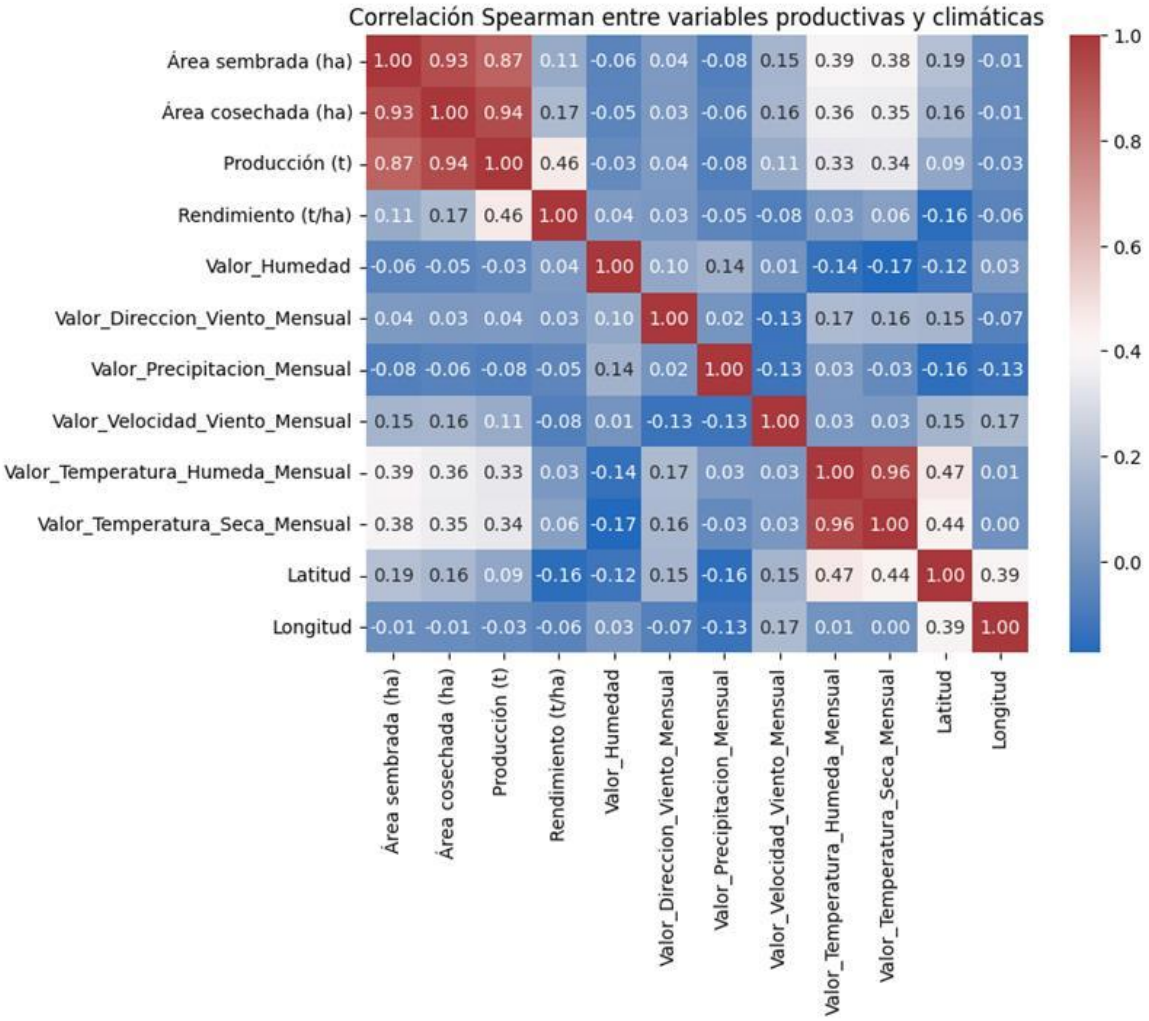
Particularmente, la temperatura húmeda mostró una correlación extremadamente alta con la temperatura seca (0,958), lo cual evidenció una elevada redundancia informacional entre ambas

variables. Ante esta situación, se optó por conservar únicamente Temp_Humeda dentro del modelo, debido a que esta variable integra simultáneamente componentes térmicos y de humedad atmosférica, proporcionando una representación más completa de las condiciones ambientales que afectan el rendimiento potencial del cultivo.

Adicionalmente, se incorporaron Vel_Viento y Dir_Viento como variables complementarias asociadas a la dinámica atmosférica local. En este tipo de arquitecturas, variables con correlaciones marginales bajas pueden aportar capacidad predictiva a través de interacciones no lineales y efectos combinados con otros factores agroclimáticos.

Finalmente, las variables espaciales de latitud y longitud no fueron incluidas explícitamente dentro del conjunto final de predictores. En su lugar, la heterogeneidad territorial fue incorporada mediante codificación binaria del identificador municipal (one-hot encoding), permitiendo capturar diferencias estructurales persistentes entre municipios relacionadas con condiciones geográficas, tecnológicas, agroecológicas y productivas no observadas directamente en las demás variables del modelo.

Figura 1 *Correlación de variables*



7.4.2. Etapa de preprocesamiento

El enfoque se desarrolló sobre la variable objetivo transformada en escala logarítmica, con el fin de estabilizar la varianza y reducir la asimetría de la distribución del rendimiento. En términos generales, si y_i representa el rendimiento observado, la transformación puede expresarse como:

$$z_i = ln\ ln\ (y_i + 1)\ = (y_i + 1) \quad (20)$$

donde ln denota el logaritmo natural (base e). equivalente a la función computacional log1p. Esta elección estabiliza la varianza y reduce la asimetría de la distribución del rendimiento. La transformación inversa, que permite recuperar las predicciones en toneladas por hectárea, es:

$$y_i = e^{zi} - 1 \quad (21)$$

implementada como expm1(·). Las métricas reportadas posteriormente en escala logarítmica corresponden a esta misma base natural.

Esta lógica permite que el modelo aprenda en una escala más estable y, posteriormente, devuelva las predicciones a la escala original para facilitar la interpretación práctica de los resultados.

La base de datos se organizó como un panel municipio–semestre, conservando la unidad territorial y temporal como estructura principal de análisis. Antes del ajuste del modelo, se realizó una depuración de nombres de columnas, estandarización de variables y construcción de identificadores necesarios para el tratamiento temporal y espacial de la información.

El procesamiento incluyó la creación de la variable municipio_id, a partir del código DANE, y la construcción de una variable temporal numérica (periodo_num) que permitió ordenar cronológicamente los semestres. Posteriormente, se definió la variable binaria semestre_B, asignando el valor 1 al segundo semestre y 0 al primero, con el fin de capturar efectos estacionales dentro del modelo.

La variable objetivo fue el rendimiento del cultivo, expresado como:

$$Rendimiento_i = Rendimiento \left(\frac{t}{h_a}\right)_i \quad (22)$$

y transformado a escala logarítmica mediante:

$$z_i = log\ log\ (1 + Rendimiento_i) \quad (23)$$

En paralelo, se seleccionaron como variables explicativas el área sembrada, la precipitación, la velocidad y dirección del viento, la humedad, la temperatura húmeda y el semestre agrícola. Adicionalmente, el identificador municipal fue incorporado mediante codificación binaria (one-hot encoding), permitiendo incluir controles territoriales dentro del proceso de aprendizaje del modelo sin asumir relaciones ordinales entre municipios. Esta estrategia facilitó la incorporación de diferencias estructurales asociadas a condiciones agroclimáticas, productivas y geográficas propias de cada territorio. Aunque este procedimiento no corresponde a un modelo de efectos fijos en sentido econométrico estricto, sí permitió controlar heterogeneidad municipal dentro de la estructura predictiva del algoritmo.

La base resultante correspondió a un panel desbalanceado, conformado por el siguiente conjunto de datos:

$$N = 31.514\ observaciones$$

La base fue segmentada respetando el orden cronológico de los datos. La partición se realizó sobre los periodos (semestres) disponibles, asignando el 80% de los semestres al conjunto de entrenamiento y el 20% restante al de evaluación externa. Este criterio garantiza que ninguna observación futura contamina la etapa de ajuste del modelo. Como resultado de esta división temporal, el conjunto de entrenamiento quedó conformado por:

- **Entrenamiento:** 24.697 registros (80%)
- **Evaluación externa:** 6.817 registros (20%)

Correspondientes respectivamente a los periodos 2006–2020 (entrenamiento) y 2021–2024 (evaluación). Los conteos de registros son consecuencia de la densidad de observaciones municipales en cada semestre, no de una selección aleatoria por filas.

La partición respetó el orden cronológico de los datos, evitando mezclar información futura con observaciones históricas, conservando una porción suficiente de datos para el ajuste y, al mismo tiempo, reservar observaciones inéditas para medir la capacidad de generalización del modelo.

7.4.3. Arquitectura del modelo

El modelo seleccionado correspondió a una implementación de XGBoost cuantílico orientada a la estimación del rendimiento potencial agrícola. La arquitectura se construyó bajo un enfoque de aprendizaje supervisado basado en árboles de decisión secuenciales, donde cada iteración busca corregir el error residual generado por los árboles anteriores.

A diferencia de modelos lineales tradicionales, la estructura del algoritmo permitió capturar relaciones no lineales, interacciones complejas entre variables climáticas y productivas, así como patrones heterogéneos entre municipios y periodos. Esta característica resultó especialmente relevante considerando la naturaleza dinámica del sistema agrícola analizado y la presencia de comportamiento diferencial entre territorios.

El modelo fue entrenado utilizando variables asociadas a condiciones productivas y meteorológicas, incluyendo área sembrada, precipitación, humedad, temperatura húmeda, velocidad del viento y

dirección del viento. Adicionalmente, se incorporaron controles territoriales mediante codificación binaria del identificador municipal, permitiendo capturar diferencias estructurales persistentes entre municipios dentro del proceso de aprendizaje del algoritmo. Aunque este procedimiento no corresponde a un modelo de efectos fijos en sentido econométrico estricto, sí permitió controlar heterogeneidad territorial dentro de la estructura predictiva del modelo.

La variable objetivo fue transformada previamente a escala logarítmica con el propósito de estabilizar la varianza y reducir problemas de asimetría presentes en la distribución del rendimiento. Posteriormente, las predicciones obtenidas fueron retransformadas a escala natural para facilitar su interpretación en toneladas por hectárea.

El componente cuantílico del modelo se implementó mediante una función de pérdida asimétrica tipo pinball loss, orientada al cuantil $\tau = 0{,}95$. Esta configuración permitió penalizar de manera diferenciada los errores de subestimación y sobreestimación, favoreciendo la aproximación a la frontera superior de la distribución condicional del rendimiento. En consecuencia, el modelo priorizó escenarios de alto desempeño productivo en lugar de estimar únicamente el comportamiento promedio de la variable objetivo.

La calibración de la arquitectura se realizó mediante un proceso de optimización de hiperparámetros, en el cual se evaluaron múltiples combinaciones asociadas a la complejidad estructural del modelo, capacidad de aprendizaje y mecanismos de regularización. En particular, se exploraron distintas configuraciones para la profundidad máxima de los árboles (max_depth), la tasa de aprendizaje (eta), el peso mínimo requerido para generar nuevas particiones (min_child_weight), la proporción de observaciones utilizadas en cada iteración (subsample) y la proporción de variables seleccionadas por árbol (colsample_bytree).

Las combinaciones evaluadas fueron las siguientes:

- max_depth: [4, 6, 8]
- eta: [0,03, 0,05]
- min_child_weight: [1, 5]
- subsample: [0,7, 0,8]
- colsample_bytree: [0,7, 0,8]

Adicionalmente, una vez ejecutada la búsqueda en grilla, se reportó la configuración final seleccionada para el modelo XGBoost cuantílico. Esta configuración corresponde a la combinación

de hiperparámetros que obtuvo el menor valor de Pinball Loss durante la validación interna. El número de árboles no hizo parte de la grilla de búsqueda, sino que fue definido como parámetro de entrenamiento del ensamble: durante la etapa de comparación de configuraciones se emplearon 150 iteraciones de boosting para reducir el costo computacional, y posteriormente el modelo final fue reentrenado con 300 iteraciones de boosting, equivalentes a 300 árboles secuenciales.

La configuración final del modelo seleccionado fue la siguiente:

Tabla 13 *Hiperparámetros seleccionados modelo XGBoost*

Hiperparámetro	Valor final	Criterio de definición
num_boost_round	300	Definido como número final de árboles del ensamble, después de la selección de hiperparámetros.
max_depth	4	Seleccionado mediante búsqueda en grilla según menor Pinball Loss.
eta	0,05	Seleccionado mediante búsqueda en grilla según menor Pinball Loss.
min_child_weight	5	Seleccionado mediante búsqueda en grilla según menor Pinball Loss.
subsample	0,7	Seleccionado mediante búsqueda en grilla según menor Pinball Loss.
colsample_bytree	0,8	Seleccionado mediante búsqueda en grilla según menor Pinball Loss.
Cuantil objetivo	0,95	Definido por el propósito de estimar una frontera superior de rendimiento potencial.

Estos hiperparámetros permitieron controlar tres aspectos centrales del modelo: la complejidad de los árboles, la velocidad de aprendizaje y los mecanismos de regularización. La profundidad máxima (max_depth) definió el nivel de complejidad de las reglas aprendidas; la tasa de aprendizaje (eta) controló el aporte incremental de cada árbol; min_child_weight reguló la creación de nuevas particiones; y subsample junto con colsample_bytree introdujeron aleatoriedad en observaciones y variables para reducir el riesgo de sobreajuste.

Este procedimiento permitió identificar configuraciones con mejor balance entre capacidad predictiva y control del sobreajuste, favoreciendo modelos con adecuada generalización fuera de muestra.

7.4.4. Proceso de entrenamiento y validación interna

El entrenamiento se diseñó bajo un esquema de validación cruzada temporal tipo rolling window o walk-forward, respetando estrictamente el orden cronológico de la información. Este punto fue fundamental, ya que la naturaleza del problema no permite mezclar observaciones futuras con observaciones pasadas sin introducir fuga de información.

En términos operativos, para cada ventana temporal se definió un conjunto de entrenamiento compuesto por los periodos anteriores al semestre a predecir, y un conjunto de prueba correspondiente al semestre siguiente. En una notación simplificada:

$$D_{train}^{(t)} = \{1,2,\dots,t-1\}, \quad D_{test}^{(t)} = \{t\} \quad (24)$$

Dentro de cada ventana temporal se realizó una partición interna del conjunto de entrenamiento en dos subconjuntos: subentrenamiento y validación. El subconjunto de subentrenamiento se utilizó para ajustar cada combinación de hiperparámetros del XGBoost cuantílico, mientras que el subconjunto de validación permitió comparar dichas combinaciones mediante la métrica Pinball Loss. La combinación con menor pérdida cuantílica fue seleccionada como la mejor configuración de la ventana.

Este procedimiento permitió separar la etapa de selección de hiperparámetros de la evaluación externa del modelo. En otras palabras, los datos de prueba no fueron utilizados para escoger la configuración del algoritmo, sino únicamente para medir su desempeño fuera de muestra. Esta estrategia redujo el riesgo de sobreajuste y mantuvo la coherencia temporal del problema, dado que el modelo aprendió con información histórica y fue evaluado sobre periodos posteriores.

El entrenamiento interno se realizó con un número fijo de iteraciones de boosting, y posteriormente se reentrenó el modelo final con la mejor configuración encontrada. La lógica seguida fue la siguiente:

1. Se construyeron los conjuntos de entrenamiento y validación con información histórica.
2. Se evaluaron distintas combinaciones de hiperparámetros.
3. Se seleccionó la configuración con menor pinball loss en validación.
4. Se ajustó el modelo final en la ventana completa de entrenamiento.
5. Se evaluó el desempeño sobre el periodo de prueba siguiente.

En la implementación computacional, cada combinación de hiperparámetros fue evaluada inicialmente con 150 iteraciones de boosting, con el fin de hacer eficiente la búsqueda en grilla. Una

vez identificada la combinación con menor Pinball Loss, el modelo fue reentrenado sobre la ventana completa de entrenamiento utilizando 300 iteraciones de boosting. Esta decisión permitió otorgar mayor capacidad de aprendizaje al modelo final, manteniendo controlado el costo computacional y evitando una complejidad excesiva. Por tanto, el número de árboles del modelo final fue de 300, mientras que la selección de los hiperparámetros estructurales se realizó mediante búsqueda en grilla. Este esquema permitió obtener una estimación más realista del comportamiento del modelo en contextos futuros.

Una vez finalizado el proceso de validación temporal y seleccionados los hiperparámetros óptimos, se procedió al entrenamiento del modelo final (*M_final*) sobre el conjunto completo de datos disponibles (2006–2024), incluyendo las observaciones reservadas para evaluación. Este procedimiento es estándar en la práctica de aprendizaje automático: el modelo de evaluación (*M_eval*) permite estimar el comportamiento fuera de muestra con datos no utilizados durante el ajuste; el modelo final aprovecha toda la información disponible para maximizar la capacidad predictiva en el horizonte de proyección. Las métricas de desempeño reportadas en este capítulo corresponden exclusivamente al modelo de evaluación y constituyen la referencia objetiva del desempeño del algoritmo.

7.4.5. Evaluación del modelo

La etapa de prueba consistió en predecir el rendimiento potencial de los semestres no observados durante el entrenamiento. Para ello, el modelo produjo predicciones en escala logarítmica, las cuales luego se transformaron a la escala natural mediante la operación inversa.

Si la predicción en escala logarítmica es $\hat{z}_i$, la predicción en escala original queda dada por:

$$\hat{y}_i = b^{\hat{z}i} - 1 \quad (25)$$

o, en la implementación del script, por la función equivalente expm1 cuando se trabaja con logaritmo natural.

De esta manera, las salidas del modelo pudieron interpretarse directamente en términos de toneladas por hectárea, que es la unidad de interés práctico para el análisis productivo.

La evaluación del conjunto de prueba se realizó con métricas que combinan precisión, cobertura y capacidad de calibración del cuantil estimado. En particular, se midió:

- Interval Score

- Coverage
- Error de Calibración
- Pinball Loss y Pinball Normalizado
- MAE
- RMSE
- Pseudo R^2

Es importante precisar que, en el caso del modelo XGBoost cuantílico seleccionado, la estimación se realizó sobre un único cuantil objetivo, τ = 0,95, utilizado como aproximación a la frontera superior del rendimiento potencial. Por tanto, el modelo final no estimó simultáneamente un intervalo intercuantílico compuesto por un límite inferior y un límite superior. En consecuencia, no se presentó el problema de cruce o traslape entre cuantiles dentro de esta especificación, ya que únicamente se modeló el cuantil superior.

La métrica de cobertura reportada para el XGBoost cuantílico debe interpretarse como la proporción de observaciones cuyo rendimiento observado quedó por debajo de la frontera superior estimada, y no como la cobertura de un intervalo predictivo bilateral. En futuras extensiones del modelo, si se estiman simultáneamente cuantiles inferiores y superiores, sería necesario incorporar mecanismos de no cruce, tales como restricciones de monotonía entre cuantiles, reordenamiento posterior de cuantiles o modelos cuantílicos conjuntos que garanticen que el límite inferior no supere al límite superior.

Adicionalmente, cuando el análisis se reporta en escala logarítmica base 10, la evaluación se realiza sobre las variables transformadas:

$$z_i = (y_i + 1)\ ,\ \hat{z}_i = (\hat{y}_i + 1) \quad (26)$$

y la pérdida cuantílica se calcula sobre dicha escala transformada. Esto garantiza consistencia con la forma en que se reportan las métricas de validación del modelo

7.4.6. Análisis de resultados

La evaluación final del modelo se efectuó sobre el subconjunto de prueba reservado desde la etapa inicial del proceso, manteniendo la proporción de separación **80/20**, con **24.697 observaciones para entrenamiento** y **6.817 para evaluación**. La base analizada corresponde a un **panel desbalanceado**, lo que implica que no todos los municipios presentan observaciones en todos los periodos. Este tipo

de estructura refuerza la pertinencia de una validación temporal, ya que permite medir el comportamiento del modelo en escenarios más cercanos a la realidad operativa.

Esta validación externa tomando del conjunto de evaluación permitió medir el comportamiento predictivo del algoritmo sobre datos no utilizados durante el entrenamiento, condición necesaria para verificar capacidad de generalización y consistencia fuera de muestra.

Dado que el modelo fue estimado sobre una transformación logarítmica natural de la variable objetivo mediante la función log1p, las métricas fueron calculadas en escala logaritmo base 10. En términos prácticos, esto significa que los errores y medidas de ajuste se interpretan sobre una versión estabilizada del rendimiento, lo que reduce efectos de asimetría, valores extremos y heterocedasticidad. Las métricas reportadas en **escala natural**, por el contrario, se expresan directamente sobre la magnitud original del fenómeno analizado y permiten una lectura operativa más cercana a la realidad productiva.

Interval Score[1]

En escala logarítmica, el modelo presentó un valor de 0,124733 en el indicador utilizado para evaluar la distancia entre la frontera cuantílica estimada y los valores observados. Dado que el XGBoost cuantílico seleccionado fue entrenado únicamente para el cuantil superior $\tau = 0,95$, este resultado no debe interpretarse como la evaluación de un intervalo bilateral con límite inferior y superior, sino como una medida complementaria de la calidad de la frontera superior estimada. En este sentido, valores menores indican una mayor cercanía entre el cuantil estimado y la estructura observada del rendimiento, manteniendo la lógica de evaluación del modelo orientada a la estimación del rendimiento potencial.

Coverage

La Coverage obtenida por el modelo XGBoost cuantílico fue de 0,910827 frente a un cuantil objetivo de (\tau = 0,95). Esto significa que aproximadamente el 91,1 % de las observaciones reales quedaron por debajo de la frontera superior estimada. En consecuencia, el modelo presentó una subcalibración moderada, equivalente a una diferencia de -0,039173 respecto al nivel nominal esperado. En términos prácticos, esto indica que la frontera estimada por el modelo es ligeramente

[1] En el caso del XGBoost cuantílico, el indicador denominado Interval Score se reporta como métrica complementaria de distancia respecto al cuantil superior, pero no corresponde a un Interval Score bilateral estricto, dado que no se estimó un intervalo intercuantílico completo.

más exigente que el cuantil teórico 0,95, dado que contiene una proporción algo menor de observaciones que la esperada.

Esta subcalibración debe interpretarse como una limitación relevante para usos estrictamente probabilísticos del modelo, especialmente si el objetivo fuera construir intervalos de predicción con garantías de cobertura exacta. Sin embargo, en el contexto de esta investigación, el uso principal del modelo no fue generar intervalos probabilísticos bilaterales, sino estimar una frontera superior de rendimiento potencial para identificar brechas productivas. Desde esta perspectiva, la ligera subcobertura observada no invalida la utilidad del XGBoost cuantílico, pero sí exige interpretar sus resultados como un umbral técnico de referencia y no como una garantía probabilística perfecta.

Error de calibración

Esa diferencia entre la cobertura observada y la esperada se refleja en el **error de calibración**, que fue de **-0,039173**. El signo negativo indica que la cobertura observada quedó por debajo del 95% esperado, pero la magnitud es reducida, lo cual sugiere que la desviación es moderada y no compromete la utilidad del modelo para fines de estimación del rendimiento potencial. La versión absoluta del mismo indicador, **0,039173**, confirma que la distancia respecto al objetivo cuantílico fue acotada.

Pinball Loss

La Pinball Loss también puede interpretarse en términos productivos al relacionarla con la variable original del estudio, expresada en toneladas por hectárea (t/ha)**.** En este caso, el valor de 0,032076 en escala natural indica que la desviación promedio ponderada entre el rendimiento potencial estimado por el modelo y el rendimiento observado fue relativamente baja. Aunque esta métrica no se interpreta exactamente como un error promedio simple debido a que penaliza de forma asimétrica según el cuantil objetivo sí permite concluir que las diferencias entre la frontera estimada y los valores reales fueron pequeñas en términos de productividad agrícola.

Llevado al contexto del cultivo de maíz, este resultado sugiere que el modelo logró aproximar con precisión el nivel superior de rendimiento técnicamente alcanzable en cada municipio y semestre analizado. En otras palabras, la distancia entre el rendimiento potencial predicho y el comportamiento observado del sistema productivo fue reducida, lo que fortalece la confiabilidad del modelo para identificar brechas de eficiencia. Si el error cuantílico hubiese sido elevado, ello implicaría una frontera poco consistente o alejada de la realidad productiva; sin embargo, el valor obtenido muestra una estimación razonablemente cercana a las condiciones reales del cultivo.

De manera complementaria, en la **escala logarítmica base 10**, utilizada durante el entrenamiento del modelo, la Pinball Loss fue de **0,009118**. Este resultado refleja el error cuantílico interno del algoritmo sobre una variable previamente transformada para reducir asimetrías y estabilizar la varianza. El menor valor observado en esta escala es esperable, ya que la transformación logarítmica comprime la dispersión de los datos y facilita el aprendizaje estadístico. En consecuencia, este indicador confirma que el modelo logró capturar adecuadamente la estructura subyacente del rendimiento potencial antes de reexpresar las predicciones en toneladas por hectárea.

Pinball los normalizado

La **Pinball Loss normalizada en escala natural**, con valor de **0,014857**, indica que el error relativo del modelo equivale aproximadamente al 1,49% de la escala de referencia utilizada en la normalización. Esto confirma que la desviación no solo es baja en términos absolutos, sino también proporcionalmente pequeña respecto al rango de productividad observado. El modelo mantiene precisión al comparar territorios con distintos niveles de rendimiento, desde municipios de baja productividad hasta zonas con mayor tecnificación agrícola.

Por su parte, en la **escala logarítmica base 10**, la Pinball Loss normalizada fue de **0,008604**, valor que representa un error relativo aún menor dentro del espacio donde el modelo fue entrenado. Esto indica que, una vez controladas las diferencias de escala, el algoritmo alcanzó una elevada precisión en la estimación del cuantil objetivo y mostró estabilidad durante la fase de validación.

MAE

Aunque el **MAE** se denomina error absoluto medio, en este contexto no debe interpretarse como una medida asociada a la predicción del rendimiento promedio. En el presente estudio, el modelo XGBoost cuantílico fue diseñado para estimar un **cuantil superior de la distribución condicional del rendimiento**, es decir, una aproximación al **rendimiento potencial o frontera productiva** del cultivo de maíz. Por tanto, el MAE mide la distancia promedio entre la **frontera estimada** y los valores observados, no la distancia respecto a una media esperada.

En la **escala logarítmica base 10**, el MAE fue de **0,130497**, lo que indica que, dentro de la escala de entrenamiento, la separación promedio entre el cuantil superior estimado y el comportamiento real observado fue reducida. Esto sugiere que el modelo logró posicionar la frontera productiva de manera consistente frente a la información histórica, conservando cercanía estadística con los valores del sistema sin perder su naturaleza de límite superior.

En la **escala natural**, el MAE alcanzó **0,407429 toneladas por hectárea**. Este resultado significa que la distancia absoluta promedio entre el rendimiento potencial estimado y el rendimiento observado fue cercana a 0,41 t/ha. En términos productivos, esta brecha representa una diferencia moderada y razonable para un modelo orientado a identificar niveles máximos técnicamente alcanzables. No debe entenderse como "error promedio de predicción convencional", sino como la separación media entre la productividad observada y la frontera estimada bajo las condiciones agroclimáticas registradas.

Desde la perspectiva del problema de investigación, un MAE contenido indica que la frontera estimada no se ubicó excesivamente lejos de la realidad productiva. Si este valor fuera muy alto, sugeriría una frontera sobredimensionada y poco creíble; si fuera excesivamente bajo, podría indicar que el modelo está reproduciendo el promedio observado y no un verdadero rendimiento potencial. En consecuencia, el resultado obtenido respalda una frontera plausible y técnicamente consistente para el cultivo de maíz.

RMSE

El **RMSE** conserva la misma lógica conceptual anterior, con la diferencia de que penaliza más intensamente las desviaciones grandes entre la frontera estimada y los valores observados. Por ello, esta métrica permite evaluar si existen municipios o periodos donde el rendimiento potencial proyectado quedó excesivamente alejado del comportamiento real.

En la **escala logarítmica base 10**, el RMSE fue de **0,182354**, superior al MAE en la misma escala, como es esperable. Esto indica la presencia de algunos casos con brechas mayores entre la frontera estimada y el rendimiento observado, aunque sin magnitudes suficientes para comprometer la estabilidad global del modelo.

En la **escala natural**, el RMSE fue de **0,633542 toneladas por hectárea**. Esto significa que, al dar mayor peso a las discrepancias altas, el modelo presenta algunas observaciones donde la distancia entre rendimiento observado y potencial estimado fue superior al promedio general. Sin embargo, el nivel alcanzado sigue siendo moderado dentro de un sistema agrícola heterogéneo como el maíz, donde influyen factores como clima, calidad del suelo, acceso tecnológico, manejo agronómico y diferencias territoriales entre municipios.

En términos de frontera productiva, estas discrepancias no necesariamente constituyen fallas del modelo. En muchos casos pueden reflejar **brechas reales de eficiencia técnica**, donde ciertas

unidades productivas operan sustancialmente por debajo de su potencial. Por ello, un RMSE mayor al MAE es coherente con la existencia de municipios con rezagos productivos más marcados.

Pseudo R^2 y capacidad explicativa del modelo

El **Pseudo R^2** constituye una medida de ajuste ampliamente utilizada en modelos no lineales y en regresión cuantílica, donde el coeficiente de determinación clásico no resulta plenamente aplicable. En este estudio, su interpretación debe realizarse en función de la capacidad del modelo para reproducir la estructura del **rendimiento potencial del cultivo de maíz**, entendido como una frontera productiva superior condicionada por factores agroclimáticos, territoriales y productivos.

En la **escala logarítmica base 10**, el modelo alcanzó un **Pseudo R^2 de 0,952697**, resultado que puede considerarse altamente satisfactorio para un algoritmo cuantílico de naturaleza no paramétrica. Este valor indica que el modelo logró capturar una proporción muy elevada de la variabilidad sistemática presente en la variable transformada, comparado con una referencia base de menor capacidad predictiva. En términos metodológicos, ello evidencia que la transformación logarítmica permitió estabilizar la distribución del rendimiento y facilitar el aprendizaje del algoritmo, especialmente frente a datos heterogéneos y asimétricos, frecuentes en sistemas agrícolas reales.

Desde una perspectiva sustantiva, este resultado sugiere que el modelo identificó adecuadamente los patrones que explican por qué ciertos municipios o periodos alcanzan niveles superiores de productividad respecto a otros. Es decir, logró reconocer relaciones complejas entre superficie sembrada, precipitación, temperatura húmeda, humedad relativa, velocidad del viento y demás variables incluidas en el sistema analizado.

Al trasladar la evaluación a la **escala natural**, el **Pseudo R^2** fue de **0,960944**, ligeramente superior al valor observado en logaritmo base 10. Este comportamiento confirma que la fortaleza predictiva del modelo no depende únicamente de la transformación estadística aplicada durante el entrenamiento, sino que se conserva al expresar los resultados en la unidad real del problema, es decir, toneladas por hectárea. En consecuencia, el modelo no solo ajusta adecuadamente en el plano matemático, sino que también mantiene coherencia y precisión al retornar a una escala directamente interpretable para fines productivos.

En términos aplicados, un **Pseudo R^2 superior al 95% en ambas escalas** implica que el comportamiento estimado de la frontera productiva del maíz responde de manera consistente a la información disponible y que el componente no explicado por el modelo es relativamente reducido.

Esto fortalece su uso para identificar brechas de productividad, comparar territorios, proyectar escenarios y priorizar intervenciones técnicas orientadas al cierre de dichas brechas.

Eficiencia técnica

La **eficiencia técnica promedio**, calculada en escala natural, alcanzó un valor de **1,283573**. Este indicador resulta especialmente relevante, ya que permite contrastar el rendimiento efectivamente observado con la frontera productiva estimada por el modelo XGBoost cuantílico. A diferencia de los enfoques orientados al promedio, en este caso la referencia no es la producción media del sistema, sino el nivel superior de desempeño técnicamente alcanzable dadas determinadas condiciones agroclimáticas y productivas.

Esto indica que, en promedio, el rendimiento potencial estimado por el modelo XGBoost cuantílico superó en un 28,4% al rendimiento efectivamente observado en los municipios analizados. En términos productivos, esta brecha representa el margen de mejora técnicamente alcanzable bajo las condiciones agroclimáticas y productivas registradas, asumiendo que los insumos y condiciones de los productores más eficientes son replicables.

Score global

El **score global del modelo** fue de **0,851424 en escala logarítmica** y de **0,795268 en escala natural**. Este indicador resume el comportamiento conjunto del modelo a partir de métricas de precisión, calibración cuantílica, estabilidad del error y capacidad explicativa. En términos prácticos, cuanto más cercano se encuentre a 1, mejor es el desempeño integral del algoritmo.

El valor de **0,851424** en escala log10 indica que el modelo presentó un desempeño alto dentro del espacio estadístico donde fue entrenado. Esto significa que, después de transformar la variable rendimiento para reducir asimetrías y estabilizar su varianza, el algoritmo logró aprender de manera eficiente las relaciones entre las variables explicativas y el rendimiento potencial. En otras palabras, el modelo identificó correctamente los patrones no lineales asociados al comportamiento superior del cultivo de maíz. Este resultado confirma que la etapa de entrenamiento fue sólida y que la estructura interna del modelo es consistente desde el punto de vista predictivo.

El valor de **0,795268** en escala natural es especialmente relevante, porque corresponde al desempeño del modelo una vez las predicciones fueron retornadas a **toneladas por hectárea (t/ha)**, es decir, a la unidad real de análisis del problema productivo. Esto significa que el modelo conservó cerca del **79,5% de desempeño global efectivo** al enfrentarse nuevamente a la variabilidad real del

rendimiento agrícola. En términos aplicados, indica que la frontera productiva estimada en toneladas por hectárea mantiene un nivel alto de confiabilidad para representar el rendimiento potencial del maíz en los municipios analizados. Aunque no se interpreta como “79,5% de toneladas explicadas”, sí puede entenderse como un nivel robusto de consistencia global entre lo que el modelo estimó como potencial productivo y lo que muestran los datos reales del sistema agrícola.

Elasticidades

Las elasticidades se presentan como una aproximación interpretativa del comportamiento del modelo y deben entenderse como indicadores de sensibilidad predictiva, no como parámetros econométricos causales. A partir del modelo entrenado, se estimaron elasticidades aproximadas a partir del modelo XGBoost cuantílico. Para ello, se aplicaron perturbaciones marginales individuales del 1 % sobre cada variable de entrada, manteniendo constantes las demás condiciones del sistema. Este procedimiento permitió evaluar la variación relativa esperada en el rendimiento potencial estimado ante pequeños cambios en cada regresor, conservando la interpretación en escala natural.

Los resultados evidenciaron que **Temp_Humeda** presentó la mayor sensibilidad en valor absoluto, con una elasticidad de −0,023150. Esto sugiere que incrementos marginales en la temperatura húmeda se asociaron con reducciones relativas en el rendimiento potencial estimado por el modelo. En segundo lugar, **Area_sembrada_ha** registró una elasticidad de −0,010618, indicando una relación inversa de baja magnitud dentro de la estructura predictiva del algoritmo. Por otro lado, **Vel_Viento** presentó una elasticidad positiva de 0,009524, mientras que **Humedad** y **Dir_Viento** mostraron efectos positivos considerablemente menores. La variable **Precipitacion** registró una elasticidad negativa cercana a cero, reflejando una sensibilidad reducida dentro del rango observado de los datos.

Tabla 14 *Elasticidades de cada variable*

Variable	Elasticidad	Interpretación marginal
Temp_Humeda	−0,023150	Un aumento de 1% reduce aproximadamente 0,023% el rendimiento potencial
Area_sembrada_ha	−0,010618	Un aumento de 1% reduce aproximadamente 0,011% el rendimiento potencial
Vel_Viento	0,009524	Un aumento de 1% aumenta aproximadamente 0,010% el rendimiento potencial
Humedad	0,001929	Efecto positivo muy pequeño sobre el rendimiento potencial

Variable	Elasticidad	Interpretación marginal
Dir_Viento	0,000955	Relación positiva prácticamente neutra
Precipitacion	−0,000643	Relación negativa prácticamente neutra

Las elasticidades reportadas en la **Tabla 14** deben interpretarse como medidas de sensibilidad del modelo y no como efectos causales directos de las variables sobre el rendimiento del cultivo. En el caso de modelos no lineales como XGBoost, estas elasticidades resumen la dirección e intensidad promedio de la respuesta estimada por el algoritmo ante variaciones marginales de las variables explicativas, manteniendo la estructura aprendida por el modelo. Por tanto, los signos obtenidos reflejan patrones predictivos presentes en los datos y pueden estar influenciados por interacciones no lineales, correlaciones entre variables climáticas, agregación temporal semestral y heterogeneidad territorial.

En particular, la elasticidad negativa asociada a la precipitación no debe interpretarse como evidencia de que la lluvia reduce siempre el rendimiento del maíz. Desde una perspectiva agronómica, la precipitación puede tener efectos positivos o negativos dependiendo de su magnitud, distribución temporal, etapa fenológica del cultivo, condiciones del suelo, drenaje y manejo agronómico. Una relación negativa promedio puede indicar que, en los registros analizados, mayores niveles acumulados de precipitación estuvieron asociados con contextos de exceso hídrico, saturación de suelos, reducción de radiación, problemas fitosanitarios o pérdidas durante etapas sensibles del cultivo. En consecuencia, el signo negativo no representa una relación agronómica universal, sino una señal predictiva condicionada por la estructura multivariada de los datos utilizados. Este resultado debe interpretarse en conexión con la multicolinealidad discutida en la sección 5.1.8, dado que precipitación, humedad relativa y temperatura húmeda capturan dimensiones relacionadas del ambiente agroclimático. Por tanto, el signo negativo de la precipitación puede reflejar interacciones y redundancias informacionales entre variables climáticas, más que un efecto agronómico causal aislado.

En términos generales, las elasticidades obtenidas mostraron magnitudes relativamente pequeñas, lo cual es consistente con la naturaleza no lineal del modelo XGBoost y con la interacción simultánea entre múltiples variables agroclimáticas y productivas.

Importancia de variables

La importancia de variables fue evaluada mediante la métrica de ganancia (gain), la cual cuantifica la contribución promedio de cada variable a la reducción del error dentro de las divisiones realizadas por el algoritmo XGBoost. Bajo este enfoque, valores más altos indican una mayor capacidad de la variable para mejorar la precisión predictiva del modelo.

Los resultados evidenciaron que la variable **Area_sembrada_ha** presentó la mayor importancia relativa, alcanzando un valor de 11,760171. Este resultado indica que la dimensión productiva asociada al tamaño del área cultivada constituyó el principal criterio de partición utilizado por el modelo para aproximar la frontera superior del rendimiento potencial. En segundo lugar, **Temp_Humeda** registró una importancia de 3,151911, seguida por **Vel_Viento** con 1,513692 y **Precipitacion** con 1,110074. Por su parte, **Humedad** y **Dir_Viento** presentaron niveles de importancia menores dentro de la estructura global de predicción.

Tabla 15 *Importancia de cada variable*

Variable	Importancia por ganancia (Gain)	Participación relativa (%)	Nivel de importancia
Area_sembrada_ha	11,760171	60,8	Muy alta
Temp_Humeda	3,151911	16,3	Moderada
Vel_Viento	1,513692	7,8	Baja
Precipitacion	1,110074	5,7	Baja
Dir_Viento	0,945387	4,9	Muy baja
Humedad	0,818919	4,2	Muy baja

En términos agronómicos, estos resultados presentan que el modelo otorgó una mayor relevancia a variables relacionadas con la escala productiva y las condiciones microclimáticas asociadas a temperatura y dinámica atmosférica. Esto revela que la estimación del rendimiento potencial estuvo influenciada tanto por factores estructurales de producción como por condiciones ambientales que afectan el desarrollo fisiológico del cultivo de maíz.

Análisis de varianza explicada

Con el propósito de identificar la relevancia relativa de las variables dentro de la estructura predictiva del modelo, se realizó un análisis de contribución explicativa basado en valores SHAP (SHapley Additive exPlanations). Este enfoque permitió cuantificar el aporte promedio de cada variable sobre la predicción del rendimiento potencial.

Los resultados evidenciaron que la variable **Area_sembrada_ha** presentó la mayor contribución explicativa dentro del modelo, registrando un valor SHAP agregado de 0,00308325. Esta variable concentró aproximadamente el 99,5 % de la capacidad explicativa total estimada por el algoritmo, indicando una alta dependencia predictiva de la escala productiva del cultivo.

En segundo lugar, la variable **Temp_Humeda** presentó una contribución de 0,00001139, seguida por Humedad con 0,00000161. Las demás variables climáticas mostraron aportes considerablemente menores dentro de la estructura global de predicción.

Tabla 16 *Varianza de cada variable*

Variable	Contribución explicativa SHAP	Participación relativa (%)	Nivel de influencia
Area_sembrada_ha	0,00308325	99,5	Muy alta
Temp_Humeda	0,00001139	0,37	Baja
Humedad	0,00000161	0,05	Baja
Vel_Viento	0,000000934	0,03	Baja
Dir_Viento	0,0000009	0,03	Baja
Precipitacion	0,00000016	0,01	Muy baja

Desde una perspectiva agronómica, los resultados sugieren que la escala de producción y las condiciones termo-higrométricas tuvieron una mayor influencia sobre la estimación del rendimiento potencial que otras variables atmosféricas de menor estabilidad dentro del sistema agrícola analizado. La importancia de una variable no representa causalidad directa, sino capacidad de aporte predictivo dentro de las divisiones y reglas internas construidas por el algoritmo. En consecuencia, una mayor contribución explicativa indica una mayor influencia sobre la capacidad predictiva del modelo, pero no necesariamente una relación causal lineal con el rendimiento agrícola.

El análisis SHAP debe interpretarse considerando la estructura predictiva del modelo. La elevada contribución de Area_sembrada_ha (99,5%) no solo refleja el efecto asociado a la escala productiva, sino también la capacidad de esta variable para captar indirectamente condiciones agronómicas y territoriales no observadas de forma explícita en la base de datos, tales como calidad del suelo, manejo agronómico, fertilización, disponibilidad hídrica, tecnificación y prácticas productivas locales. Al tratarse de la principal variable agroproductiva incorporada en el modelo, su comportamiento termina sintetizando múltiples condiciones estructurales del sistema agrícola en interacción con la heterogeneidad municipal capturada mediante la codificación territorial. En este contexto, los valores SHAP representan capacidad explicativa dentro de la lógica predictiva del algoritmo y no relaciones causales directas. Aunque este comportamiento favoreció la precisión en la estimación de la frontera superior del rendimiento (τ=0,95), también sugiere que parte de la variabilidad climática quedó absorbida por componentes territoriales y productivos agregados, aspecto que podría profundizarse en futuras investigaciones mediante segmentación agroecológica o incorporación de variables edáficas y de manejo agronómico más específicas.

Criterio de selección del modelo final e interpretación aplicada del cuantil 0,95

La selección del modelo final se realizó considerando la finalidad aplicada del estudio: estimar una frontera superior de rendimiento potencial que permitiera identificar brechas productivas municipales. En consecuencia, el criterio de decisión no se orientó únicamente a escoger el modelo con mejor cobertura probabilística aislada, sino aquel que ofreciera el mejor equilibrio entre capacidad para representar la frontera superior, estabilidad fuera de muestra, precisión en la estimación del cuantil y utilidad territorial de los resultados.

Bajo esta lógica, el cuantil 0,95 fue interpretado como un umbral técnico de referencia, no como una predicción promedio ni como una garantía probabilística exacta para cada municipio. Su función dentro del estudio fue aproximar un nivel alto de rendimiento empíricamente alcanzable, útil para comparar el desempeño observado frente a una frontera productiva plausible. Por tanto, la confiabilidad del umbral no depende exclusivamente de que la cobertura empírica sea exactamente igual al 95 %, sino de que el modelo mantenga una frontera coherente, estable y suficientemente cercana al comportamiento observado en los niveles altos de rendimiento.

En este contexto, la subcalibración moderada observada en el XGBoost cuantílico no invalida su uso como modelo principal, pero sí condiciona la forma en que deben interpretarse sus resultados. La cobertura obtenida indica que la frontera estimada es ligeramente más exigente que el nivel nominal esperado; por ello, las brechas productivas derivadas del modelo deben leerse como

señales de oportunidad y priorización territorial, no como pérdidas determinísticas ni como metas obligatorias de corto plazo. Esta precisión es especialmente importante si los resultados se utilizan como insumo para programas de asistencia técnica, inversión pública o planificación agropecuaria.

La comparación con modelos alternativos permitió reconocer que algunos enfoques podían presentar ventajas puntuales en calibración, pero con menor capacidad para representar con precisión la frontera productiva o para discriminar diferencias territoriales relevantes. Por esta razón, se priorizó el XGBoost cuantílico como modelo preferente, manteniendo explícita la advertencia de que su salida corresponde a un umbral técnico de rendimiento potencial. En aplicaciones futuras orientadas a decisiones operativas o de política pública, este umbral debería complementarse con validación territorial, análisis de sensibilidad y estrategias de recalibración probabilística, tales como métodos conformales, calibración posterior de cuantiles o ensambles que integren la precisión del XGBoost con modelos de mejor comportamiento distribucional.

De esta manera, la elección del XGBoost cuantílico no desconoce la importancia de la calibración, sino que establece una jerarquía coherente con el objetivo del proyecto: primero, estimar una frontera productiva útil para el análisis de brechas; segundo, asegurar desempeño fuera de muestra y estabilidad temporal; y tercero, documentar explícitamente las limitaciones probabilísticas del umbral estimado para orientar una interpretación responsable de los resultados.

7.4.7. Interpretación territorial de brechas productivas e impacto económico del modelo XGBoost cuantílico

Una vez seleccionado el XGBoost cuantílico como modelo final, las predicciones del rendimiento potencial fueron integradas en un tablero analítico territorial con el fin de transformar la salida del modelo en indicadores útiles para la toma de decisiones. Dado que el horizonte de predicción corresponde al año 2025 y no se dispone aún de un rendimiento observado definitivo para todos los municipios en dicho periodo, se construyó un rendimiento observado de referencia a partir del promedio histórico municipal del rendimiento real. Para preservar la estructura temporal del cultivo, este promedio fue calculado por municipio y semestre.

La brecha productiva se definió como la diferencia entre el rendimiento potencial estimado por el modelo y el rendimiento observado de referencia. En los casos donde dicha diferencia fue negativa, la brecha económica se truncó en cero, con el propósito de evitar interpretar como pérdida aquellos municipios cuyo promedio histórico se ubicó por encima de la frontera estimada.

$$Brecha_{i,s} = \hat{Y}_{i,s}^{potencial} - Y_{i,s}^{ref} \quad (27)$$

$$Brecha\ positiva_{i,s} = \max(\text{Brecha}_{i,s}, 0) \quad (28)$$

$$Brecha\ porcentual_{i,s} = \frac{\text{Brecha positiva}_{i,s}}{Y_{i,s}^{ref}} \times 100 \quad (29)$$

Para cuantificar el impacto económico, la brecha positiva en toneladas por hectárea fue multiplicada por el área agrícola sembrada de referencia del municipio. Posteriormente, las toneladas no producidas fueron valoradas usando precios mayoristas SIPSA, expresados en pesos colombianos por tonelada. Por tanto, el indicador económico corresponde a un valor bruto de producción potencial no alcanzada, no a una pérdida neta contable:

$$Toneladas\ no\ producidas_{i,s} = Brecha\ positiva_{i,s} \times A_{i,s}^{ref} \quad (30)$$

$$Costo\ de\ oportunidad_{i,s} = Toneladas\ no\ producidas_{i,s} \times P_{i,s}^{SIPSA} \quad (31)$$

Los resultados muestran que la brecha productiva presenta una distribución territorial heterogénea. En algunos municipios, la diferencia entre el rendimiento potencial y el rendimiento observado de referencia es reducida, lo que indica cercanía a la frontera productiva estimada. En otros territorios, la brecha es amplia, indicando oportunidades de mejora asociadas a manejo agronómico, adopción tecnológica, asistencia técnica o condiciones agroclimáticas locales. Al valorar esta brecha con precios SIPSA, se identifican departamentos y municipios donde la oportunidad económica agregada es mayor, no necesariamente porque tengan la mayor brecha en t/ha, sino porque combinan brechas positivas con mayor área sembrada y precios de referencia relevantes.

Tabla 17 *Resumen nacional 2025*

Indicador 2025	Valor
Rendimiento potencial	3.989,8 t/ha
Rendimiento observado de referencia	2.447,8 t/ha
Brecha productiva	1541,9 t/ha
Brecha porcentual	76,17%
Área de referencia	217.276,1 ha
Toneladas no producidas estimadas	251.367,1 t

Indicador 2025	Valor
Precio SIPSA promedio usado	$ 2.060.015,8 COP/t
Costo de oportunidad bruto	$ 509.143.093.163,1 COP

Este análisis permite priorizar territorios para programas de extensión rural, inversión pública, estrategias de adaptación agroclimática y focalización de instrumentos financieros o aseguramiento agropecuario. No obstante, los valores económicos deben interpretarse como aproximaciones de costo de oportunidad bruto, dado que el modelo no incorpora costos de producción, calidad del grano, restricciones logísticas, disponibilidad de insumos, tipo de semilla ni prácticas específicas de manejo agronómico.

Tabla 18 *Ranking de departamentos con mayor brecha económica en 2025*

Departamento	Brecha t/ha	Rendimiento potencial	Rendimiento observado ref.	Brecha media %	Toneladas no producidas	Costo oportunidad COP
Boyacá	256,5	540,3	283,8	100,65 %	11.802,6	$ 25.797.255.827,8
Antioquia	205,1	493,1	288,0	86,22 %	20.422,5	$ 40.792.348.969,2
Cundinamarca	131,9	401,9	270,0	56,54 %	13.080,6	$ 28.168.838.983,9
Bolívar	106,2	251,7	145,5	79,32 %	51.855,0	$ 103.620.667.642,2
Santander	100,4	342,0	241,6	46,29 %	4.744,0	$ 9.465.302.686,2
Magdalena	88,2	173,1	85,0	111,13 %	36.524,5	$ 75.257.757.840,1
Nariño	86,9	231,7	144,9	72,24 %	4.732,4	$ 9.600.080.852,5
Chocó	82,4	140,3	57,9	173,52 %	8.484,7	$ 16.935.484.681,5
Norte De Santander	69,7	167,1	97,4	79,92 %	4.308,9	$ 8.607.620.440,0
Cauca	54,8	151,5	96,7	64,42 %	8.692,7	$ 16.931.718.601,7
Atlántico	53,0	110,0	57,1	107,52 %	10.606,4	$ 21.934.445.230,4
Cesar	39,7	103,1	63,4	72,77 %	15.568,6	$ 29.100.993.775,3
La Guajira	34,0	77,6	43,6	82,41 %	13.085,7	$ 22.706.391.577,6

Departamento	Brecha t/ha	Rendimiento potencial	Rendimiento observado ref.	Brecha media %	Toneladas no producidas	Costo oportunidad COP
Tolima	33,4	125,5	92,1	42,97 %	2.803,0	$ 6.046.324.570,4
Caquetá	29,8	65,3	35,6	97,05 %	8.613,7	$ 19.034.871.467,7
Putumayo	27,5	62,7	35,2	92,60 %	7.622,3	$ 14.783.927.500,3
Córdoba	25,1	78,3	53,2	52,92 %	7.589,4	$ 15.684.136.087,5
Sucre	22,8	121,9	99,1	25,55 %	4.885,7	$ 10.118.707.443,4
Arauca	17,2	37,5	20,3	89,08 %	11.827,6	$ 25.676.733.929,1
Caldas	15,5	78,0	62,5	26,67 %	266,0	$ 539.326.417,1
Quindío	10,5	54,0	43,5	26,86 %	164,5	$ 335.291.514,7
Casanare	10,3	40,5	30,2	39,76 %	578,8	$ 1.264.346.949,4
Valle Del Cauca	10,2	46,4	36,2	30,57 %	137,2	$ 278.570.874,4
Risaralda	6,9	20,0	13,1	57,59 %	168,6	$ 342.067.206,7
Guaviare	6,6	15,7	9,1	75,53 %	1.133,0	$ 2.475.247.471,6
Vaupés	4,1	9,9	5,8	76,62 %	160,9	$ 353.261.269,2
Amazonas	4,0	10,2	6,2	64,22 %	166,3	$ 361.905.628,2
Meta	3,1	12,7	9,6	35,95 %	1.201,7	$ 2.661.306.998,5
Vichada	2,6	15,2	12,6	21,87 %	93,2	$ 194.843.553,2
San Andrés Y Providencia	1,7	2,6	1,0	168,75 %	1,4	$ 2.866.221,8
Guainía	1,3	4,8	3,6	36,10 %	31,6	$ 70.450.951,8

Figura 2 *Evolución de rendimiento observado, potencial y brecha*

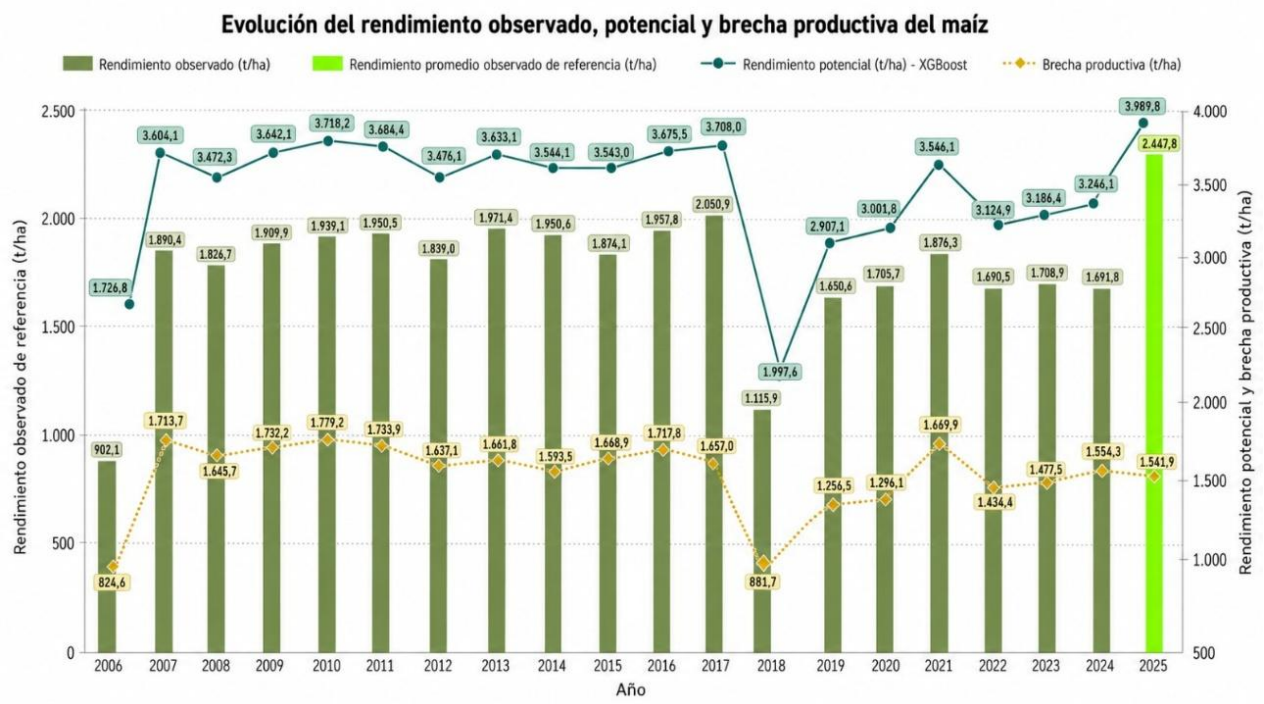
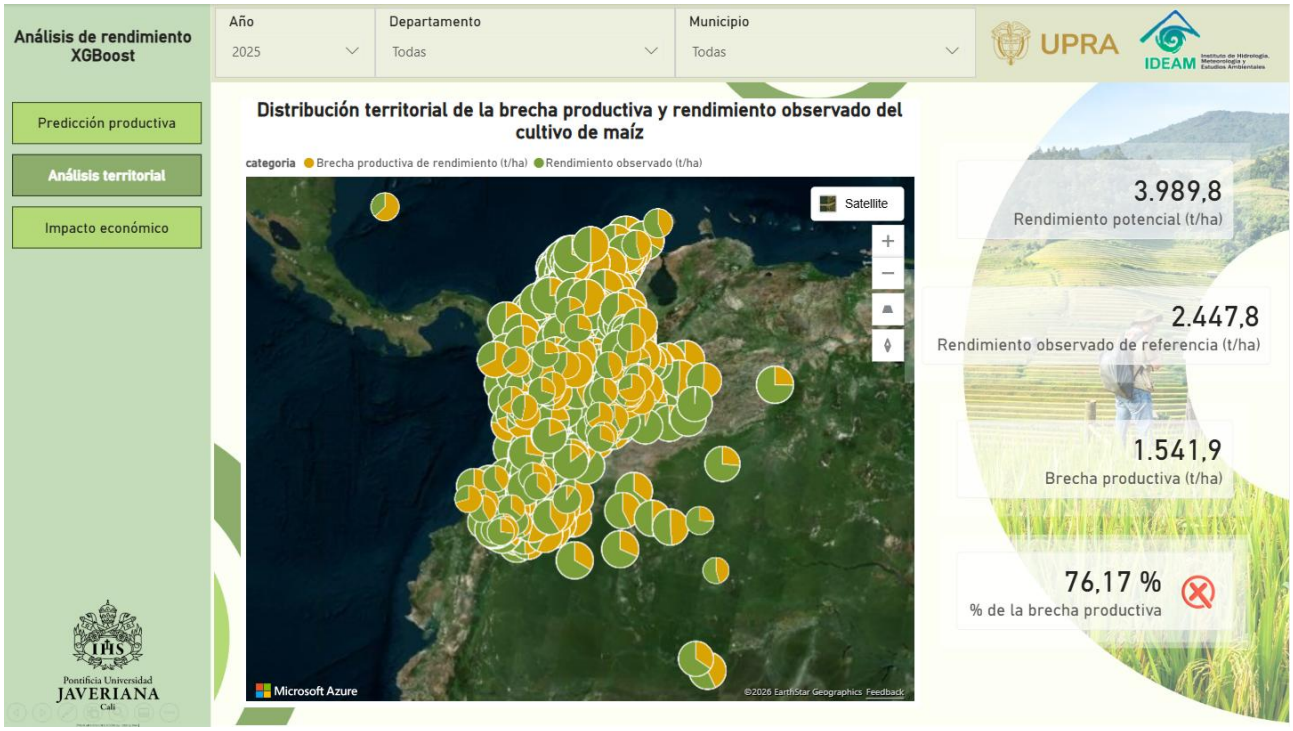


Figura 3 *Evolución del rendimiento observado, rendimiento potencial y brecha productiva del cultivo de maíz, 2006–2025 (PowerBi)*

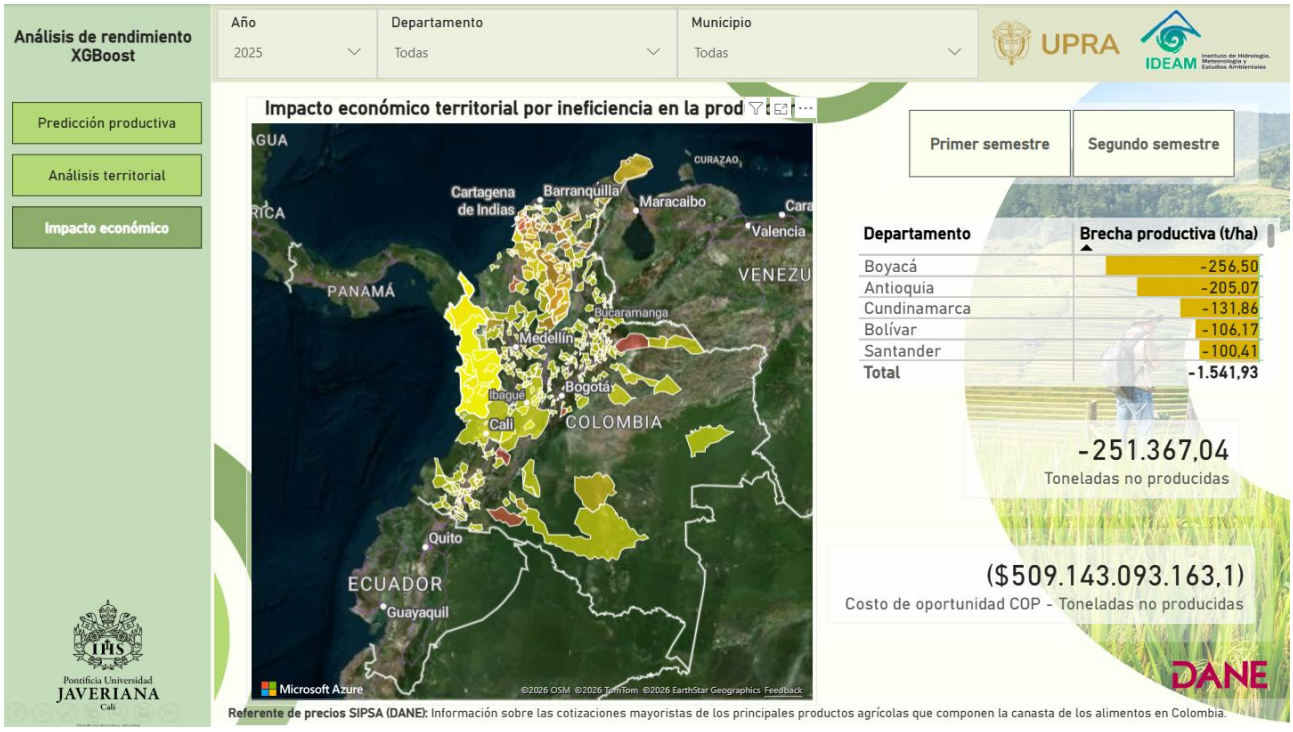
La **Tablero analítico para la comunicación territorial de resultados en Power BI**
Figura 3 permite comparar la trayectoria histórica del rendimiento observado con la frontera potencial estimada por el modelo XGBoost cuantílico. La distancia entre ambas series representa la brecha productiva, entendida como el margen de mejora entre el desempeño observado o de referencia y el rendimiento potencial estimado. Para 2025, el rendimiento observado corresponde a un valor de referencia construido a partir del promedio histórico municipal, dado que no se dispone de un rendimiento observado definitivo para dicho año.

Figura 4 *Mapa municipal de brechas de rendimiento del cultivo de maíz estimadas mediante XGBoost cuantílico, 2025 (PowerBi)*



La **Figura 4** evidencia la heterogeneidad espacial de la brecha productiva del maíz. Los municipios con mayor diferencia entre el rendimiento potencial estimado y el rendimiento observado de referencia representan territorios donde existe una mayor oportunidad de mejora productiva. Esta visualización permite pasar de una evaluación global del modelo a una lectura territorial útil para la priorización de asistencia técnica, inversión pública y estrategias de adaptación agroclimática.

Figura 5 *Estimación territorial de toneladas no producidas y costo de oportunidad por brecha productiva del maíz, 2025 (PowerBi)*



La **Figura 5** traduce la brecha productiva estimada en toneladas no producidas y en valor económico potencial, utilizando precios de referencia SIPSA - DANE. Este indicador debe interpretarse como un costo de oportunidad bruto asociado al rendimiento potencial no alcanzado, no como una pérdida neta contable, dado que no incorpora costos de producción, transporte, comercialización ni diferencias de calidad del grano.

8. CONCLUSIONES Y TRABAJOS FUTUROS

8.1. CONCLUSIONES.

El presente trabajo permitió desarrollar y validar un modelo predictivo para estimar el rendimiento potencial del cultivo de maíz a escala municipal en Colombia, integrando información agroproductiva y agroclimática proveniente de fuentes oficiales abiertas. La construcción de una base tipo panel municipio-semestre para el periodo 2006–2024 representó un aporte metodológico central, al transformar datos públicos dispersos en una estructura analítica trazable, reproducible y útil para el análisis del rendimiento agrícola.

El XGBoost cuantílico fue seleccionado como modelo preferente por presentar el mejor balance entre precisión predictiva, ajuste de la frontera superior de rendimiento y utilidad para la estimación de brechas productivas municipales. Esta selección no implica que el modelo sea superior en todas las dimensiones de evaluación, sino que responde a la jerarquía de criterios definida para el propósito aplicado del estudio. En particular, la calibración probabilística se reconoce como una dimensión relevante para futuras aplicaciones operativas o de política pública, por lo que los resultados deben interpretarse como una frontera técnica de referencia y no como una garantía probabilística absoluta. Su capacidad para capturar relaciones no lineales, interacciones entre variables y heterogeneidad territorial resultó especialmente adecuada para estimar una frontera superior de productividad mediante el cuantil 0,95, este enfoque permitió aproximar niveles de rendimiento técnicamente alcanzables, coherentes con el propósito de identificar brechas productivas.

Los resultados de evaluación evidenciaron un desempeño robusto del modelo en precisión, pérdida cuantílica, capacidad explicativa y generalización temporal. La validación tipo rolling window fortaleció la confiabilidad del proceso, al respetar la secuencia cronológica de los datos y evitar fuga de información. Asimismo, el análisis de variables mostró que el rendimiento potencial está asociado a factores productivos, climáticos y territoriales, destacando el área sembrada como una variable que puede capturar indirectamente diferencias estructurales entre municipios.

Uno de los principales aportes aplicados fue traducir la predicción del rendimiento potencial en una medida de brecha productiva, comparando la frontera estimada con un rendimiento observado de referencia. Esta aproximación permitió identificar territorios con mayor margen de mejora y convertir el modelo en una herramienta útil para orientar asistencia técnica, planificación agrícola y priorización territorial. De manera complementaria, la valoración económica con precios SIPSA permitió aproximar el costo de oportunidad bruto asociado a la producción potencial no alcanzada.

No obstante, los resultados deben interpretarse considerando las restricciones derivadas de la disponibilidad y resolución de la información climática. La agregación semestral de variables meteorológicas y la imputación basada en proximidad espacio-temporal permitieron construir una base integrada y funcional para el modelado, pero no sustituyen una caracterización agroclimática de alta resolución ajustada a la fenología del cultivo ni a la complejidad topográfica del territorio colombiano.

En síntesis, la investigación demuestra el potencial de la ciencia de datos para fortalecer la toma de decisiones en el sector agropecuario colombiano. La integración de datos abiertos, aprendizaje automático cuantílico, validación temporal e interpretación territorial permitió construir una metodología replicable para estimar rendimiento potencial y brechas productivas, aportando evidencia cuantitativa para una agricultura más informada, diferenciada y basada en datos.

8.2. TRABAJOS FUTUROS

Como línea futura, se recomienda incorporar variables edáficas y de manejo agronómico que no estuvieron disponibles en la base consolidada, como tipo de suelo, pH, materia orgánica, fertilización, variedad sembrada, densidad de siembra, disponibilidad de riego y nivel de tecnificación. Estas variables permitirían mejorar la interpretación del modelo y reducir la dependencia de variables agregadas, como el área sembrada o la codificación municipal, que actualmente pueden capturar de forma indirecta condiciones estructurales del sistema productivo.

También sería pertinente fortalecer el componente espacial del análisis mediante técnicas de autocorrelación espacial, análisis de residuos georreferenciados o modelos que incorporen relaciones de vecindad entre municipios. Esto permitiría evaluar si existen patrones espaciales no capturados por el modelo y mejorar la interpretación territorial de las brechas productivas.

Otra línea relevante consiste en incorporar explícitamente indicadores de variabilidad climática extrema, como índices ENSO, anomalías de temperatura, anomalías de precipitación o clasificaciones oficiales de años El Niño y La Niña. Esto permitiría evaluar de manera más directa cómo los eventos climáticos extremos afectan el rendimiento potencial y las brechas productivas del maíz. Asimismo, sería pertinente pasar de agregaciones semestrales generales a variables climáticas fenológicas, construidas según etapas críticas del cultivo, como emergencia, floración y llenado de grano. Esta mejora permitiría capturar con mayor precisión los efectos de déficits hídricos, excesos de precipitación o estrés térmico en momentos determinantes del ciclo productivo. Este tipo de análisis permitiría estimar cómo podrían variar el rendimiento potencial y

las brechas productivas bajo condiciones de estrés climático, aportando información útil para la gestión del riesgo agroclimático y la planificación territorial.

Adicionalmente, futuras investigaciones podrían incorporar información satelital y de sensores remotos, como índices de vegetación, humedad del suelo, anomalías térmicas o indicadores de estrés hídrico. Estas fuentes mejorarían la resolución espacial y temporal del modelo, especialmente en municipios con baja disponibilidad de estaciones climáticas o registros agroproductivos incompletos.

Como extensión metodológica, se recomienda incorporar estrategias de recalibración probabilística del modelo cuantílico, especialmente si sus resultados se orientan a usos operativos, normativos o de política pública. Aunque el XGBoost cuantílico presentó una calibración razonablemente cercana al nivel nominal del cuantil 0,95, futuras aplicaciones podrían fortalecer la confiabilidad probabilística del umbral estimado mediante calibración posterior de cuantiles, regresión cuantílica conformal, entrenamiento conjunto de cuantiles inferiores y superiores con restricciones de no cruce, o esquemas de ensamble que combinen la precisión del XGBoost con la capacidad distribucional de otros modelos cuantílicos.

Finalmente, la metodología desarrollada puede escalarse a otros sistemas productivos del país. El flujo de trabajo propuesto integración de datos abiertos, modelado cuantílico, validación temporal, estimación de brechas y análisis territorial podría adaptarse a cultivos estratégicos como arroz, café, cacao, papa, fríjol, plátano, caña panelera u otros productos priorizados. Esta escalabilidad representa una oportunidad para construir herramientas analíticas comparables que apoyen la planificación rural y el uso estratégico de datos públicos en Colombia.

9. RECURSOS A EMPLEAR

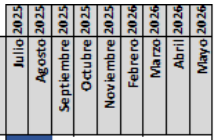
9.1. RECURSOS HUMANOS

Director del proyecto: David Arango Londoño

10. CRONOGRAMA DE ACTIVIDADES

Figura 6 *Cronograma de actividades*

Nº	Fase	Nº	Actividad	Duración estimada	Fecha inicio	Fecha Final
1	**Fase 1: Construcción de base de datos**	1	**Integración y construcción de la base de datos agroclimática y agroproductiva**	**6 semanas**	**1/7/2025**	**11/8/2025**
	Fase 1: Construcción de base de datos	1.1	Recolección de datos	2 semanas	1/7/2025	14/7/2025
	Fase 1: Construcción de base de datos	1.2	Procesamiento, georreferenciación e imputación espacio-temporal de datos	5 semanas	15/7/2025	12/8/2025
	Fase 1: Construcción de base de datos	1.3	Documentación y almacenamiento	1 semana	12/8/2025	18/8/2025
2	**Fase 2: Análisis y selección de variables**	2	**Selección y análisis de variables predictoras**	**5 semanas**	**19/8/2025**	**22/9/2025**
	Fase 2: Análisis y selección de variables	2.1	Análisis exploratorio de datos (EDA)	5 semanas	19/8/2025	22/9/2025
	Fase 2: Análisis y selección de variables	2.2	Selección preliminar de variables predictoras			
3	**Fase 3: Evaluación de relevancia estadística**	3	**Evaluación de relevancia estadística y explicativa**	**3 semanas**	**23/9/2025**	**13/10/2025**
	Fase 3: Evaluación de relevancia estadística	3.1	Análisis de correlaciones de variables	1 semana	23/9/2025	29/9/2025
	Fase 3: Evaluación de relevancia estadística	3.2	Detección de efectos no lineales	1 semana	30/9/2025	6/10/2025
	Fase 3: Evaluación de relevancia estadística	3.3	Identificación de interacciones	1 semana	7/10/2025	13/10/2025
4	**Fase 4: Modelado predictivo**	4	**Desarrollo de modelos predictivos y cuantificación de incertidumbre**	**6 semanas**	**21/10/2025**	**1/12/2025**
	Fase 4: Modelado predictivo	4.1	Preparación de conjuntos de datos	1 semana	21/10/2025	27/10/2025
	Fase 4: Modelado predictivo	4.2	Entrenamiento de modelos predictivos	5 semanas	28/10/2025	1/12/2025
	Fase 4: Modelado predictivo	4.3	Ajuste de modelos cuantílicos			
5	**Fase 5: Enfoque estocástico**	5	**Introducción de enfoque estocástico**	**3.5 semanas**	**1/2/2026**	**26/2/2026**
	Fase 5: Enfoque estocástico	5.1	Estimación de intervalos de predicción	1 semana	1/2/2026	7/2/2026
	Fase 5: Enfoque estocástico	5.2	Cuantificación de la incertidumbre	1 semana	8/2/2026	14/2/2026
6	**Fase 6: Validación y evaluación del modelo**	6	**Evaluación y validación del modelo final**	**3.5 semanas**	**27/2/2026**	**20/3/2026**
	Fase 6: Validación y evaluación del modelo	6.1	Validación cruzada	2 semanas	27/2/2026	12/3/2026
	Fase 6: Validación y evaluación del modelo	6.2	Construcción de ventanas temporales de validación (Rolling Window) temporale	3 días	13/3/2026	17/3/2026
	Fase 6: Validación y evaluación del modelo	6.3	Reentrenamiento y evaluación temporal iterativa	3 días	18/3/2026	20/3/2026
	Fase 6: Validación y evaluación del modelo	6.4	Evaluación por municipio y año	2 días	23/3/2026	24/3/2026
	Fase 6: Validación y evaluación del modelo	6.5	Agregación y resumen de resultados	2 días	25/3/2026	26/3/2026
7	**Fase 7: Métricas de evaluación**	7	**Métricas de evaluación**	**2 semanas**	**27/3/2026**	**9/4/2026**
	Fase 7: Métricas de evaluación	7.1	Definición de criterios de evaluación	3 días	27/3/2026	31/3/2026
	Fase 7: Métricas de evaluación	7.2	Diseño del esquema de validación	3 días	1/4/2026	3/4/2026
	Fase 7: Métricas de evaluación	7.3	Aplicación del proceso de evaluación	3 días	6/4/2026	8/4/2026
8	**Fase 8: Análisis de error (ineficiencia)**	8	**Análisis territorial de errores y brechas de rendimiento**	**3 semanas**	**13/4/2026**	**1/5/2026**
	Fase 8: Análisis de error	8.1	Identificación de municipios con mayor desviación	5 días	13/4/2026	17/4/2026
	Fase 8: Análisis de error	8.2	Caracterización de errores (ineficiencias)	5 días	20/4/2026	24/4/2026
	Fase 8: Análisis de error	8.3	Evaluación territorial de precisión	3 días	27/4/2026	29/4/2026
	Fase 8: Análisis de error	8.4	Correlación de errores con variables explicativas	1 semana	30/4/2026	7/5/2026
9	**Fase 9: Interpretación y estimación del rendimiento**	9	**Interpretación del modelo y estimación del rendimiento potencial**	**1.5 semanas**	**8/5/2026**	**15/5/2026**
	Fase 9: Interpretación y estimación del rendimiento	9.1	Estimación del rendimiento potencial	4 días	8/5/2026	13/5/2026
10	**Fase 10: Conclusiones**	10	**Conclusiones**	**2 días**	**14/5/2026**	**15/5/2026**
	Fase 10: Conclusiones	10.1	Conclusiones	2 días	14/5/2026	15/5/2026



11. REFERENCIAS BIBLIOGRÁFICAS

[1] Departamento Administrativo Nacional de Estadística (DANE), “Encuesta nacional agropecuaria (ENA) Históricos.”

[2] M. y E. A. (IDEAM) Instituto de Hidrología, “Boletín agroclimático nacional.”

[3] G.-H. Morán-Figueroa, D.-F. Muñoz-Pérez, J.-L. Rivera-Ibarra, and C.-A. Cobos-Lozada, “Model for Predicting Maize Crop Yield on Small Farms Using Clusterwise Linear Regression and GRASP,” *Mathematics*, vol. 12, no. 21, p. 3356, Oct. 2024, doi: 10.3390/math12213356.

[4] S. I. Golik, S. Larran, G. S. ; Gerard, and M. C. Fleitas, *Cereales de verano*. Editorial de la Universidad Nacional de La Plata (EDULP), 2018. doi: 10.35537/10915/68613.

[5] Unidad de planificación rural agropecuaria (UPRA), “Evaluaciones agropecuaria municipales 2021,” 2021.

[6] J. C. Ovalle Másmela, “Tecnologías emergentes para el agro y su aplicación en Colombia,” Mosquera, Cundinamarca, 2023. doi: 10.21930/agrosavia.estudiodevigilancia.2023.2.

[7] M. A. Altieri, *Seguimiento de los progresos relativos a los indicadores de los ODS relacionados con la alimentación y la agricultura 2023*. FAO, 2023. doi: 10.4060/cc7088es.

[8] CIAT and CIMMYT, “Maíz para Colombia vision 2030,” 2019.

[9] M. K. van Ittersum and R. Rabbinge, “Concepts in production ecology for analysis and quantification of agricultural input-output combinations,” *Field Crops Res.*, vol. 52, no. 3, pp. 197–208, Jun. 1997, doi: 10.1016/S0378-4290(97)00037-3.

[10] D. Aigner, C. A. K. Lovell, and P. Schmidt, “Formulation and estimation of stochastic frontier production function models,” *J. Econom.*, vol. 6, no. 1, pp. 21–37, Jul. 1977, doi: 10.1016/0304-4076(77)90052-5.

[11] D. B. Lobell, K. G. Cassman, and C. B. Field, “Crop Yield Gaps: Their Importance, Magnitudes, and Causes,” *Annu. Rev. Environ. Resour.*, vol. 34, no. 1, pp. 179–204, Nov. 2009, doi: 10.1146/annurev.environ.041008.093740.

[12] J. van Wart *et al.*, “Use of agro-climatic zones to upscale simulated crop yield potential,” *Field Crops Res.*, vol. 143, pp. 44–55, Mar. 2013, doi: 10.1016/j.fcr.2012.11.023.

[13] R. Koenker and G. Bassett, “Regression Quantiles,” *Econometrica*, vol. 46, no. 1, p. 33, Jan. 1978, doi: 10.2307/1913643.

[14] *An Introduction to Efficiency and Productivity Analysis*. New York: Springer-Verlag, 2005. doi: 10.1007/b136381.

[15] I. A. Canay, “A simple approach to quantile regression for panel data,” *Econom. J.*, vol. 14, no. 3, pp. 368–386, Oct. 2011, doi: 10.1111/j.1368-423X.2011.00349.x.

[16] N. Meinshausen, “Quantile Regression Forests,” *Journal of machine learning research*, vol. 7, pp. 983–999, 2006.

[17] T. Chen and C. Guestrin, “XGBoost,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[18] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Conference on Neural Information Processing Systems*, California, 2017.

[19] W. Meeusen and J. van Den Broeck, “Efficiency Estimation from Cobb-Douglas Production

Functions with Composed Error," *Int. Econ. Rev. (Philadelphia).*, vol. 18, no. 2, p. 435, Jun. 1977, doi: 10.2307/2525757.

[20] G. E. Battese and T. J. Coelli, "Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data," *J. Econom.*, vol. 38, no. 3, pp. 387–399, Jul. 1988, doi: 10.1016/0304-4076(88)90053-X.

[21] R. Koenker and J. A. F. Machado, "Goodness of Fit and Related Inference Processes for Quantile Regression," *J. Am. Stat. Assoc.*, vol. 94, no. 448, p. 1296, Dec. 1999, doi: 10.2307/2669943.

[22] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *J. Am. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, Mar. 2007, doi: 10.1198/016214506000001437.

[23] R. L. Winkler, "A Decision-Theoretic Approach to Interval Estimation," *J. Am. Stat. Assoc.*, vol. 67, no. 337, p. 187, Mar. 1972, doi: 10.2307/2284720.

[24] R. Stull, "Wet-Bulb Temperature from Relative Humidity and Air Temperature," *J. Appl. Meteorol. Climatol.*, vol. 50, no. 11, pp. 2267–2269, Nov. 2011, doi: 10.1175/JAMC-D-11-0143.1.

[25] D. B. Lobell, M. Bänziger, C. Magorokosho, and B. Vivek, "Nonlinear heat effects on African maize as evidenced by historical yield trials," *Nat. Clim. Chang.*, vol. 1, no. 1, pp. 42–45, Apr. 2011, doi: 10.1038/nclimate1043.

[26] J. L. Hatfield and J. H. Prueger, "Temperature extremes: Effect on plant growth and development," *Weather Clim. Extrem.*, vol. 10, pp. 4–10, Dec. 2015, doi: 10.1016/j.wace.2015.08.001.

[27] G. Poveda, D. M. Álvarez, and Ó. A. Rueda, "Hydro-climatic variability over the Andes of Colombia associated with ENSO: a review of climatic processes and their impact on one of the Earth's most important biodiversity hotspots," *Clim. Dyn.*, vol. 36, no. 11–12, pp. 2233–2249, Jun. 2011, doi: 10.1007/s00382-010-0931-y.

[28] L. Anselin, *Spatial Econometrics: Methods and Models*, vol. 4. Dordrecht: Springer Netherlands, 1988. doi: 10.1007/978-94-015-7799-1.

[29] T. G. Conley, "GMM estimation with cross sectional dependence," *J. Econom.*, vol. 92, no. 1, pp. 1–45, Sep. 1999, doi: 10.1016/S0304-4076(98)00084-0.

[30] K. D. MORENO RUIZ and S. P. MORI JULCA, "Medición de la eficiencia técnica del maíz amarillo duro en los distritos de chongoyape, lagunas-mocupe y nueva arica: un análisis con frontera estocástica," Universidad católica Santo Toribio de mogrovejo , Chiclayo, 2017.

[31] H. Lamos-Díaz, D. E. Puentes-Garzón, and D. A. Zarate-Caicedo, "Comparison Between Machine Learning Models for Yield Forecast in Cocoa Crops in Santander, Colombia," *Revista Facultad de Ingeniería*, vol. 29, no. 54, p. e10853, May 2020, doi: 10.19053/01211129.v29.n54.2020.10853.

[32] A. , Morales-Ruiz and E. Díaz-López, "Influencia de la temperatura, precipitación y radiación solar en el rendimiento de maíz en el valle de toluca, méxico." Accessed: May 25, 2025. [Online]. Available: https://agrociencia-colpos.org/index.php/agrociencia/article/view/1933/2068

[33] Y. Yan, Y. Wang, J. Li, J. Zhang, and X. Mo, "Crop Yield Time-Series Data Prediction Based on

Multiple Hybrid Machine Learning Models," *Applied and Computational Engineering*, vol. 133, no. 1, pp. 215–221, Feb. 2025, doi: 10.54254/2755-2721/2025.20800.

[34] Y. Romano, E. Patterson, and E. J. Candès, "Conformalized Quantile Regression," in *Conference on Neural Information Processing Systems*, Vancouver, 2019.

[35] R. Wirth and Hipp Jochen, "CRISP-DM: Towards a Standard Process Model for Data Mining," 2000.

[36] Vijay. Kotu and Balachandre. Deshpande, *Data science : concepts and practice*, 2nd ed. Morgan Kaufmann Publishers, an imprint of Elsevier, 2019.

[37] Andrew. Gelman, *Bayesian Data Analysis*. CRC Press, 2013.

[38] J. K. . Kruschke, *Doing Bayesian data analysis : a tutorial with R, JAGS, and Stan*. Academic Press, 2015.

[39] Gobierno de Colombia, "EVALUACIÓN DE DAÑOS, PÉRDIDAS E IMPACTOS ASOCIADOS A LA OCURRENCIA DEL FENÓMENO DE LA NIÑA 2021-2023," 2023. Accessed: Jun. 15, 2026. [Online]. Available: https://www.fondoadaptacion.gov.co/images/2024/Gestion_del_Conocimiento/Estudios_Investigaciones_Publicaciones/Evaluacion_Danos_Perdidas_e_Impactos_Fenomeno_de_la_Nina_2021-2023.pdf