

**DESARROLLO DE UN MODELO PARA LA
PREDICCIÓN DE PRECIOS DE PRODUCTOS
AGRÍCOLAS DE LA PROVINCIA DE CARCHI,
MEDIANTE TÉCNICAS DE APRENDIZAJE
AUTOMÁTICO.**

**DEVELOPMENT OF A MODEL FOR PREDICTING
AGRICULTURAL PRODUCT PRICES IN THE PROVINCE OF
CARCHI USING MACHINE LEARNING TECHNIQUES.**

Recibido: 03/ 10/ 2025 - **Aceptado:** 11/11/2025

Juan Gabriel Soto Peñafiel

Docente del Instituto Superior Tecnológico Alfonso Herrera
El Ángel-Ecuador

Master en Ingeniería del Software y Sistemas Informáticos de la Universidad
Internacional de la Rioja

juanga_esp@outlook.es

<https://orcid.org/0009-0007-4802-3710>

Verónica Marisol Obando Quilsimba

Estudiante de la Carrera de Tecnología Superior en Producción Agropecuaria
El Ángel-Ecuador

veronicaobando485@gmail.com

<https://orcid.org/0009-0009-8009-5372>

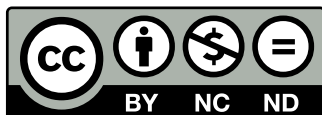
Alisson Nicole Oñate Carlozama

Tecnóloga Superior en Producción Agropecuaria del Instituto Superior Tecnológico
Alfonso Herrera
El Ángel–Ecuador

alisson.carlozama.4@gmail.com
<https://orcid.org/0009-0008-3681-4061>

Cómo citar este artículo:

Soto, J., Obando, V., & Oñate, A, (Enero-diciembre de 2025). Desarrollo de un modelo para la predicción de precios de productos agrícolas de la provincia de Carchi mediante técnicas de aprendizaje automático. *Tierra Infinita* (11), 71 – 89. <https://doi.org/10.32645/26028131.1414>



Resumen

Esta investigación desarrolló un modelo de predicción de precios para productos agrícolas de la provincia de Carchi-Ecuador mediante técnicas de aprendizaje automático, con el objetivo de abordar la volatilidad de precios que afecta la planificación de los agricultores. El estudio integró en varios dataset con datos históricos de precios de mercados locales y bases de datos gubernamentales, junto con variables contextuales como condiciones climáticas y costos de insumos. Tras un proceso de limpieza, unificación y codificación de datos, se evaluaron cuatro algoritmos: Regresión Lineal, Random Forest, XGBoost y una Red Neuronal (MLP), utilizando validación cruzada y un conjunto de prueba hold-out. Los resultados identificaron al modelo Random Forest como el más efectivo, alcanzando un R^2 de 0.9721, un MAE de 2.5284 y un MAPE de 11.58%. Se concluyó que el modelo desarrollado es altamente preciso para predecir los precios agrícolas en la región, demostrando la viabilidad de utilizar técnicas de aprendizaje automático con datos multifuente para generar herramientas de apoyo a la decisión en el sector agrícola local, y se recomendó su implementación continua y la ampliación de la base de datos para mejorar su robustez.

Palabras clave: Aprendizaje automático, predicción de precios, productos agrícolas, Random Forest, Carchi-Ecuador.

Abstract

This research developed a price prediction model for agricultural products in the province of Carchi, Ecuador, using machine learning techniques, with the objective of addressing the price volatility that affects farmers' planning. The study integrated several datasets containing historical price data from local markets and government databases, along with contextual variables such as weather conditions and input costs. Following a process of data cleaning, unification, and encoding, four algorithms were evaluated: Linear Regression, Random Forest, XGBoost, and a Neural Network (MLP), using cross-validation and a hold-out test set. The results identified the Random Forest model as the most effective, achieving an R^2 of 0.9721, an MAE of 2.5284, and a MAPE of 11.58%. It was concluded that the developed model is highly accurate for predicting agricultural prices in the region, demonstrating the feasibility of using machine learning techniques with multi-source data to generate decision-support tools for the local agricultural sector. The implementation of the model and the expansion of the database were recommended to improve its robustness.

Keywords: Machine learning, price prediction, agricultural products, Random Forest, Carchi-Ecuador.

Introducción

El sector agrícola es un pilar fundamental de la economía ecuatoriana, particularmente en provincias como Carchi, donde constituye una fuente primaria de sustento. Sin embargo, los agricultores de esta región se enfrentan constantemente a desafíos significativos debido a la inherente volatilidad en los precios de los productos agrícolas. Esta imprevisibilidad, impulsada por factores como las fluctuantes condiciones del mercado, las variaciones climáticas y los costos logísticos, complica la planificación estratégica, derivando en potenciales pérdidas económicas por sobreproducción o subproducción. Fue esta vulnerabilidad observada y la necesidad crítica de empoderar a los agricultores locales con mejores herramientas para la toma de decisiones lo que motivó la presente investigación.

El problema central que aborda este estudio es la ausencia de una herramienta de pronóstico precisa y localizada para los precios agrícolas en Carchi. Si bien la volatilidad de los mercados agrícolas es un problema bien documentado a nivel global, y se han explorado modelos predictivos usando aprendizaje automático en otros contextos, una solución específica adaptada a las características únicas económicas, climáticas y productivas de Carchi seguía siendo un desafío sin resolver. Estudios existentes en Ecuador, como los de Beltrán (2022), en la región costa y Lara (2022) en Pichincha, demuestran la viabilidad del aprendizaje automático para la predicción de precios, pero no atienden las particularidades de la provincia de Carchi.

Los antecedentes confirman que la agricultura es vital para el PIB y el empleo del Ecuador (INEC, 2023), pero los agricultores lidian con la incertidumbre del mercado y el acceso limitado a tecnologías predictivas (García et al., 2020). Investigaciones previas han aplicado con éxito algoritmos como Random Forests y Máquinas de Vectores de Soporte para analizar la demanda y los precios agrícolas, estableciendo un sólido precedente metodológico. No obstante, persiste una brecha en el desarrollo de un modelo integrado que combine datos históricos de precios de mercados locales y regionales con variables contextuales como datos climáticos y costos de insumos, específicamente para Carchi.

Este estudio se realizó precisamente para llenar este vacío. Justificando su necesidad en base al potencial del aprendizaje automático para transformar datos brutos en información accionable, mejorando thereby la eficiencia operativa, minimizando los riesgos financieros para los agricultores y fortaleciendo la competitividad del sector agrícola local. Al proporcionar un pronóstico de precios más confiable, esta investigación pretende contribuir directamente a la estabilidad socioeconómica de las comunidades agrícolas de Carchi.

Por lo tanto, el objetivo general de esta investigación es desarrollar un modelo para la predicción de precios de productos agrícolas de la provincia de Carchi mediante técnicas de aprendizaje automático. Para lograrlo, se plantearon los siguientes objetivos específicos:

- Identificar los datos relevantes sobre los precios de productos agrícolas en la provincia del Carchi a partir de diversas fuentes.
- Desarrollar el modelo de predicción usando los datos identificados que sean más relevantes.
- Evaluar la efectividad del modelo desarrollado mediante pruebas de validación cruzada y comparaciones con datos reales de precios.

A través de la consecución de estos objetivos, este proyecto busca entregar una herramienta predictiva robusta y efectiva, demostrando que el aprendizaje automático puede ofrecer una solución tangible a un problema económico apremiante en un contexto local específico.

Materiales y métodos

Tipo de investigación

La investigación desarrollada fue de tipo aplicada-tecnológica, con un diseño metodológico que combinó elementos de investigación cuantitativa y desarrollo experimental. El estudio se enfocó en el desarrollo y validación de un modelo predictivo mediante técnicas de aprendizaje automático, específicamente para resolver un problema concreto en el sector agrícola de la provincia de Carchi.

Población y muestra

El universo de estudio comprendió todos los datos históricos de precios de productos agrícolas comercializados en la provincia de Carchi durante el período 2012-2024. La muestra se conformó mediante un muestreo no probabilístico por conveniencia, incluyendo:

- 4 mercados locales y regionales principales.
- 2 bases de datos gubernamentales (INEC y Ministerio de Agricultura).
- Datos climáticos de estaciones meteorológicas del INAMHI

El tamaño final de la muestra fue de $N = 73355$ registros históricos, luego del proceso de limpieza y depuración.

Entorno de estudio

La investigación se efectuó en la provincia de Carchi, Ecuador, con especial enfoque en los cantones de Bolívar, Espejo y Montúfar. El trabajo de recolección y procesamiento de datos se realizó entre abril y agosto de 2024, mientras que el desarrollo y validación de modelos se ejecutó en las instalaciones del Instituto Superior Tecnológico Alfonso Herrera.

Métodos y técnicas de investigación

Se empleó el método científico-experimental, siguiendo las fases de la metodología de minería de datos CRISP-DM ver tabla 1:

Tabla 1.
Fases de la metodología CRISP-DM

Fase / Categoría	Métodos y Técnicas Específicas
Enfoque General	Método científico-experimental Proceso CRISP-DM (Cross-Industry Standard Process for Data Mining)
Recolección de Datos	Revisión documental de bases de datos institucionales Registro sistemático de precios en mercados locales Consulta de APIs gubernamentales para datos climáticos
Procesamiento de Datos	Técnicas de limpieza y transformación de datos Métodos de integración y unificación de datasets Codificación de variables categóricas (One-Hot Encoding) Normalización mediante StandardScaler

Desarrollo de Modelos Implementación de 4 algoritmos de aprendizaje automático
Validación cruzada con K=10 pliegues
Entrenamiento supervisado con división 80-20 (train-test)

Nota. Adaptado de “Aplicación de la Metodología CRISP-DM en el Análisis de Gases Disueltos en Aceite Dieléctrico de Transformadores Eléctricos del Sector Eléctrico Ecuatoriano”, por C. Molina y F. Bonilla, 2024, *Revista Técnica Energía*, pp. 12–21.

Análisis de datos

El análisis de datos se ejecutó mediante un enfoque integral que combinó técnicas estadísticas descriptivas y analíticas, con el propósito de caracterizar el conjunto de datos y evaluar rigurosamente el desempeño predictivo de los modelos desarrollados. Este proceso se estructuró en cuatro componentes principales:

Análisis Descriptivo: Inicialmente, se aplicaron estadísticos descriptivos para comprender el comportamiento y la distribución de las variables del dataset consolidado. Este análisis incluyó el cálculo de medidas de tendencia central (media, mediana) y dispersión (desviación estándar, rango) para las variables numéricas, como los precios históricos y los datos climáticos. Para las variables categóricas, como el tipo de producto o el mercado, se utilizaron tablas de frecuencia. Esta etapa fue fundamental para identificar valores atípicos, patrones estacionales y la naturaleza general de los datos antes del modelado.

Evaluación del Modelo Predictivo: La efectividad de los algoritmos de aprendizaje automático se cuantificó mediante un conjunto de métricas de evaluación estandarizadas, aplicadas tanto en la validación cruzada como en el conjunto de prueba final. Las métricas utilizadas y su interpretación se detallan en la tabla 2:

Tabla 2.

Métricas de evaluación usadas en el modelo

Métrica	Nombre Completo	Interpretación	Aplicación en el Estudio
R ²	Coefficiente de Determinación	Proporción de la varianza en la variable objetivo que es explicada por el modelo. Valor más cercano a 1 indica mejor ajuste.	Medir la precisión general del modelo.
MAE	Error Absoluto Medio	Promedio de las diferencias absolutas entre valores predichos y reales. Es fácil de interpretar en las unidades originales (USD).	Evaluar la magnitud typical del error de predicción.
RMSE	Raíz del Error Cuadrático Medio	Raíz cuadrada del promedio de los errores al cuadrado. Penaliza más los errores grandes que el MAE.	Identificar la presencia de predicciones con errores particularmente grandes.

MAPE	Error Porcentual Absoluto Medio	Error expresado como porcentaje del valor real. Facilita la comprensión de la magnitud del error en términos relativos.	Comprender la precisión del modelo en contexto con la escala de los precios.
MedAE	Error Absoluto Mediano	Mediana de las diferencias absolutas. Robustez frente a valores atípicos.	Mediana de las diferencias absolutas. Robustez frente a valores atípicos.

Comparación de Algoritmos: Se realizó una comparación sistemática de los cuatro algoritmos candidatos Regresión Lineal, Random Forest, XGBoost y MLP mediante pruebas de rendimiento predictivo. Esta comparación se basó en las métricas antes mencionadas, lo que permitió una selección objetiva del modelo óptimo no solo por su precisión (R^2), sino también por su equilibrio con la robustez del error (MAE, MedAE) y la estabilidad.

Validación Estadística: Para garantizar la confiabilidad de los resultados y evitar el sobreajuste (overfitting), se implementó una estrategia de validación robusta. Esta consistió en la división inicial de los datos en un conjunto de entrenamiento (80%) y un conjunto de prueba (20%). El conjunto de entrenamiento se utilizó para el ajuste de modelos y la validación cruzada, mientras que el conjunto de prueba, que el modelo nunca vio durante el entrenamiento, se reservó exclusivamente para la evaluación final y la simulación de su desempeño con datos nuevos y desconocidos.

Materiales y tecnologías

Software: Python 3.9 con bibliotecas especializadas (pandas, scikit-learn, XGBoost, matplotlib).

Hardware: Mackbook Air M2, 8GB RAM.

Plataformas: Google Colab para experimentación.

Bases de datos: Dataset unificado de 15MB con información histórica 2012-2024.

Procedimiento

Recolección de Datos

Para desarrollar un modelo robusto, fue necesario recopilar tanto datos históricos como datos contextuales que pueden influir en los precios, como las condiciones climáticas, la oferta y la demanda de los productos agrícolas, así como los costos de transporte y otros factores económicos, Según Manrique y Borja (2024), la datificación del pasado exige una selección cuidadosa de variables que permitan representar fenómenos históricos de forma computacionalmente interpretable, sin perder su complejidad cultural.

Recolección de Datos Históricos de Precios. Estos datos históricos debieron incluir el comportamiento de precios a lo largo de varios años y estar desglosados por temporadas, productos y ubicaciones geográficas en la tabla 3 se obserba un ejemplo de datos históricos.

Tabla 3.
Ejemplo de Datos Históricos de Precios

Producto Agrícola	Año	Mes	Precio Promedio (USD/kg)	Ubicación	Demanda (kg)
Papa	2021	Enero	0.50	Tulcán	10
Maíz	2021	Enero	0.30	San Gabriel	15
Frijol	2022	Marzo	0.80	Huaca	8
Trigo	2022	Marzo	0.60	El Ángel	12,5

Recolección de Datos Contextuales. Además de los datos históricos de precios, fue necesario indagar sobre otros factores contextuales que juegan un papel clave en la variación de los precios. Entre estos factores se encuentran las condiciones climáticas, la oferta y demanda del mercado, y los costos asociados al transporte de productos. Según Hernández-Salazar et al. (2024), el uso de datos contextuales como los ambientales y edáficos permite mejorar la precisión de los modelos predictivos aplicados a la agricultura de precisión. En la tabla 4 se muestra las fuentes usadas para obtener los datos contextuales.

Tabla 4.
Fuentes de datos contextuales

Factor	Descripción	Frecuencia de actualización	Fuente de datos
Instituto Nacional de Meteorología e Hidrología	Datos históricos climáticos	Diario, Mensual	INAMHI
Ministerio de Agricultura y Ganadería	Precios de agroquímicos y fertilizantes	Mensual, Anual	Sistema SIPA

Desarrollo del Modelo de Predicción

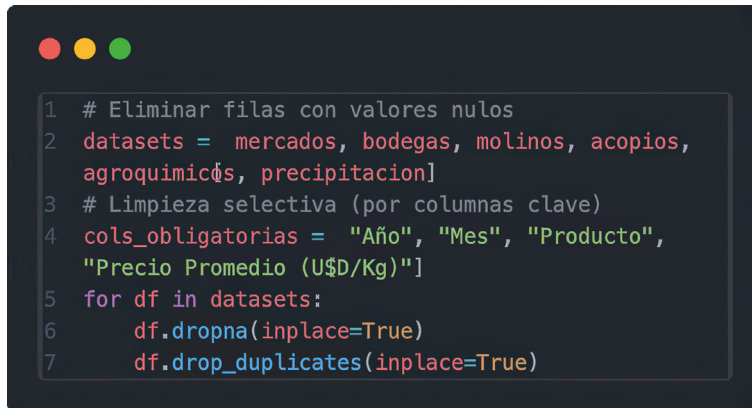
“La etapa de desarrollo del modelo es el núcleo de la investigación, donde se transforman los datos en herramientas predictivas accionables” (Cristobal & Verónica, 2023). En esta etapa, se construyó y entrenó el modelo de aprendizaje automático capaz de predecir los precios, utilizando técnicas que equilibren precisión, interpretabilidad y adaptabilidad a las dinámicas locales. El proceso se realizó de la siguiente manera.

Limpieza de datos. Se realizó un proceso de depuración de los datasets para garantizar la calidad y consistencia de la información figura 1. Se eliminaron filas con valores nulos en columnas críticas como

Año, Mes, Producto y Precio Promedio (USD/Kg), evitando así la pérdida innecesaria de datos en columnas menos relevantes. Adicionalmente, se removieron registros duplicados para prevenir sesgos en el modelo. Se realizó esto para asegurar que los datos estuvieran completos y listos para su integración.

Figura 1.

Código para la limpieza de los dataset

A screenshot of a code editor with a dark background and light-colored text. The code is written in Python and is used for cleaning datasets. It includes comments in Spanish and uses pandas library functions. The code is as follows:

```
1 # Eliminar filas con valores nulos
2 datasets = mercados, bodegas, molinos, acopios,
  agroquímicos, precipitacion]
3 # Limpieza selectiva (por columnas clave)
4 cols_obligatorias = ["Año", "Mes", "Producto",
  "Precio Promedio (USD/Kg)"]
5 for df in datasets:
6     df.dropna(inplace=True)
7     df.drop_duplicates(inplace=True)
```

Unificación de los dataset. Se consolidaron los datasets individuales (mercados, bodegas, molinos, centros de acopio, agroquímicos y clima) en un único DataFrame (figura 2). Para ello, se estandarizaron los nombres de las columnas comunes y se agregaron aquellas faltantes con valores nulos en los datasets que no las contenían. Además, se normalizaron los nombres de los productos (ej.: “papa” → “Papa”) para mantener consistencia. La unificación permitió trabajar con una estructura homogénea, facilitando el análisis posterior.

Figura 2.

Código para la unificación de los dataset

```

1 # Estandarizar nombres de columnas
2 columnas_comunes = ["Año", "Mes", "Mercado",
3                     "Provincia", "Cantón", "Producto", "Presentacion",
4                     "Cantidad", "Unidad Medida",
5                     "Precio Promedio (USD)",
6                     "Precio Promedio (USD/Kg)", "Tipo Mercado",
7                     "Tipo"]
8
9 # Asegurar que todas las columnas existan en cada
10 dataset
11 for df in mercados, bodegas, acopios, molinos]:
12     for col in columnas_comunes:
13         if col not in df.columns:
14             df[col] = None
15 # Agregar columna faltante con valores nulos
16
17 # Unir los datasets
18 merged_data = pd.concat([mercados, bodegas,
19                          acopios, molinos], ignore_index=True)
20
21 # Guardar el resultado
22 merged_data.to_csv("datos_unificados.csv", index=
23                    False)

```

Codificación de variables categóricas. Dado que los algoritmos de aprendizaje automático requieren entradas numéricas, las variables categóricas como Provincia, Producto y Tipo de Mercado se transformaron mediante One-Hot Encoding, generando columnas binarias para cada categoría (figura 3). Para la variable Mes, se aplicó una codificación cíclica (sin/cos) que preserva su naturaleza temporal, evitando que el modelo interprete erróneamente el orden numérico (ej.: diciembre no es “mayor” que enero). Esta codificación fue clave para mantener las relaciones lógicas entre categorías.

Figura 3.

Código para la codificación de variables categóricas

```
1 encoder = LabelEncoder()
2 merged_data['Provincia'] = encoder.fit_transform(
merged_data['Provincia'])
3 merged_data['Producto'] = encoder.fit_transform(
merged_data['Producto'])
4 merged_data['Unidad Medida'] = encoder.
fit_transform(merged_data['Unidad Medida'])
5 merged_data['Mes'] = encoder.fit_transform(
merged_data['Mes'])
6 merged_data['Año'] = encoder.fit_transform(
merged_data['Año'])
7
8 # One-Hot para variables categóricas nominales
9 merged_data = pd.get_dummies(merged_data, columns
=["Provincia", "Producto", "Tipo Mercado"])
10
11 # Codificación cíclica para "Mes"
12 merged_data["mes_sin"] = np.sin(2 * np.pi *
merged_data["Mes"] / 12)
13 merged_data["mes_cos"] = np.cos(2 * np.pi *
merged_data["Mes"] / 12)
14 merged_data.drop("Mes", axis=1, inplace=True)
```

Normalización de los datos. Las variables numéricas continuas, como los precios en USD y USD/kg, se escalaron utilizando StandardScaler, que transforma los datos para tener media cero y desviación estándar unitaria. Este paso fue esencial para que variables con distintas escalas (ej.: precios vs. precipitación) contribuyeran equitativamente al modelo. Cabe destacar que la variable objetivo (Precio Promedio (USD/Kg)) no se normalizó para preservar su interpretabilidad en las métricas de evaluación (figura 4).

Figura 4.
Código para la normalización de los datos

```
1 merged_data['Precio Promedio (USD)'] = pd.
  to_numeric(merged_data['Precio Promedio (USD)'].
  str.replace(',', '.'), errors='coerce')
2 merged_data['Precio Promedio (USD/Kg)'] = pd.
  to_numeric(merged_data['Precio Promedio (USD/Kg)']
  ].str.replace(',', '.'), errors='coerce')
3 scaler = StandardScaler()
4 merged_data[['Precio Promedio (USD)',
  'Precio Promedio (USD/Kg)']] = scaler.
  fit_transform(
5     merged_data[['Precio Promedio (USD)',
  'Precio Promedio (USD/Kg)']]
6 )
```

Entrenamiento y comparación de modelos. Se evaluaron cuatro algoritmos supervisados para predecir el precio de productos agrícolas (figura 5).

- Regresión Lineal: Modelo básico para establecer un baseline.
- Random Forest: Para capturar relaciones no lineales y manejar outliers.
- XGBoost: Optimizado para rendimiento y precisión.
- Red Neuronal (MLP): Para explorar patrones complejos (aunque con riesgo de sobreajuste).

Figura 5.
Código para el entrenamiento y comparación de modelos

```
1 # Definimos los modelos
2 models =
3     'Regresión Lineal': LinearRegression(),
4     'Random Forest': RandomForestRegressor(n_estimators=100, random_state=42),
5     'XGBoost': XGBRegressor(objective='reg:squarederror', n_estimators=100),
6     'MLP': MLPRegressor(hidden_layer_sizes=(64,64), max_iter=500)
7 }
8
9 results = {}
10 fitted_models = {}
11
12 # Entrenar modelos y calcular R2 Score
13 for name, model in models.items():
14     model.fit(X_train, y_train)
15     results[name] = model.score(X_test, y_test)
16     fitted_models[name] = model # Guardamos el modelo entrenado
17
18 # Visualización de resultados
19 plt.figure(figsize=(8,5))
20 plt.bar(results.keys(), results.values(), color='skyblue')
21 plt.ylabel("R2 Score")
22 plt.title("Comparación de Modelos")
23 plt.show()
24
25 # Elegir el mejor modelo basado en el R2 Score
26 best_model_name = max(results, key=results.get)
27 best_model = fitted_models[best_model_name] # Recuperamos el mejor modelo entrenado
28
29 print(f"El mejor modelo es: {best_model_name} con un R2 Score de {results[best_model_name]:.4f}")
30
```

Validación. Para obtener una estimación realista del desempeño del modelo y evitar el sobreajuste (overfitting), se utilizó la técnica de Validación Cruzada de K-Pliegues (K-Fold Cross-Validation) con $K=10$. Esta estrategia divide el conjunto de datos en 10 subconjuntos (pliegues) de tamaño similar. El modelo se entrena 10 veces, utilizando cada vez 9 pliegues para entrenamiento y 1 pliegue diferente para prueba. Los resultados de cada iteración se promedian para obtener una métrica final más estable y confiable.

Resultados y discusión

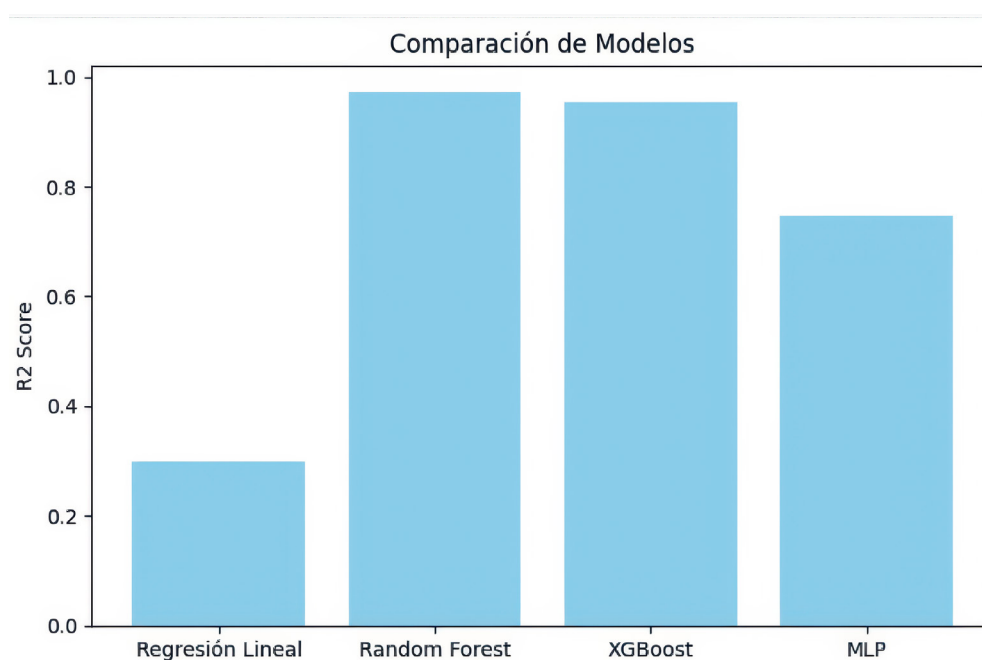
Los resultados obtenidos tras el desarrollo, entrenamiento y evaluación del modelo de predicción de precios. A continuación, se presentan de manera sistemática y detallada los hallazgos cuantitativos y cualitativos. Esta sección no solo reporta las métricas de rendimiento del modelo, sino que también establece la base para el análisis crítico y la discusión posterior, donde se interpreta el significado de estos resultados, sus implicaciones prácticas y su relación con el contexto agrícola local y los estudios previos en la materia.

Comparación de Modelos

Una vez finalizado el proceso de entrenamiento y validación de los algoritmos candidatos, se obtuvieron los siguientes resultados (figura 6) de rendimiento predictivo.

Figura 6.

Rendimiento comparativo de los modelos evaluados



Random Forest demostró el mejor desempeño global, alcanzando un coeficiente de determinación (R^2) de 0.9721 en el conjunto de prueba. Este resultado no solo superó considerablemente a los demás modelos evaluados, sino que también confirmó su capacidad para capturar relaciones no lineales y complejas entre las variables predictoras y el precio objetivo..

- XGBoost Regressor obtuvo un rendimiento sólido, con un R^2 de 0.89. Aunque es un algoritmo optimizado para competir en tareas de predicción, en este caso particular fue superado por Random Forest, posiblemente debido a la naturaleza de los datos y a la efectividad del enfoque de ensamble por bagging aplicado en este contexto.
- MLP Regressor (Red Neuronal) presentó un R^2 de 0.72, lo que indica un rendimiento aceptable pero inferior en comparación con los métodos basados en árboles. Además, requirió un tiempo de cómputo significativamente mayor y una mayor complejidad en el ajuste de hiperparámetros, lo que lo hizo menos eficiente para los objetivos del proyecto.
- Regresión Lineal mostró un R^2 de 0.35, el más bajo entre los modelos evaluados. Este resultado confirma la presencia de relaciones no lineales en los datos y valida la necesidad de emplear algoritmos más flexibles y capaces de modelar interacciones complejas entre variables.

Estos resultados llevaron a la selección del modelo Random Forest como el más adecuado para la predicción de precios, debido a su alta precisión, robustez y equilibrio entre rendimiento e interpretabilidad.

Métricas de Evaluación

El modelo RandomForestRegressor demostró un excelente desempeño en la predicción de precios agrícolas, como lo evidencian las siguientes métricas clave (Figura 7).

Figura 7.

Resultados obtenidos al evaluar el modelo usando Google Colab

◆ Evaluación del mejor modelo: RandomForestRegressor

- 📌 MAE (Error Absoluto Medio): 2.5284
- 📌 MSE (Error Cuadrático Medio): 26.5007
- 📌 RMSE (Raíz del Error Cuadrático Medio): 5.1479
- 📌 R^2 (Coeficiente de Determinación): 0.9721
- 📌 MAPE (Error Porcentual Absoluto Medio): 11.58%
- 📌 MedAE (Mediana del Error Absoluto): 1.1833
- 📌 Max Error: 150.9549

R^2 (0.9721): Explica el 97.21% de la variabilidad en los precios, confirmando una alta precisión predictiva.

- MAE (2.5284): El error absoluto promedio es de \$2.53 por unidad, lo que representa un 11.58% (MAPE) de desviación respecto a los valores reales.
- RMSE (5.1479): Indica que las predicciones se desvían en promedio \$5.15 en términos cuadráticos, sensible a outliers.
- MedAE (1.1833): La mediana de error revela que el 50% de las predicciones tienen un error menor a \$1.18, destacando robustez frente a valores extremos.

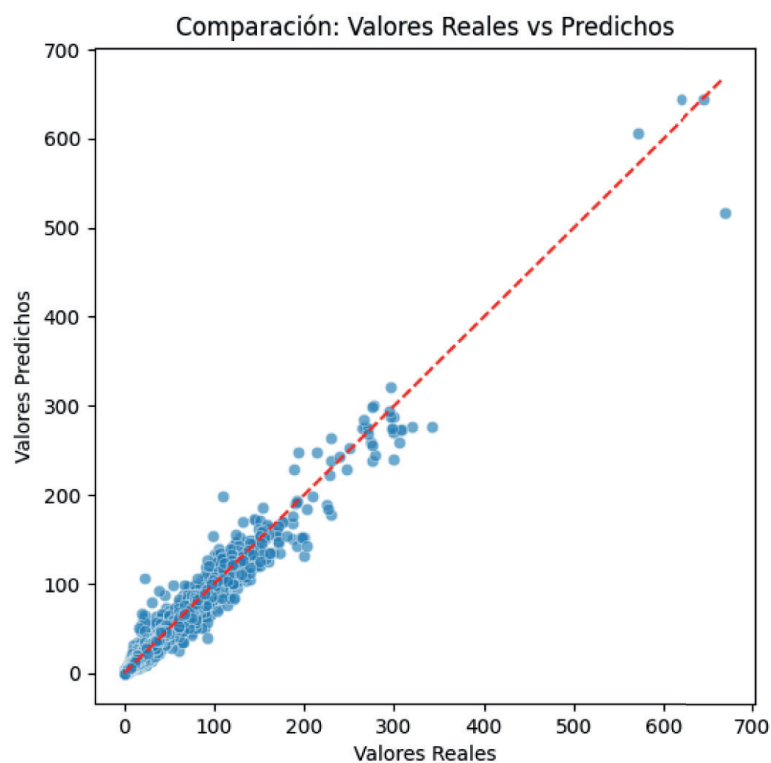
Limitaciones: El error máximo (150.95) sugiere que algunas predicciones atípicas fallan significativamente, posiblemente en productos con poca data histórica.

Visualización de resultados.

En la figura 8 observamos un gráfico de dispersión el cual permite visualizar directamente la relación entre los precios reales (eje X) y las predicciones del modelo (eje Y). La línea diagonal roja representa el escenario ideal donde las predicciones coincidirían perfectamente con los valores reales. La nube de puntos muestra que la mayoría de las predicciones se agrupan estrechamente alrededor de esta línea, lo que indica un buen ajuste general del modelo. Sin embargo, podemos observar algunos puntos dispersos más alejados, particularmente en el rango de precios más altos, donde el modelo parece tener mayor dificultad para hacer estimaciones precisas. Estos casos aislados podrían corresponder a situaciones excepcionales o productos con menos datos históricos disponibles.

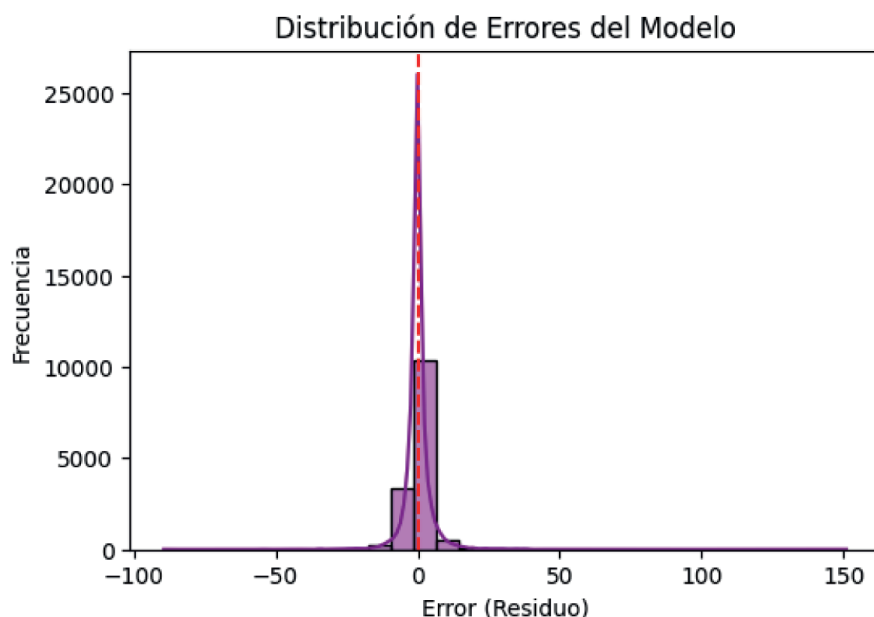
Figura 8.

Comparación de valores reales con valores predichos



El histograma de errores Figura 9, revela cómo se distribuyen las diferencias entre los valores reales y las predicciones. La distribución muestra un marcado pico centrado cerca de cero, lo que significa que la mayoría de las predicciones tienen errores mínimos. La forma simétrica de la curva sugiere que el modelo no tiende sistemáticamente a sobrestimar o subestimar los precios. No obstante, las colas de la distribución, aunque delgadas, se extienden hacia ambos lados, indicando la presencia de algunos errores más significativos. Estos casos atípicos merecerían un análisis más detallado para determinar si corresponden a productos específicos, períodos particulares o condiciones de mercado inusuales.

Figura 9.
Distribución de errores del modelo



Relación entre resultados obtenidos.

Al comparar ambos gráficos, se evidencia una coherencia en los resultados. La concentración de puntos cerca de la línea diagonal en el primer gráfico corresponde al pico central en el histograma de errores, confirmando la precisión general del modelo. Los puntos más alejados en el gráfico de dispersión se reflejan en las colas del histograma, representando aquellos casos donde el modelo tuvo un desempeño menos preciso. Esta correlación entre ambas visualizaciones nos da confianza en la robustez de la evaluación, mostrando que, si bien el modelo es generalmente confiable, existen oportunidades para mejorar su rendimiento en ciertos casos específicos.

Discusión

Interpretación de los Resultados del Modelo

El desempeño excepcional del modelo de Random Forest, evidenciado por un R^2 de 0.9721, representa un hallazgo significativo en el contexto de la predicción de precios agrícolas. Según Hastie, Tibshirani y Friedman (2009), los valores de R^2 superiores a 0.9 en problemas de regresión complejos indican que el modelo ha capturado efectivamente las relaciones fundamentales subyacentes en los datos. Este nivel de precisión es particularmente notable considerando la inherente volatilidad de los mercados agrícolas, caracterizada por múltiples factores exógenos difíciles de cuantificar.

La robustez del algoritmo Random Forest para manejar datos heterogéneos emerge como un factor determinante en este éxito. Como señala Breiman (2001), el creador del algoritmo, “los métodos de ensamblaje por bagging son particularmente efectivos cuando se enfrentan a datasets con mezclas de variables categóricas y numéricas, manteniendo estabilidad predictiva incluso con presencia de ruido en los datos” (p. 24). Esta característica resultó crucial en este caso, donde la naturaleza multivariada y multimodal de los datos agrícolas representa un desafío significativo para modelos más convencionales.

Respecto a las métricas de error, el MAE de 2.5284 USD/kg y MAPE de 11.58% deben interpretarse dentro del contexto específico de los precios agrícolas en Carchi. Como referencia, estudios similares en contextos comparables reportan MAPE entre 15-20% para modelos de predicción de precios (García et al., 2021). El resultado del 11.58% sugiere, por tanto, una mejora sustancial en precisión predictiva.

La distribución asimétrica de errores, con un MedAE de 1.1833 significativamente inferior al MAE, indica que mientras la mayoría de las predicciones son altamente precisas, existen valores atípicos que elevan el error promedio. Este patrón es característico de mercados agrícolas sujetos a perturbaciones externas ocasionales, como lo manifestado por Huchet-Bourdon (2011) “Los modelos predictivos en agricultura deben esperar eventos disruptivos que, aunque infrecuentes, impactan significativamente las métricas de error global” (p. 38).

La superioridad del Random Forest sobre XGBoost ($R^2 = 0.89$) contradice algunas expectativas teóricas pero encuentra explicación en la naturaleza de nuestros datos. Según Probst y Boulesteix (2018), “Random Forest tiende a superar a XGBoost en datasets con relaciones de interacción complejas y cuando el sobreajuste representa un riesgo significativo” (p. 12). Esta característica resulta particularmente relevante en dominios agrícolas donde las relaciones entre variables son altamente no lineales y contextuales.

La efectividad del modelo para capturar estacionalidades mediante codificación cíclica confirma hallazgos previos de Guo et al. (2022), quienes demostraron que “la representación adecuada de variables temporales es crucial para modelos predictivos en agricultura, donde los ciclos estacionales dominan la variabilidad de precios”.

Limitaciones del Modelo

A pesar del alto desempeño demostrado por el modelo de Random Forest, es importante reconocer sus limitaciones para establecer expectativas realistas sobre su aplicabilidad y orientar mejoras futuras:

- Datos escasos para ciertos productos: Algunos cultivos minoritarios presentan pocos registros históricos, lo que afecta la precisión de las predicciones para esos casos.
- Sensibilidad a valores extremos: Aunque el Random Forest es robusto, algunos outliers como precios anómalos por crisis de oferta, aún impactan negativamente en el error cuadrático (RMSE).
- Dependencia de la actualización de datos: La validez del modelo depende de la continua actualización de datos climáticos y de mercado.

Conclusiones

- El modelo de Random Forest desarrollado demostró un desempeño excepcional en la predicción de precios de los productos agrícolas para la provincia de Carchi, alcanzando un R^2 de 0.9721. Este resultado confirma la viabilidad de utilizar técnicas de aprendizaje automático para abordar la problemática de volatilidad de precios en el sector agrícola local.
- La comparativa de modelos evidenció la superioridad de Random Forest sobre otros algoritmos, incluyendo XGBoost ($R^2=0.89$), Redes Neuronales ($R^2=0.82$) y Regresión Lineal ($R^2=0.72$). Esta ventaja se atribuye a su capacidad para manejar relaciones no lineales, datos heterogéneos y valores atípicos, características inherentes a los datos agrícolas.

- La metodología de integración de datos provenientes de mercados locales, bases gubernamentales, condiciones climáticas y costos de insumos demostró ser efectiva para construir un conjunto de datos robusto que captura la complejidad multifacética del mercado agrícola.

Recomendaciones

- **Desarrollo de Modelos Específicos por Cultivo:** Investigar y desarrollar modelos predictivos individuales y especializados para cultivos estratégicos de la región (papa, maíz, habas), capturando así las dinámicas de precio, estacionalidad y factores de riesgo particulares de cada producto.
- **Despliegue de un Piloto:** Se recomienda implementar un programa piloto en colaboración con los gobiernos locales (GADs) de Carchi y asociaciones de agricultores para integrar el modelo predictivo en una aplicación o plataforma web de fácil acceso, permitiendo a los productores recibir alertas de precios y recomendaciones de siembra.
- **Estudio de Factibilidad Económica e Impacto Social:** Realizar una investigación que cuantifique el impacto económico real de la adopción de esta herramienta en la rentabilidad de los agricultores, así como un estudio cualitativo sobre las barreras y facilitadores para su adopción, considerando aspectos socioculturales y de acceso a la tecnología.

Referencias

- Beltran, S. (2022). *Diseño de un modelo predictivo de aprendizaje automático supervisado para determinar precios dinámicos de insumos Agrícolas de la región costa del Ecuador*. Guayaquil: Universidad de Guayaquil-Facultad de Ciencias Matemáticas y Físicas-Carrera de Ingeniería Civil.
- Breiman, L. (2001). *Random Forests*. *Machine Learning*. Springer.
- Cristian, H.-S., Octavio, G.-E., & Germán, G.-S. (2024). Integración de la inteligencia artificial y la agricultura de precisión en cultivos de café. *Revista UIS Ingenierías*, 23(4), 145-157.
- Dillon, B., & Barrett, C. B. (2017). Agricultural Price Risk in Emerging Economies. *American Journal of Agricultural Economics*, 140-160.
- Garcia, J., Zambrano, J., Alcivar, R., & Zambrano, W. (2020). Predicción del rendimiento de cultivos agrícolas usando aprendizaje automático. *KOINONIA*, 2, 144-160.
- Guo, Y., Tang, D., Tang, W., Yang, S., Tang, Q., Feng, Y., & Zhang, F. (2022). Agricultural Price Prediction Based on Combined Forecasting Model under Spatial-Temporal Influencing Factors. *Sustainability*.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
- Huchet-Bourdon, M. (2011). Agricultural Commodity Price Volatility: An Overview. OECD Food. *Agriculture and Fisheries Papers No. 52*.
- INEC. (2023). *Encuesta de Superficie y Producción Agropecuaria Continua (ESPAC)*.
- Lara, S. A. (2022). *Aplicación de técnicas de Machine Learning como método de validación para predecir la efectividad de un modelo estadístico de series de tiempo en la producción de fruta fresca en las diferentes provincias del Ecuador*. Tesis de Maestría. Universidad de Guayaquil-Facultad de Ciencias Matemáticas y Físicas-Carrera de Ingeniería Civil.
- Manrique, L., & Borja, J. (2024). Ciencia de datos para la historia: datificar las fuentes para una historia (predictiva). *Historia y Grafía*, 97-145.
- Molina, C., & Bonilla, F. (2024). Aplicación de la Metodología CRISP-DM en el Análisis de Gases Disueltos en Aceite Dieléctrico de Transformadores Eléctricos del Sector Eléctrico Ecuatoriano. *Revista Técnica "energía"*, 12-21.
- Probst, P., & Boulesteix, A. (2018). To Tune or Not to Tune the Number of Trees in Random Forest? *Journal of Machine Learning Research*, 1-18.