

Modelo de predicción de heladas con XGBoost integrado a APIs meteorológicas en tiempo real para zonas altoandinas del Perú

Frost prediction model with XGBoost integrated into real-time meteorological APIs for high Andean areas of Peru



Universidad Nacional Micaela Bastidas de Apurímac – Perú
Riqchary, revista de investigación en ciencia y tecnología

ISSN: 2810-8124 (en línea) / ISSN: 2706-543x
Vol. 8 Núm. 1 (2026) - Publicado: 01/04/2026
doi.org/10.57166/riqchary/v8.n1.2026

Páginas: 1- 10

Recibido 01/01/2025 ; Aceptado 01/02/2025
<https://doi.org/10.57166/riqchary/v8.n1.2026.1>

Jorge L. Anco Coaquira

Universidad Nacional del Altiplano, PE.
<https://orcid.org/0009-0002-1563-8632>
60066564@est.unap.edu.pe

Abad Poma Maquera

Universidad Nacional del Altiplano, PE.
<https://orcid.org/0009-0006-7203-8800>
60732897@est.unap.edu.pe

Gianmarco Alain Bruno Machaca Cahuana

Universidad Nacional del Altiplano, PE.
72073749@est.unap.edu.pe
<https://orcid.org/0009-0006-5062-2002>

Aldo H. Zanabria Gálvez

Universidad Nacional del Altiplano, PE.
aldo.zanabria@unap.edu.pe
<https://orcid.org/0000-0003-3314-8768>

Resumen— Las heladas representan uno de los principales riesgos climáticos para la agricultura en las zonas altoandinas del Perú, debido a su impacto directo en la productividad agrícola y la seguridad alimentaria de las poblaciones rurales. Debido a esta situación, se abordó el problema evaluando el desempeño de un modelo de aprendizaje automático basado en XGBoost con el objetivo de predecir eventos de helada, integrando datos meteorológicos históricos y en tiempo real. El análisis se realizó utilizando un conjunto de datos compuesto por 1 672 825 observaciones en 25 años provenientes de once regiones altoandinas, obtenidas de SENAMHI y la plataforma NASA POWER que fué sometida a procesos de limpieza, integración y balanceo. El modelo fue entrenado con el 80% de los datos y evaluado con el 20% restante mediante un esquema de validación estratificada, empleando métricas como accuracy, precision, recall, F1-score y AUC-ROC. Los resultados muestran un desempeño sobresaliente, con una exactitud del 95.17% y un AUC-ROC de 0.9938, evidenciando una alta capacidad para detectar eventos de helada incluso en escenarios de clases desbalanceadas. Finalmente, la implementación del modelo en un sistema de predicción en tiempo real con APIs meteorológicas y su visualización a través de un dashboard interactivo demuestran su potencial como herramienta de apoyo para sistemas de alerta temprana y la gestión del riesgo agrícola en zonas altoandinas.

Palabras clave: aprendizaje automático, heladas, predicción meteorológica, zonas altoandinas.

Abstract— Frost represents one of the main climatic risks for agriculture in the high Andean areas of Peru, due to its direct impact on agricultural productivity and food security for rural populations. In response to this situation, the problem was addressed by evaluating the performance of a machine learning model based on XGBoost with the aim of predicting frost events, integrating historical and real-time meteorological data. The analysis was performed using a dataset consisting of 1,672,825 observations over 25 years from eleven high Andean regions, obtained from SENAMHI and the NASA POWER platform, which was subjected to cleaning, integration, and balancing processes. The model was trained with 80% of the data and evaluated with the remaining 20% using a stratified validation scheme, employing metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. The results show outstanding performance, with an accuracy of 95.17% and an AUC-ROC of 0.9938, demonstrating a high capacity to detect frost events even in unbalanced class scenarios. Finally, the implementation of the model in a real-time prediction system with meteorological APIs and its visualization through an interactive dashboard demonstrate its potential as a support tool for early warning systems and agricultural risk management in high Andean areas.

Keywords: machine learning, frosts, weather forecasting, high Andean areas.

1 INTRODUCCIÓN

Las heladas son fenómenos atmosféricos que ocurren por la disminución de la nubosidad en los Andes y se produce cuando la temperatura ambiental disminuye más allá de los cero grados centígrados, durante los meses de junio, julio y agosto [1]. Este fenómeno se origina en su mayoría en las zonas ubicadas a más de 3000 metros sobre el nivel del mar, lo que corresponde a la zona altoandina del Perú [2]. Por lo cual representa una de las principales amenazas para la agricultura en las zonas altoandinas del Perú, debido a los daños irreversibles que ocasiona en los cultivos, la reducción significativa de los rendimientos agrícolas y su impacto directo en la seguridad alimentaria de poblaciones rurales [3]. En el 2024 el informe técnico del Centro Nacional de Estimación, Prevención y Reducción del Riesgo de Desastres (Cenepred) destaca que 1,642,026 hectáreas (ha) de superficie agrícola de 328 distritos enfrentan niveles de riesgo muy alto debido a las heladas [4]. Estos registros evidencian la necesidad de sistemas de monitoreo y predicción que permitan anticipar oportunamente sus efectos adversos.

En lo que respecta al contexto socioeconómico peruano, las heladas afectan de manera desproporcionada a los pequeños productores agrícolas, quienes presentan una limitada capacidad de adaptación frente a la variabilidad climática [5], [6]. Estudios previos han señalado que la exposición recurrente a eventos extremos incrementa el riesgo de pérdidas económicas donde las papas nativas son su principal fuente de alimentación [7].

A pesar de la importancia y el uso de sensores climatológicos, los métodos tradicionales de predicción empleados en zonas altoandinas presentan limitaciones significativas [8]. Existen sectores donde los factores meteorológicos y físicos tienden a ser difíciles para adquirir datos, particularmente en regiones caracterizadas por una topografía compleja y alta variabilidad microclimática [9]. Asimismo, los modelos estadísticos clásicos, como la regresión lineal múltiple o la regresión logística, han sido utilizados históricamente para la predicción de heladas; sin embargo, diversos estudios han demostrado que su desempeño es limitado [10], ya que dependen de series históricas densas y de una red extensa de estaciones meteorológicas, condiciones que no siempre están disponibles en zonas rurales de altura.

Visto que, el uso de técnicas de machine learning ha mostrado un desempeño superior en la predicción y clasificación de fenómenos climáticos complejos [11]. Modelos como Random Forest y Extreme Gradient Boosting (XGBoost) [12], han sido ampliamente utilizados en estudios recientes, demostrando una mayor capacidad para

capturar relaciones no lineales y manejar grandes volúmenes de datos heterogéneos [13].

Una de las principales fortalezas de XGBoost radica en la optimización de una función objetivo, lo que contribuye a reducir el sobreajuste y mejorar la capacidad de generalización del modelo [14]. Adicionalmente, el algoritmo incorpora aproximaciones de gradiente de segundo orden [15], técnicas de submuestreo y una estructura de datos optimizada (DMatrix), lo que permite acelerar el proceso de entrenamiento y mejorar la eficiencia computacional, incluso en escenarios con grandes volúmenes de datos y características heterogéneas [16].

Además, la capacidad de este modelo radica en capturar interacciones complejas entre variables meteorológicas — como temperatura, humedad relativa, punto de rocío, precipitación y velocidad del viento — permite una representación más precisa de los procesos físicos involucrados en la formación de heladas [12]. Por otra parte, la incorporación de mecanismos de regularización L1/L2 contribuye a mitigar el sobreajuste, fortaleciendo la estabilidad del modelo en regiones con alta variabilidad climática y topografía compleja [14]. Estudios han reportado que XGBoost supera en precisión y robustez a modelos como regresión logística, árboles de decisión y Random Forest [15] en tareas de predicción climática y análisis de series temporales, particularmente en la estimación de temperatura mínima y la detección de eventos extremos [16].

No obstante, a pesar de estos avances, aún existe una limitada aplicación de modelos de aprendizaje automático integrados con fuentes de datos meteorológicos en tiempo real, especialmente orientados a la predicción de heladas en zonas altoandinas con infraestructura meteorológica restringida [17].

En este escenario, la disponibilidad de plataformas de datos meteorológicos abiertos, como los servicios del SENAMHI y la base de datos NASA POWER, ha impulsado el desarrollo de metodologías predictivas más accesibles y replicables. Investigaciones previas [18] han demostrado que los datos provenientes de NASA POWER presentan una alta correspondencia con observaciones terrestres, particularmente en variables como temperatura, humedad y punto de rocío, lo que los convierte en una alternativa viable para regiones con escasa cobertura instrumental. En consecuencia, el objetivo de la presente investigación es evaluar la adaptabilidad y el desempeño de un modelo de machine learning basado en XGBoost para la predicción de heladas en zonas altoandinas del Perú, integrándolo a datos meteorológicos en tiempo real. Se busca generar información predictiva confiable que contribuya a la imple-

mentación de estrategias de alerta temprana y a la reducción del riesgo asociado a eventos de heladas en comunidades agrícolas vulnerables.

En este contexto, se plantea la siguiente hipótesis de investigación: la integración de un modelo de machine learning basado en XGBoost con datos meteorológicos en tiempo real provenientes de plataformas abiertas permite mejorar significativamente la precisión en la predicción de eventos de heladas en zonas altoandinas del Perú, en comparación con métodos estadísticos tradicionales.

3 METODOLOGÍA

En esta conjunto se presenta la arquitectura general del modelo (véase la Figura 1), la cual describe el flujo completo desde la adquisición y preprocesamiento de datos meteorológicos hasta el entrenamiento y evaluación del modelo XGBoost para la predicción de heladas.

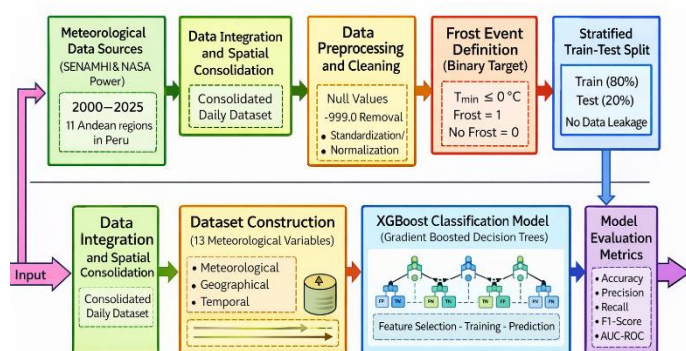


Figura 1. Arquitectura general del modelo XGBoost para la predicción de heladas

La arquitectura integra fuentes de datos meteorológicos abiertos y organiza el proceso de modelado dentro de un pipeline reproducible orientado a la detección de eventos de helada.

3.1 Área de estudio y fuentes de datos meteorológicos

La investigación se desarrolló en once regiones altoandinas del Perú con alta incidencia histórica de heladas: Áncash, Apurímac, Arequipa, Ayacucho, Cusco, Huancavelica, Huánuco, Junín, Moquegua, Puno y Tacna. Para cada región se seleccionaron entre 15 y 25 puntos de estudio meteorológico, definidos en función de la frecuencia histórica de ocurrencia de heladas y de la disponibilidad de registros climáticos continuos.

Los datos meteorológicos fueron obtenidos a partir del repositorio oficial del Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI) y de la plataforma

NASA POWER, la cual proporciona información reanalizada de variables atmosféricas como temperatura, humedad y punto de rocío. En conjunto, se recopilieron series temporales diarias correspondientes al periodo comprendido entre el 1 de enero de 2000 y el 14 de noviembre de 2025, lo que equivale aproximadamente 25 años de observaciones diarias y continuas por cada punto seleccionado; el conjunto de datos utilizado en esta investigación fue consolidado y publicado en la plataforma Kaggle [19].



Figura 2. Distribución espacial de los puntos de observación

3.2 Construcción del conjunto de datos

Los registros meteorológicos correspondientes a las once regiones de estudio fueron integrados en un único conjunto de datos unificado, con el objetivo de disponer de una base continua, homogénea y adecuada para el entrenamiento del modelo de aprendizaje automático [19]. Este proceso permitió consolidar la información climática proveniente de distintas zonas altoandinas en un solo archivo estructurado, facilitando su posterior procesamiento y análisis.

El conjunto de datos final estuvo compuesto por un total de 1 672 825 observaciones válidas, correspondientes a trece variables meteorológicas consideradas relevantes para la predicción de heladas. El tamaño y la diversidad del dataset resultante proporcionaron una base robusta para el entrenamiento, evaluación y validación del modelo XGBoost, asegurando estabilidad estadística y representatividad temporal.

3.3 Variables meteorológicas consideradas

TABLA 1
Lista de variables meteorológicas utilizadas

Variable	Identificador de variable
Punto de rocío a 2m	punto_rocio_2m
Temperatura a 2m	temp_2m
Altitud	altitud
Temperatura máxima a 2m	temp_max_2m
Humedad relativa 2m	humedad_rel_2m
Latitud	latitud
Longitud	longitud
Día	día
Precipitación	precipitacion
Viento a 10m	viento_10m
Viento máximo 10m	viento_max_10m
Año	año
Viento mínimo 10m	viento_min_10m

Para el desarrollo del modelo predictivo se consideran trece variables meteorológicas, incluyendo temperatura del aire a 2 m, temperatura máxima diaria, punto de rocío, humedad relativa, precipitación, velocidad del viento, así como variables geográficas y temporales como altitud, latitud, longitud, día y año. La selección de estas variables se realizó considerando su relevancia reportada en estudios previos sobre la predicción de heladas y fenómenos climáticos extremos.

3.4 Preprocesamiento de datos

Una vez consolidado el conjunto de datos principal, se aplicó un proceso de limpieza y preprocesamiento con el fin de garantizar la calidad de la información utilizada en el entrenamiento del modelo. Se eliminaron los registros que presentaban valores nulos o inválidos, especialmente aquellos codificados como -999.0, valor comúnmente utilizado por plataformas meteorológicas como NASA POWER para indicar datos faltantes o errores de medición. La implementación del pipeline de preprocesamiento y del modelo XGBoost se desarrolló en Python y se encuentra disponible en un repositorio público [20].

Tras completar el proceso de limpieza, se obtuvo un conjunto de datos consistente y libre de registros erróneos, apto para su uso en las etapas posteriores de entrenamiento y evaluación del modelo.

3.5 Definición del evento de helada

En este estudio, un evento de helada se definió como la ocurrencia de una temperatura mínima diaria igual o inferior a 0 °C a 2 m de altura. Esta definición permitió formular el problema como una tarea de clasificación binaria, donde la variable objetivo toma el valor de 1 cuando ocurre una helada y 0 en caso contrario.

3.6 División de datos y estrategia de validación

El conjunto de datos final fue preparado para el entrenamiento del modelo mediante la definición explícita de una variable objetivo-binaria denominada *helada*. Implementado un script en python para nuestro pipeline de preparación se etiquetó como evento de helada toda observación cuya temperatura mínima diaria a 2 m de altura fue igual o inferior a 0 °C, asignándose el valor 1 en caso afirmativo y 0 en caso contrario. Esta definición permitió formular el problema como una tarea de clasificación binaria orientada a la detección de eventos de baja frecuencia.

Posteriormente, se seleccionaron exclusivamente las variables meteorológicas y geográficas relevantes para el modelo predictivo, incluyendo temperatura media y máxima del aire, punto de rocío, humedad relativa, precipitación, velocidad del viento, así como latitud, longitud, altitud y variables temporales (año y día del año). Con ello se construyó un subconjunto de datos depurado y consistente, utilizado de manera exclusiva en la fase de modelado.

El conjunto de datos fue dividido en subconjuntos de entrenamiento y prueba empleando un esquema de retención que corresponde a nuestro pipeline de entrenamiento, asignando el 80 % de las observaciones al entrenamiento y el 20 % restante al pipeline de evaluación. Dado el carácter inherentemente desbalanceado del fenómeno de estudio, la partición se realizó de manera estratificada respecto a la variable objetivo, garantizando que la proporción de eventos de helada y no helada se mantuviera constante en ambos subconjuntos.

Con el fin de asegurar la reproducibilidad de los resultados, se estableció una semilla aleatoria fija durante el proceso de partición. Asimismo, se garantizó el aislamiento total del conjunto de prueba, el cual no fue utilizado en ninguna etapa del ajuste de parámetros ni del entrenamiento del modelo, evitando así problemas de data leakage. El subconjunto de prueba fue reservado exclusivamente para la evaluación final del desempeño del modelo mediante las métricas definidas en la sección correspondiente.

Esta estrategia de división y validación permitió obtener una evaluación objetiva y metodológicamente sólida del modelo XGBoost, especialmente adecuada para la predicción de eventos climáticos extremos de baja frecuencia en regiones altoandinas.

3.7 Métricas de evaluación

La evaluación del desempeño del modelo predictivo se realizó utilizando un conjunto independiente de prueba, reservado exclusivamente para esta etapa y no empleado durante el entrenamiento. Dado que la predicción de heladas constituye un problema de clasificación binaria con clases desbalanceadas, se seleccionó un conjunto de métricas que permiten evaluar no solo la exactitud global del modelo, sino también su capacidad para identificar correctamente eventos de helada, considerados de mayor relevancia desde el punto de vista agrícola.

La exactitud (Accuracy) se empleó como una medida general del porcentaje de clasificaciones correctas. La precisión (Precision) y la recuperación (Recall) fueron utilizadas para cuantificar la capacidad del modelo de identificar correctamente los eventos de helada y minimizar tanto las falsas alarmas como las omisiones. Asimismo, se calculó la **puntuación F1**, proporcionando una evaluación balanceada del rendimiento del modelo.

Adicionalmente, se construyó la matriz de confusión, con el objetivo de evaluar la capacidad discriminativa del modelo independientemente del umbral de decisión. Se calculó el Área Bajo la Curva ROC (AUC-ROC) a partir de las probabilidades estimadas por el modelo, siendo especialmente adecuada para problemas de clasificación con eventos raros.

Todas las métricas fueron calculadas a partir de las predicciones obtenidas sobre el conjunto de prueba, garantizando una evaluación objetiva y reproducible del desempeño del modelo XGBoost. Asimismo, los resultados obtenidos fueron almacenados en formatos tabulares y gráficos para su posterior análisis y presentación.

4 RESULTADOS

4.1 Características del dataset final

El dataset final está compuesto por 1 672 825 observaciones válidas, resultantes del proceso de integración de los registros meteorológicos obtenidos para las 11 regiones altoandinas del Perú. En total, se consolidaron 13 variables climáticas provenientes de las plataformas SENAMHI y NASA POWER, lo que permitió contar con un conjunto de datos amplio, heterogéneo y representativo de distintas condiciones atmosféricas y topográficas.

Durante la construcción del dataset se identificó un marcado desbalance en la variable objetivo, evidenciando que los días con helada son considerablemente menos frecuentes que los días sin helada, característica inherente a la naturaleza del fenómeno. La gran extensión temporal y espacial de los datos permitió capturar eventos de helada en diversos escenarios climáticos, garantizando suficiente variabilidad para entrenar y evaluar modelos de aprendizaje automático con solidez estadística.

4.2 Importancia de variables meteorológicas

TABLA 2
Importancia de las variables meteorológicas

Ranking	Variable	Importancia (%)
1	punto_rocio_2m	50.10%
2	temp_2m	21.53%
3	altitud	7.88%
4	temp_max_2m	5.60%

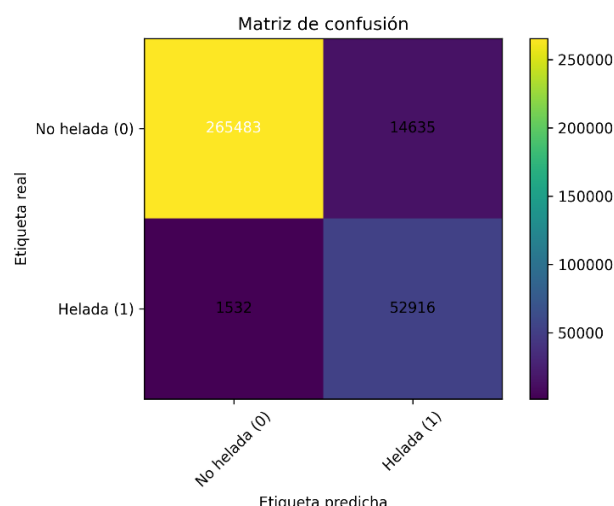
5	humedad_rel_2m	5.37%
6	latitud	2.59%
7	longitud	1.71%
8	dia	1.26%
9	precipitacion	1.08%
10	viento_10m	0.89%
11	viento_max_10m	0.73%
12	año	0.71%
13	viento_min_10m	0.53%

La Tabla 2 representa las variables de estudio y su porcentaje de importancia utilizadas en el modelo predictivo de heladas y los resultados nos muestra que el punto de rocío a 2 metros que es la variable más importante con un (50,10%), lo que influye bastante con la ocurrencia de heladas debido a su relación directa en la formación de hielo superficial. Luego, le sigue la temperatura del aire a 2 metros (21,53%), lo cual esta variable es esencial para determinar temperaturas bajas. En menor medida, la altitud (7.88%) y la temperatura máxima diaria (5.60%) que también contribuyen un mejor entrenamiento y aprendizaje del modelo al revelar condiciones topográficas y variabilidad térmica.

En segundo orden, le sigue las variables como la humedad relativa, latitud, longitud y día del año que presentan porcentajes de importancia (entre 1% y 5%), mientras que la precipitación y distintos casos de viento aportan menos del (1%), lo que propone que tiene menor impacto en la predicción de heladas. Dicho esto, los resultados confirman que las condiciones termodinámicas locales son las más esenciales a la hora de determinar comportamientos del modelo para la predicción de heladas.

4.3 Desempeño del modelo XGBoost

El modelo XGBoost mostró un rendimiento sobresaliente en la predicción de eventos de helada, alcanzando una exactitud global de 95.17% y un AUC-ROC de 0.9938, lo que evidencia una excelente capacidad discriminativa



entre condiciones de helada y no helada. Estas métricas confirman que el modelo captura de manera efectiva relaciones no lineales presentes en el comportamiento térmico de las zonas altoandinas.

Figura 3. Matriz de confusión

La matriz de confusión (véase la Figura 3) muestra un adecuado equilibrio entre sensibilidad y especificidad. El modelo logró identificar correctamente 52 916 eventos reales de helada, manteniendo un número reducido de falsos negativos (1 532 casos), lo cual es especialmente relevante para aplicaciones agrícolas donde la omisión de una helada puede generar pérdidas severas. Asimismo, el número de falsos positivos se mantuvo dentro de un rango manejable (14 635 casos), preservando la utilidad del sistema como herramienta preventiva.

TABLA 3

Métricas de desempeño del modelo XGBoost

Clase / Métrica	Preci-sion	Recall	F1-score	Support
0	0.9943	0.9478	0.9705	280118
1	0.7833	0.9719	0.8675	54448
Accuracy	—	—	0.9517	334566
Macro avg	0.8888	0.9598	0.919	334566
Weighted avg	0.9599	0.9517	0.9537	334566

Adicionalmente, el reporte de clasificación (véase la Tabla 3) evidencia que la clase minoritaria (helada) obtuvo un recall de 97.19% y un F1-score de 0.8675, lo que demuestra que el modelo mantiene un alto nivel de sensibilidad incluso en presencia de desbalanceo en las clases. La curva ROC (véase la Figura 4) confirma este comportamiento al mostrar una separación casi perfecta entre ambas clases.

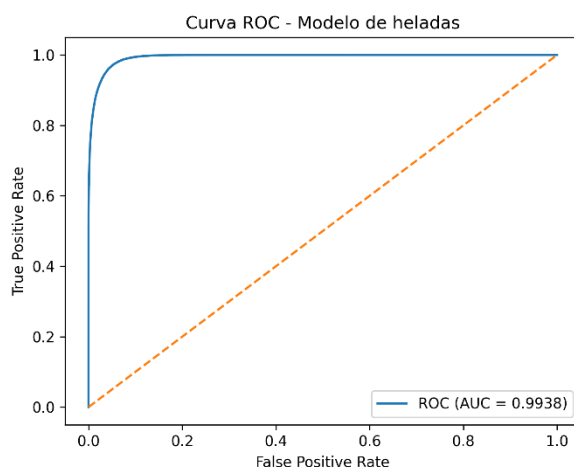


Figura 4. Curva ROC

4.4 Resultados de validación cruzada

La validación cruzada se realizó utilizando un esquema k-fold con $k = 5$, aplicando particiones aleatorias del conjunto de entrenamiento para evaluar la estabilidad general del modelo XGBoost. Este procedimiento permitió

obtener métricas independientes por cada pliegue, incluyendo *accuracy*, *precision*, *recall*, *F1-score* y *AUC*, evitando cualquier sesgo asociado a una única división entrenamiento-prueba. Los resultados mostraron una alta consistencia entre pliegues, con variaciones mínimas en todas las métricas evaluadas, lo que evidencia que el modelo mantiene un comportamiento estable al ser expuesto a distintas particiones del dataset.

TABLA 4

Resultados promedio de validación cruzada

Fold	Accu-racy	Preci-sion	Recall	F1-score	AUC
1	0.970069	0.919912	0.893185	0.906351	0.994495
2	0.97001	0.923448	0.890026	0.906429	0.994622
3	0.969531	0.921736	0.888246	0.904681	0.994478
4	0.969931	0.920494	0.891514	0.905772	0.994439
5	0.970421	0.922569	0.89408	0.908101	0.994619
Promedio	0.969992	0.921632	0.89141	0.906267	0.99453

La baja dispersión observada en las métricas sugiere ausencia de sobreajuste y confirma que el modelo generaliza adecuadamente sobre los datos, lo que respalda su uso para predicciones operacionales en condiciones reales (véase la Tabla 4).

4.5 Comparación con otros modelos

Para evaluar el rendimiento relativo del algoritmo XGBoost, se realizó una comparación directa con modelos convencionales utilizados en clasificación binaria: regresión logística y random forest. La comparación se ejecutó bajo un esquema de validación cruzada de tres pliegues y utilizando un subconjunto representativo de 200 000 observaciones para acelerar el procesamiento. Las métricas evaluadas incluyeron *accuracy*, *precision*, *recall*, *F1-score* y *AUC*, garantizando una valoración integral del desempeño de cada modelo sobre las mismas particiones del dataset.

TABLA 5

Comparación de modelos

Modelo	Accu-racy	Preci-sion	Re-call	F1-score	AUC
Logistic Regression	0.936775	0.865866	0.721873	0.787337	0.972283
Random Forest	0.959595	0.903637	0.840426	0.870879	0.989113
XGBoost	0.96602	0.909484	0.877795	0.893358	0.9925

Los resultados obtenidos mostraron que XGBoost alcanzó sistemáticamente los valores más altos en F1-score, recall y AUC, superando a los modelos lineales y a los métodos de ensamble tradicionales (véase la Tabla 5).

4.6 Evaluación en escenarios de predicción en tiempo real

El sistema de predicción en tiempo real integra servicios externos para obtener de manera automática las variables meteorológicas necesarias para alimentar el modelo XGBoost previamente entrenado. Para cada consulta, el usuario ingresa las coordenadas geográficas y la fecha de interés; a partir de esta información, se ejecuta un proceso de geocodificación inversa mediante la API Nominatim de OpenStreetMap, lo que permite identificar el departamento, provincia y distrito asociados al punto consultado. Esta información contextualiza territorialmente el pronóstico generado.

Posteriormente, el sistema obtiene las variables meteorológicas diarias a través de la API *Open-Meteo*, incluyendo temperatura máxima, temperatura media, punto de rocío, humedad relativa, velocidad del viento en sus distintos valores característicos (media, mínima y máxima), precipitación acumulada y altitud del punto geográfico evaluado. Estos parámetros representan la dinámica atmosférica local y son devueltos por la API para la fecha solicitada.

```
model = pickle.load(f)
Modelo cargado correctamente.

Introduce los datos para obtener el pronóstico y la predicción de helada:

Latitud (grados decimales, ej. -15.84): -13.53
Longitud (grados decimales, ej. -70.02): -71.93
Fecha a predecir (formato YYYY-MM-DD, ej. 2025-08-15): 2025-12-11

Fecha ingresada: 2025-12-11 (año=2025, día_juliano=345)

Consultando ubicación aproximada (Nominatim / OpenStreetMap)...
Ubicación aproximada detectada:
Departamento: Cusco
Distrito/Localidad: San Sebastián

Llamando a la API de Open-Meteo...

Variables diarias obtenidas de la API:
temp_max_2m = 19.5 °C
temp_2m (media) = 13.4 °C
punto_rocio_2m = 8.3 °C
humedad_rel_2m = 74 %
viento_10m (media) = 8.0 m/s
viento_min_10m = 2.2 m/s
viento_max_10m = 18.8 m/s
precipitacion = 7.3 mm
altitud (API) = 3305.0 m

Vector de entrada al modelo:
año día temp_max_2m temp_2m ... precipitacion latitud longitud altitud
0 2025 345 19.5 13.4 ... 7.3 -13.53 -71.93 3305.0
[1 rows x 13 columns]
```

Figura 5. Flujo del sistema en tiempo real

Con los datos recopilados, se construye un vector de características estructurado en el mismo formato empleado durante el entrenamiento del modelo. Este vector es ingresado al clasificador XGBoost, almacenado en formato *pickle*,

```
RESULTADO DE LA PREDICCIÓN
-----
Ubicación aproximada de la predicción:
Departamento: Cusco
Distrito/Localidad: San Sebastián

Probabilidad de helada (clase 1): 0.00 %
✅ El modelo PREDICE: NO habrá helada en la fecha indicada.
```

a partir del cual se obtiene la probabilidad de ocurrencia de helada y la clasificación correspondiente. Todo el flujo garantiza coherencia entre los datos históricos utilizados en la etapa de entrenamiento y los datos operativos obtenidos en tiempo real, lo que permite generar predicciones confiables para diversas zonas altoandinas.

Figura 6. Resultado de la figura en tiempo real

4.7 Implementación aplicada del dashboard

La implementación aplicada del sistema se desarrolló mediante una interfaz gráfica tipo dashboard utilizando Streamlit, con el objetivo de proporcionar un modelo intuitivo y sencillo de usar. Esta interfaz integra de manera automática los servicios externos de adquisición de datos y el modelo predictivo entrenado, permitiendo al usuario realizar predicciones para cualquier punto geográfico dentro del Perú utilizando únicamente coordenadas y una fecha objetivo (véase la Figura 7).



Figura 7. Pantalla general del dashboard

Este dashboard incorpora un módulo de geocodificación inversa basado en la API Nominatim de OpenStreetMap, mediante el cual se identifican las divisiones político-administrativas asociadas a las coordenadas ingresadas (departamento y provincia). De forma paralela, la interfaz consulta en tiempo real la API Open-Meteo, desde donde se extraen las variables diarias necesarias para alimentar al modelo XGBoost, tales como temperatura máxima y media a 2 metros, punto de rocío, humedad relativa, velocidad del viento en sus valores característicos, precipitación acumulada y la altitud del punto consultado (véase la Figura 8).

	Variable	Valor
0	Temp. máxima (°C)	11.8
1	Temp. media (°C)	5.6
2	Punto de rocío (°C)	2.6
3	Humedad relativa (%)	83
4	Viento medio (m/s)	6.4
5	Viento mínimo (m/s)	2.3
6	Viento máximo (m/s)	11.4
7	Precipitación (mm)	25.7

Figura 8. Módulo de variables meteorológicas consultadas

Una vez descargadas, las variables son estructuradas mediante un constructor interno que garantiza el orden y formato requerido por el modelo. El sistema genera automáticamente la probabilidad de ocurrencia de helada y su clasificación correspondiente, utilizando un umbral ajustado y codificado según niveles de riesgo (bajo, moderado o alto). Los resultados se muestran en un panel principal junto con la ubicación detectada y las variables meteorológicas procesadas.

El dashboard incluye un componente cartográfico basado en Folium, donde el punto consultado es representado mediante un marcador codificado por color según el nivel de riesgo estimado. Esta integración facilita la interpretación espacial del fenómeno y permite su uso en aplicaciones territoriales de monitoreo y gestión agrícola. La estructura modular implementada en Streamlit permite extender el sistema hacia plataformas web o móviles sin modificar la lógica de predicción ni los servicios de adquisición de datos.

5 DISCUSIONES

La discusión de los hallazgos se centra en los factores meteorológicos identificados por el modelo y su coherencia con la literatura existente sobre la predicción de heladas. En este estudio, el punto de rocío y la temperatura del aire a 2 m emergen como los predictores más influyentes, evidenciando que los procesos termodinámicos locales son determinantes en la ocurrencia de heladas en las zonas altoandinas. Este resultado se alinea con lo reportado por Pérez Tárraga, quien señala que las variables asociadas al balance térmico superficial presentan una mayor capacidad explicativa frente a otros factores meteorológicos [21]. De igual manera, la relevancia de la altitud refuerza el carácter local del fenómeno, ya que condiciona la dinámica térmica en regiones de alta montaña, aspecto que también ha sido destacado en estudios clásicos de predicción de heladas agrícolas en Córdoba, Argentina [22].

Desde una perspectiva metodológica, las evidencias confirman que el uso de XGBoost permite capturar relaciones no lineales complejas entre variables meteorológicas, superando las limitaciones de los modelos tradicionales empleados históricamente en la predicción de heladas. A diferencia de enfoques basados en umbrales fijos o modelos lineales, el modelo desarrollado mostró una elevada capacidad para identificar eventos poco frecuentes [23], priorizando la reducción de falsos negativos, lo cual resulta crítico en aplicaciones agrícolas. Este comportamiento es consistente con estudios que señalan que los métodos basados en gradient boosting presentan ventajas significativas en la detección temprana de heladas en contextos agrícolas de altura [24]. Asimismo, el aprendizaje automático aplicado a la predicción meteorológica ofrece una mayor capacidad de generalización en escenarios de alta variabilidad climática [25].

La principal innovación del estudio radica en la integración de un modelo de aprendizaje automático de alto

desempeño con fuentes de datos meteorológicos en tiempo real, aplicado específicamente a zonas altoandinas del Perú. Desde el punto de vista teórico, los mecanismos de regularización y el uso de gradientes de segundo orden propios de XGBoost permiten un mejor control del sobreajuste y una mayor estabilidad del modelo en escenarios complejos [26]. Estudios comparativos desarrollados por Li y Gómez coinciden en que XGBoost ofrece un mejor equilibrio entre sensibilidad y precisión frente a otros algoritmos de clasificación en aplicaciones agrícolas [27], [28]. No obstante, es relevante considerar que la eficiencia del modelo depende de la calidad y resolución de los datos meteorológicos disponibles, por lo que sus resultados deben interpretarse como una herramienta de apoyo a la toma de decisiones. En este sentido, el modelo constituye un soporte operativo para sistemas de alerta temprana orientados a la gestión del riesgo agrícola y la mitigación del impacto de las heladas.

6 CONCLUSIONES

En síntesis, la investigación se desarrolló y evaluó un modelo de predicción de heladas basado en XGBoost centrado en las zonas altoandinas del Perú, por su gran historial climatológico respecto al fenómeno de heladas, manifestando el potencial de los enfoques de aprendizaje automático para abordar fenómenos climáticos complejos y de alta variabilidad espacial. Los resultados obtenidos afirman el enfoque del modelo propuesto cumpliendo de manera satisfactoria el objetivo general de la investigación, logrando un elevado desempeño predictivo, reflejado en una exactitud del 95.17% y un AUC-ROC de 0.9938.

El análisis de importancia de variables que se reveló al transcurso de metodología mostró que las variables termodinámicas locales, en particular la temperatura del aire a 2 m y el punto de rocío, son los factores más determinantes en la ocurrencia de heladas, mientras que el viento, la altura y la precipitación presentan una contribución secundaria. Asimismo, la comparación con modelos tradicionales mostró que XGBoost supera en gran medida a modelos más tradicionales como la regresión logística y Random Forest, lo que refuerza su idoneidad para aplicaciones operativas en climatología agrícola.

Desde una perspectiva aplicada, los hallazgos de este estudio evidencian la viabilidad de implementar sistemas de predicción de heladas en tiempo real mediante el uso de modelos de aprendizaje automático y datos meteorológicos provenientes de APIs abiertas. Esta aproximación demarca una herramienta relevante para la construcción de sistemas de alerta temprana, así como también contribuye a la toma de decisiones en el sector agrícola, con el potencial de reducir las pérdidas productivas asociadas a eventos extremos.

Es necesario decir que, el estudio presenta algunas limitaciones, principalmente relacionadas con la cobertura geográfica del conjunto de datos y la disponibilidad de varia-

bles atmosféricas de mayor resolución. En este sentido, futuras investigaciones deberían ampliar la validación con una muestra más general del País utilizando de modelo a otras regiones no solo altoandinas y explorar la incorporación de nuevas variables, como índices de radiación solar o características orográficas, con el fin de fortalecer la robustez y generalización del sistema propuesto.

En conjunto, este trabajo aporta evidencia empírica de que los modelos basados en XGBoost constituyen una alternativa eficaz y escalable para la predicción de heladas, de igual importancia abriendo nuevas oportunidades para la aplicación de técnicas de inteligencia artificial en la gestión del riesgo climático en regiones vulnerables.

REFERENCIAS

- [1] SENAMHL, «PLATAFORMA DE ESTADO PERUANO,» 2025. [En línea]. Available: <https://www.senamhi.gob.pe/?p=heladas-y-friajes-preguntas>.
- [2] R. R. A. S. Caminada Vallejo, «El eterno retorno del fenómeno de las heladas en el Perú: ¿Existen adecuadas políticas para combatir dicho fenómeno en el Perú?,» *ALICIA*, 2015.
- [3] ComexPerú, «ComexPerú.com,» 5 julio 2024. [En línea]. Available: <https://www.comexperu.org.pe/articulo/mas-de-un-millon-de-hectareas-agricolas-en-riesgo-por-heladas>.
- [4] P. y. R. d. R. d. D. (. Centro Nacional de Estimación, «ESCENARIOS DE RIESGO POR HELADAS Y FRIAJES EN EL MARCO DEL PLAN MULTISECTORIAL 2025 - 2027,» Lima, 2024.
- [5] S. N. d. M. e. H. d. Perú., «Temporada de bajas temperaturas 2024,» 2024. [En línea]. Available: <https://www.gob.pe/institucion/senamhi/campañas/1281-temporada-de-bajas-temperaturas-2022>.
- [6] E. Zegarra, «Heladas y perfil de riesgo de los agricultores peruanos (Working paper),» *GRADE*, 2019.
- [7] L. J. L. V. C. G. C. A. H. D. R. R. O. G.-R. Evelin Zarate Aller, «Efecto de las bajas temperaturas en la selección por tolerancia para progenies híbridas de papa (*Solanum* sp.) con introgresión de *Solanum commersonii*, en la Comunidad de Occopata, Cusco,» *Revista de Investigaciones Altoandinas*, 30 Abril 2023.
- [8] J. A. C. Robinson Castillo Méndez, «Desarrollo de un modelo de machine learning para determinar el comportamiento de heladas en la Sabana de Bogotá,» *Vía Innova*, 2024.
- [9] S. N. d. M. e. H. d. P. Perú., «Atlas de heladas del Perú,» *IDSEEP*, 2010.
- [10] J. & K. H. Lee, «Prediction of Frost Occurrences Using Statistical Modeling Approaches,» *DOAJ*, 2016.
- [11] L. A. S. P. C. Monteiro, «Assessment of NASA/POWER satellite-based weather system for Brazilian conditions and its impact on sugarcane yield simulation,» *International Journal of Climatology*, pp. 1571-1581, 2018.
- [12] T. & G. C. Chen, «XGBoost: A scalable tree boosting system,» *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 785-794, 2016.
- [13] D. R. M. & P. J. Torres, «Performance improvement in meteorological forecasting using optimized XGBoost models,» *Environmental Modelling & Software*, pp. 156-168, 2022.
- [14] J. H. Friedman, «Greedy function approximation: A gradient boosting machine,» *Annals of Statistics*, p. 1189-1232, 2001.
- [15] G. S. K. & W. Y. Zhang, «Evaluation of machine learning methods for meteorological prediction tasks,» *Atmospheric Research*, pp. 105-155, 2021.
- [16] X. L. Y. & Z. M. Qian, «Real-time meteorological prediction using gradient boosting and sensor networks,» *Sensors*, pp. 52-81, 2022.
- [17] M. Á. G. Navarro, «Construcción de un modelo eficiente de predicción de heladas en entornos locales mediante técnicas del análisis inteligente en contextos iot,» *Dialnet*, pp. 18-21, 2021.
- [18] H. & A. S. Aboelkhair, «Assessment of agroclimatology NASA POWER reanalysis datasets for temperature types and relative humidity at 2 m against ground observations over Egypt,» *Advances in Space Research*, pp. 1036-1049, 2019.
- [19] A. P. B. M. J.L. Anco, «Kaggle,» Diciembre 2025. [En línea]. Available: <https://www.kaggle.com/datasets/abadpomama/quera/meteorological-data-for-frost-prediction-in-peru>.
- [20] A. P. B. M. J.L. Anco, «GitHub,» Diciembre 2025. [En línea]. Available: <https://github.com/Abad-cyber/riesgo-heladas-xgboost-peru.git>.
- [21] J. Pérez Tárraga , M. Castillo-Cara , E. Arias-Antúnez y D. Dujovne , «Frost forecasting through machine learning algorithms,» *Earth Science Informatics*, vol. 18, n° 183, pp. 1-16, 2025.
- [22] G. Ovando , M. Bocco y S. Sayago , «Redes neuronales para modelar predicción de heladas,» *Agricultura Técnica (Chile)*, p. 65, 2005.
- [23] T. Chen y C. Guestrin , «XGBoost: A scalable tree boosting system,» *Proceedings of the 22nd ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.

- [24] R. Castillo Méndez y J. A. Camacho Castro , «Predicción de heladas y variables meteorológicas relevantes en agricultura en la Sabana de Bogotá usando machine learning,» *Ingeniería y Desarrollo*, vol. 43, n° 1, pp. 124-139, 2025.
- [25] C. J. Talsma , K. C. Solander , M. K. Mudunuru , B. Crawford y M. R. Powell , «Frost prediction using machine learning and deep neural network models,» *Frontiers in Artificial Intelligence*, vol. 5, n° 963781, pp. 1-16, 2023.
- [26] K.-Q. Li y H.-L. He, «Towards an improved prediction of soil-freezing characteristic curve based on extreme gradient boosting model,» *Geoscience Frontiers*, vol. 15, 2024.
- [27] D. Gómez , G. Araujo , F. E. Martínez , A. O. Rodríguez , J. M. Estupiñán y L. Y. Deantonio , «Early prediction of frost events in high altitude crops, using machine learning methods,» *Revista Ceres*, vol. 71, n° 4, pp. 1-22, 2024.
- [28] H. Li , J. Gao , . Y. Guo y X. G. Yuan , «Application of XGBoost model and multi-source data for winter wheat yield prediction in Henan Province of China,» *Big Data and Information Analytics*, vol. 9, pp. 29-47, 2025.
- [29] G. R. Huancavelica, «gob.pe,» 27 septiembre 2022. [En línea]. Available: <https://www.gob.pe/institucion/regionhuancavelica/noticias/654544-mas-de-7-mil-agricultores-afectados-por-las-heladas-son-beneficiados-con-el-seguro-agrario>.
- [30] D. F. O. Carráscal, «Helada intertropical, sus impactos y mitigación.,» *Revista Ingenio*, pp. 45-60.
- [31] K. S. M. M. B. C. M. P. Carl Talsma, «Predicción de heladas mediante aprendizaje automático y modelos de redes neuronales profundas,» *Frontiers*, 11 Enero 2023.

ANEXOS

Dataset subido a Kaggle:

<https://www.kaggle.com/datasets/abadpomama-quera/meteorological-data-for-frost-prediction-in-peru>

Repositorio en GitHub del modelo y los scripts desarrollados:

<https://github.com/Abad-cyber/riesgo-heladas-xgboost-peru.git>