“Modelo de Machine Learning para la predicción del precio del limón en base a niveles de producción, desastres naturales y las condiciones climáticas asociadas de la región de Piura, Perú”

Autor/Autores: Sanchez Boza, R (20221221); Lau Lu, O (20220884); Sanchez Pastor, R (20201230);Diaz Meoño, F (20222110); Tenipuclla Flores, G (20224918)

Resumen - En este proyecto, se propone el desarrollo un modelo de Machine Learning para predecir el precio del limón producido en la región de Piura, Perú. Los principales datos empleados en el proyecto son reportes de los precios del Gran Mercado Mayorista de Lima, reportes de desastres naturales y niveles de producción del insumo. Con el fin de presentar las mejores estimaciones se analizan diversos algoritmos de aprendizaje supervisado, realizando la comparación y la decisión del mejor modelo. Este trabajo busca contribuir a la toma de decisiones en el sector agrícola y comercial, facilitando la planificación de precios y estrategias de producción ante la alta volatilidad del mercado del limón.

1. Introducción

De acuerdo con el reporte “Situación del limón en el Perú, un análisis económico” desarrollado por el Ministerio de Desarrollo Agrario y Riego en marzo de 2025, la producción de limón de la región de Piura representa una participación de aproximadamente 63% de la producción total nacional del limón. Esta cifra se tradujo en el 2024 en una producción neta de 236,272 toneladas de limón y un ingreso de 226,821,120 soles para la región. Sin embargo, a pesar de la importancia de este insumo dentro del consumo nacional, tiene precios extremadamente variables en el tiempo. Esto afecta primordialmente a los ingresos de los agricultores de la región y los precios de productos derivados o empleando este insumo.

Dentro de los principales factores que afectan esta producción debemos enfatizar la importancia de las condiciones ambientales. De acuerdo con la “Guía para la implementación de buenas prácticas agrícolas (BPA) para el cultivo de limón” elaborado por Servicio Nacional de Sanidad Agraria del Perú, se identifica que el cultivo de limón debe realizarse entre los 23 °C y 30 °C y considerando un nivel de riego de apropiado entre 16,000 y 18,000 m3/Ha. Con esta información, se puede identificar desastres naturales y fenómenos meteorológicos como “El Niño”, “La Niña” o anticiclones pueden tener un efecto directo sobre variables cruciales para la calidad del cultivo del limón.

Con todo lo establecido previamente, podemos plantear como hipótesis principal que los efectos de fenómenos naturales sobre las condiciones climáticas sobre las áreas de producción principal de la región Piura en conjunto con conocimiento pasado de la variabilidad temporal de los precios van a permitir una predicción futura eficiente del precio del limón utilizando como referencia el Gran Mercado Mayorista de Lima.

Por ello el principal objetivo del presente proyecto es la elaboración, calibración y evaluación de diversos modelos de regresión para hacer la predicción diaria del precio del limón. Este modelo busca servir como herramienta de apoyo estratégico para comerciantes mayoristas, intermediarios, empresas agroindustriales y negocios que utilizan el limón como insumo principal, permitiéndoles anticipar la volatilidad de precios, optimizar decisiones de compra y ajustar sus estrategias comerciales ante fluctuaciones del mercado.

2. Trabajos relacionados

Food prices and production in the aftermath of natural disasters: the case of Peru

Esta investigación explica la relación relevante entre los factores meteorológicos y los desastres naturales, principalmente lluvias y sequías, comparándolos con datos de precios proporcionados por el INEI. Los resultados son obtenidos por medio de una prueba empírica de regresión lineal bidireccional.

Enhancing agricultural commodity price forecasting with deep learning

En este reporte científico se utilizan modelos predictivos entrenados con Machine Learning para el pronóstico de la variabilidad de precios en productos agrícolas en las regiones de India. Este

subraya la eficacia del enfoque de aprendizaje profundo para este tipo de modelos predictivos.

Machine Learning para predecir la demanda del limón en el Mercado Mayorista de Lima

En esta investigación se evaluó el desempeño de técnicas de Machine Learning para la predicción del comportamiento de la demanda de limones en el Mercado Mayorista de Lima. Se concluyó que esos modelos son idóneos para pronosticar comportamientos en el mercado.

3. Metodología

- Enfoque(s) propuesto:

Este proyecto se basa en la construcción de un modelo predictivo que estima el precio de la bolsa de limones utilizando un enfoque regresional bajo dos métodos: Regresión lineal lasso y Random Forest Regression. Para los cuales se utilizará el siguiente enfoque.

El comportamiento de entrada consiste en un vector de 37 características numéricas (X) preprocesadas y diseñadas, que incluyen: atributos temporales (Año, Día, componentes cíclicos del mes como seno y coseno, variables dummy para Estación y Trimestre), condiciones climáticas (Temp Max, Temp Min, Humedad, Precipitaciones, con características de rezago de 10 días para Temp Min, Temp Max, Humedad y Precipitaciones), eventos disruptivos (Incendio Forestal, Lluvias Intensas, Inundación, Huaico, Fenomeno_Niño), volúmenes de ingreso ('Masa ingreso Bolsa(T)'), y un conjunto robusto de características temporales derivadas de los precios históricos de Bolsa: valores rezagados de 1, 3, 7, 14, 21 y 30 días, así como medias móviles y volatilidades calculadas sobre ventanas de 7, 14 y 30 días. La salida de ambos modelos es un valor numérico continuo (Y) que representa el Precios Bolsa predicho para un día específico. Lo anterior descrito se puede evidenciar con la siguiente notación:

$\chi \in R^{37}$ *el vector de características de entrada para un dia t*

$Y \in R$ *el valor real del Precio Bolsa en el día t*

Definimos el conjunto de datos como:

$$D = \{(X_t, Y_t)\}_{t=1}^{n}$$

Donde:

- $X_t = \left(x_t^{(1)}, x_t^{(2)}, x_t^{(3)}, \ldots, x_t^{(37)}\right)$
- $Y_t = Precio\ Bolsa_t$

el objetivo es aprender una función aproximadora:

$$f: R^{37} \rightarrow R$$

tal que:

$$\widehat{Y}_t = f(X_t)$$

y minimizar el error de predicción bajo una métrica MSE:

$$\frac{MIN}{f} \frac{1}{N} \sum_{t=1}^{N} (Y_t - f(X_t))^2$$

Se definió los siguientes algoritmos utilizados para poder hacer la predicción de precios futuros:

1. **Lasso Regression**
   Modelo de regresión lineal regularizado que penaliza las variables menos relevantes a tener menor influencia en la función solución.
   $$min\left(\sum_{i=1}^{n}(y_i - x_i^{\top}\beta)^2 + \lambda \sum_{j=1}^{37}|\beta_j|\right)$$
   donde λ controla la intensidad de la regulación

2. **Random Forest Regressor**
   Es un algoritmo de aprendizaje basado en la construcción de múltiples árboles de decisión entrenados sobre muestras aleatorias del conjunto de datos y subconjuntos aleatorios de características. La predicción final se obtiene como el promedio de las predicciones de todos los árboles, lo que reduce la varianza y mejora la capacidad de generalización del modelo.

Baseline:
Con el fin de tener una referencia para comparar estos algoritmos se optó por utilizar la media de móvil de los últimos 14 días del precio de bolsa.

$$\widehat{Y_t} = \frac{1}{14} \sum_{k=1}^{14} Y_{t-k}$$

- Adquisición de datos:

Los datos han sido obtenidos desde el módulo de mercados mayoristas de la página web de MIDAGRI. En este dataset se registran los precios diarios del limón durante los años 2022 al 2024. Además, se le ha agregado información del SENAMHI sobre las temperaturas, medición de humedad y precipitación de la región. En particular, se decidió emplear una estación meteorológica ubicada en la zona de Sullana, Piura que se encuentra en un campo de cultivo de limón propiamente

En la siguiente tabla se puede apreciar un ejemplo de los datos obtenidos en el mes de Enero del año 2023

Fecha	Bolsa (s/.)	Caja (s/.)	Temp Max	Temp Min	Humedad	Precipitación
2023-01-01	2.08	2.34	33	21.8	59.8	0
2023-01-02	2.08	2.34	33.6	22	63.7	0

Figura 1: Ejemplo del dataset con los atributos a evaluar.

En adición a la información previamente mencionada, se agregaron variables binarizadas sobre la presencia de ciertos fenómenos meteorológicos y/o desastres naturales. Entre estos se incluyen el Fenómeno del Niño, incendios forestales, lluvias de alta intensidad, inundación y huaicos todos según la definición del INDECI. Estos datos permiten hacer las predicciones considerando factores externos que son de principal impacto sobre la producción de limón y por ende su venta.

Finalmente, para la inclusión de datos de niveles de producción, se incluyó la masa de ingreso de bolsas y cajas de limón al Mercado Mayorista de Lima medido en toneladas. Esta información fue obtenida de la misma página del MIDAGRI.

Preprocesamiento de datos

Durante la exploración de datos se identificó que los precios presentan alta variabilidad, especialmente en la presentación de cajón. Para ilustrar esta situación, se generó un boxplot comparativo.

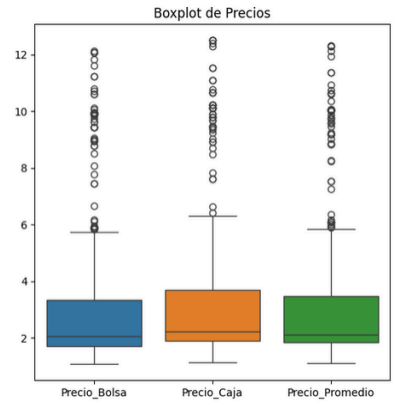

Figura 2: Boxplot de los datos brutos extraídos.
Fuente: Elaboración propia

Con el fin de analizar los episodios de precios elevados, se aplicó un umbral de 2 soles a la serie de precios de la bolsa y las cajas. Esto permitió concentrarse en los días donde el precio superó dicho valor.

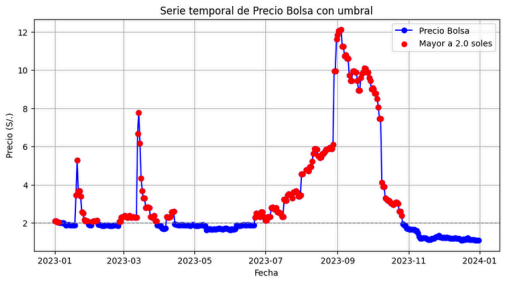

Figura 3: Evolución de los precios de una bolsa de limón.
Fuente: Elaboración propia

Figura 4: Evolución de los precios de una caja de limón.
Fuente: Elaboración propia

Para complementar el análisis, se construyó un histograma de los precios de bolsa y caja superiores a 2 soles, con el fin de observar la frecuencia de ocurrencia de estos valores.

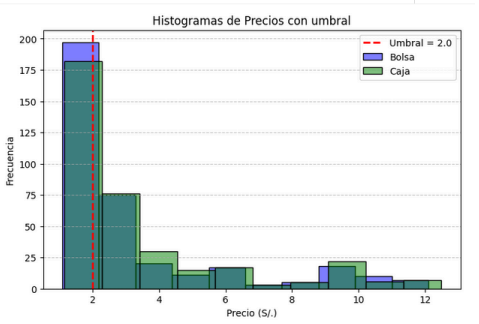

Figura 5: Histograma de los precios de una bolsa y una caja de limón.
Fuente: Elaboración propia

Con esta información de la variabilidad de los datos objetivos (precios), podemos determinar la importancia de mantener esta información presente al momento de entrenar los modelos, en particular porque reflejan tiempos afectados por el Fénomeno del Niño. El preprocesamiento de la data siguió el siguiente flujo:

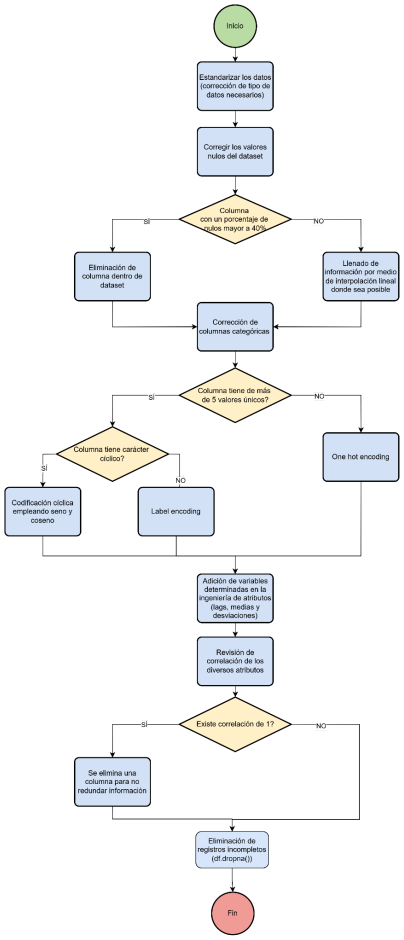

Figura 6: Diagrama de flujo del proceso de preprocesamiento de datos.
Fuente: Elaboración propia

Estandarización de tipos de datos:
Se corrigieron inconsistencias en el formato de datos numéricos, reemplazando comas por puntos decimales y convirtiendo columnas tipo object a float64 mediante pd.to_numeric().

Tratamiento de valores nulos:
Con respecto a los valores nulos observados en precios, variables climáticas y volúmenes de ingreso, se optó por la imputación mediante interpolación lineal (df[col].interpolate(method='linear')) para variables temporales (precios, temperatura, humedad, precipitaciones y masa de ingreso), aprovechando la naturaleza secuencial de los datos para estimar valores faltantes basándose en la tendencia temporal. Sin embargo, en caso se evidencie una columna con más de 40% de datos nulos se evaluará la eliminación de la columna. Para el caso propuesto, sólo se realiza la eliminación de la columna "'Masa ingreso Caja(T)" dado que posee más de 45% de nulos.

Tratamiento de valores atípicos:
Se identificaron valores atípicos en los precios, pero se decidió conservarlos al corresponder a fluctuaciones reales del mercado.

Codificación de variables categóricas:
Con respecto a las variables categóricas, los meses se codificaron cíclicamente para que el modelo pueda identificar su periodicidad. Por otra parte, las estaciones y los trimestres se codificaron mediante one hot encoding, porque representan categorías mutuamente excluyentes y no ordinales.

Ingeniería de características:
- Lags climáticos: Mediante df[col].shift(10) se generaron 'Temp_Min_lag10', 'Temp_Max_lag10', 'Humedad_lag10' y 'Precipitaciones_lag10' para capturar información de 10 días previos.
- Lags de precios: Aplicación de df['Precios Caja'].shift(n) creando 'Precio_lag1', 'Precio_lag3', 'Precio_lag7', 'Precio_lag14', 'Precio_lag21' y 'Precio_lag30'.
- Medias móviles: Cálculo de df['Precios Caja'].rolling(window=n).mean() generando 'Precio_roll7', 'Precio_roll14' y 'Precio_roll30' que suavizan fluctuaciones de corto plazo.
- Volatilidades móviles: Implementación de df['Precios Caja'].rolling(window=n).std() produciendo 'Precio_vol7' y 'Precio_vol30' como medidas de dispersión del precio.
- Términos de interacción: Se considerarán términos de interacción como Temp_Min × Precipitación y Temp_Max x Precipitación para capturar la conexión de la temperatura con las precipitaciones que son dos atributos primordiales a las buenas prácticas de cultivo del limón.

Selección por correlación:
Se eliminaron 'Precios Caja', 'Temp_MinxPrecipitacion' y 'Temp_MaxxPrecipitacion' por correlación perfecta (r=1.0) con otras variables. Los lags de precios se conservaron por su valor predictivo único.

Eliminación de registros incompletos:
Finalmente, tras la generación de características de desfase temporal (lags), se eliminaron los primeros 366 registros que contenían valores NaN producto de las ventanas de cálculo de las características temporales (lags de 30 días y ventanas móviles de 30 días), reduciendo el dataset final a 730 observaciones utilizables y 37 características para el entrenamiento.

■ Entrenamiento del modelo

La regresión lineal múltiple y Random Forest Regressor se presentan como los modelos más adecuados en principio para este proyecto debido a su capacidad de establecer relaciones cuantificables entre variables climáticas (temperatura, humedad, precipitaciones),

frecuencia de desastres naturales y el precio del limón.

Si bien la regresión lineal simple presenta limitaciones para capturar relaciones no lineales y efectos temporales retardados inherentes a los ciclos agrícolas, estas restricciones pueden mitigarse significativamente mediante transformaciones estratégicas de variables como se mencionó en la ingeniería de atributos. Estas incorporaciones amplían la capacidad predictiva del modelo, manteniendo su ventaja principal: la interpretabilidad y la simplicidad computacional.

La regresión Lasso es una técnica de regresión lineal regularizada que utiliza una penalización L1, reduciendo el riesgo de sobreajuste y realizando selección automática de características. Al penalizar los coeficientes, Lasso elimina aquellas variables que no aportan significativamente al modelo. En este estudio, se utilizó Lasso para identificar las variables más influyentes en la predicción y mejorar la capacidad de generalización del modelo considerando la gran cantidad de variables.

Por otro lado, el desarrollo de estas variables sirven a su vez como una herramienta de expansión de datos (feature engineering) para Random Forest Regressor. El uso de Random Forest complementa a la regresión lineal, ya que permite explorar interacciones entre variables y efectos no lineales que el modelo lineal, incluso transformado, podría no capturar completamente.

Implementación:
- División de datos:
La división inicial para la estrategia train_test_split sin shuffle se realizó en 2 conjuntos: 80% para entrenamiento y 20% para validación, asegurando la preservación del orden temporal. Adicionalmente, se implementó la Walk-Forward Validation, que también respeta el orden temporal, para evaluar el rendimiento predictivo. En este último tipo de validación, el modelo solo tiene la información pasada y se entrena periódicamente. El tamaño de la ventana de entrenamiento consta de 365 días para la captura de patrones estacionales y para el periodo de evaluación se centró en los últimos 180 días del conjunto de datos.

- Formulación matemática del modelo:
El entrenamiento del modelo de Regresión Lasso se llevó a cabo utilizando la implementación sklearn.linear_model.Lasso de la biblioteca scikit-learn en Python. Este modelo es una variante de la regresión lineal múltiple, donde se busca estimar los coeficientes que relacionan las características (features) con la variable objetivo (Precios Bolsa) la Regresión Lasso introduce una penalización L1 (la suma de los valores absolutos de los coeficientes) a la función de costo (la suma de errores cuadráticos). El modelo fue ajustado (fit) utilizando el conjunto de entrenamiento permitiéndole calcular los pesos óptimos que mejor representan la relación entre las variables de entrada y el precio objetivo.

Por otro lado, dada la naturaleza de serie temporal de los datos, el entrenamiento del modelo Random Forest (RandomForestRegressor de libería scikit-learn en Python) se llevó a cabo utilizando una estrategia de validación walk-forward, la cual es particularmente adecuada para evaluar el rendimiento predictivo en escenarios de pronóstico.

Validación y selección de modelo:
Se realizarán análisis estadísticos los cuales serán evaluados con el fin de identificar la significancia de los parámetros escogidos y obtener el modelo más parsimonioso sin sacrificar capacidad predictiva.

4. Experimentación y Resultados

- Setup experimental:

El conjunto de datos empleado se obtuvo gracias del Gran Mercado Mayorista de Lima e información climática del SENAMHI. Adicionalmente, se incorporaron datos de eventos naturales provenientes de la plataforma del Instituto Nacional de Defensa Civil (INDECI). Los datos contienen precios diarios de la bolsa de limón, condiciones meteorológicas, volumen de ingreso al mercado, y variables indicadoras de eventos naturales como lluvias intensas, huaycos, incendios forestales, inundaciones y presencia del Fenómeno del Niño.

Para la evaluación del desempeño de los modelos se emplearon cuatro métricas principales: el Error Cuadrático Medio (MSE) y el Error Absoluto Medio (MAE), que miden la diferencia entre los valores reales y predichos; el Explained Variance Score, que indica cuanto del comportamiento de la variable objetivo es explicado por el modelo; y el Coeficiente de Determinación (R^2). En conjunto, estas métricas permiten comparar de manera objetiva el desempeño predictivo de diferentes modelos.

Se implementaron tres modelos para determinar el enfoque más adecuado: un baseline basado en la media móvil de los últimos 14 días, el cual sirve como referencia mínima de desempeño, Lasso Regression, un modelo de regresión lineal con regularización, y Random Forest Regressor, un modelo para basado en árboles de decisión capaz de capturar relaciones no lineales complejas. Cada modelo fue entrenado bajo las mismas condiciones para garantizar una comparación justa.

Particularmente para el modelo Lasso, se realizaron dos esquemas de validación: uno con walk-forward validation y otro con train_test_split sin shuffle, con el fin de evaluar la robustez del modelo bajo diferentes configuraciones y verificar su comportamiento en escenarios de validación

estándar. En el caso del modelo Random Forest, se empleó el esquema walk-forward validation.

Adicionalmente, se evaluó el impacto de diferentes combinaciones de variables en el desempeño del modelo con el objetivo de determinar si la inclusión de más variables contribuye a mejorar la capacidad predictiva del modelo o, por el contrario, introduce redundancia. Además, se hizo la calibración respectiva en el caso de Random Forest Regressor para encontrar la combinación óptima de niveles de profundidad y número de árboles.

Resultados y Discusión:

Se obtuvieron los siguientes resultados numéricos tras la evaluación de los modelos:

Modelo/ Métrica	MSE	MAE	Varianza explicada	R^2
Lasso (train_test_split sin shuffle)	0.0934	0.2529	0.7504	0.4058
Lasso (Walk-Forward)	0.0253	0.1164	0.8138	0.8117
Random Forest (Walk-Forward)	0.0171	0.0849	0.8724	0.8723
Baseline (media movil de 14 días)	0.0492	0.1539	0.6291	0.6275

Figura 7: Tabla de los resultados de los modelos evaluados
Fuente: Elaboración propia

Los resultados presentados en la figura 7 permiten establecer un orden en el rendimiento. El Baseline establecer el umbral mínimo con un R^2 de 0.6275. Se observa que la implementación del modelo Lasso con validación Walk-Forward logra una mejora sustancial respecto a la línea base, reduciendo el MSE a 0.0253 y elevando el R^2 a 0.8117. Sin embargo, el modelo Random Forest Regresor exhibe un desempeño mayor, con un R^2 de 0.8723 con menor MAE de 0.0849, lo que implica que el modelo aprovecha de mejor manera las relaciones no lineales.

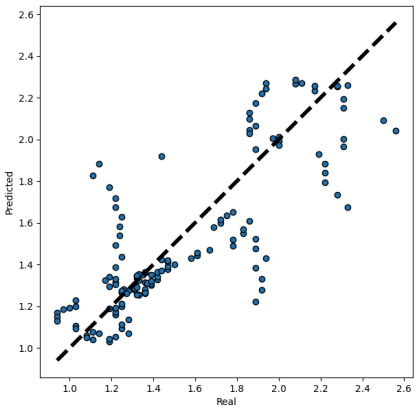

Figura 8: Scatterplot que representa los resultados obtenidos del baseline
Fuente: Elaboración propia

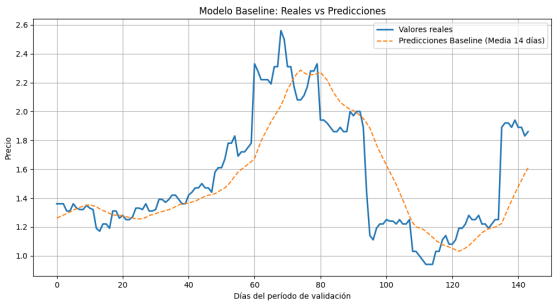

Figura 9: Scatterplot que representa la variación de las predicciones y los precios reales en el baseline.
Fuente: Elaboración propia

Como punto de referencia inicial, el modelo Baseline presenta en el gráfico de dispersión (Figura 8) una variabilidad considerable alrededor de la diagonal, lo que implica que no puede predecir la variación correctamente. Adicionalmente, la serie temporal (Figura 9) revela que la curva de predicción es visiblemente más suave que las de los valores reales. Esta característica provoca graves imprecisiones y retrasos en la respuesta ante cambios de tendencia y revela una incapacidad para capturar los picos y valles de alta volatilidad. Esto confirma que, si bien el Baseline sigue la tendencia general, no puede modelar los cambios de la serie, lo que justifica la necesidad de modelos multivariados más complejos, como Lasso o Random Forest.

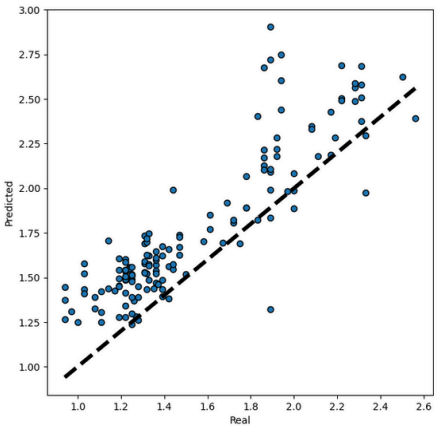

Figura 10: Scatterplot que representa los resultados obtenidos con el regresor lineal con train_test_split con shuffle=False
Fuente: Elaboración propia

En la figura 10 se puede apreciar que para el modelo lineal con train_test_split sin shuffle, las predicciones presentan un comportamiento que sugiere relación lineal, sin embargo, se observan valores en su mayoría dispersos por encima de la línea ideal. Esta dispersión evidencia las limitaciones del modelo para capturar adecuadamente la variabilidad temporal de los

precios cuando se entrena una única vez con datos históricos.

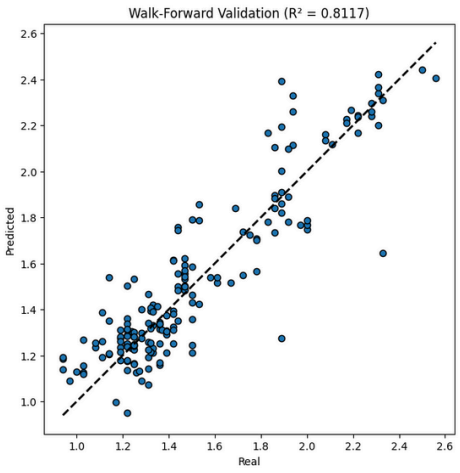

Figura 11: Scatterplot que representa los resultados obtenidos con el regresor lineal Lasso con Walk-Forward
Fuente: Elaboración propia

Por otro lado, en la figura 11 se muestra una línea de puntos más cercana a la diagonal de 45 grados, reflejando una mejora sustancial en la capacidad de predicción. Este resultado es coherente con el hecho de que el Walk-Forward validation utiliza únicamente información de las semanas pasadas para las predicciones, reentrenando el modelo progresivamente y evitando así fugas de información.

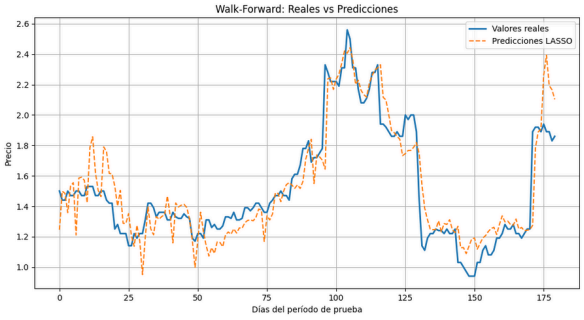

Figura 12: Diagrama que representa la variación de las predicciones y los precios reales con walk forward
Fuente: Elaboración propia

De igual modo, en la figura 12 se puede apreciar un gráfico temporal de las predicciones y su variación respecto a los datos reales en validación. En este se observa que el modelo Lasso con walk-forward realiza predicciones relativamente cercanas a los valores reales.

Existen algunas desviaciones considerables, pues tiende a sobreestimar el precio del limón, especialmente en los picos del precio de la bolsa de limón. En cuanto a los valles, cuando el limón cuesta entre 1.2 y 1.6, tiende a tener un comportamiento errático. Esto significa que el modelo predice bien las tendencias de mediano plazo, pero presenta limitaciones para modelar los extremos de la serie temporal.

Estos resultados son correspondientes a lo esperado para el modelo Lasso. Esto es porque, al ser lineal y regularizado, tiende a suavizar las predicciones. Por consiguiente, no logra capturar completamente las variaciones propias del precio del limón Además, la dependencia de factores no lineales dificultan que un modelo lineal represente correctamente los extremos.

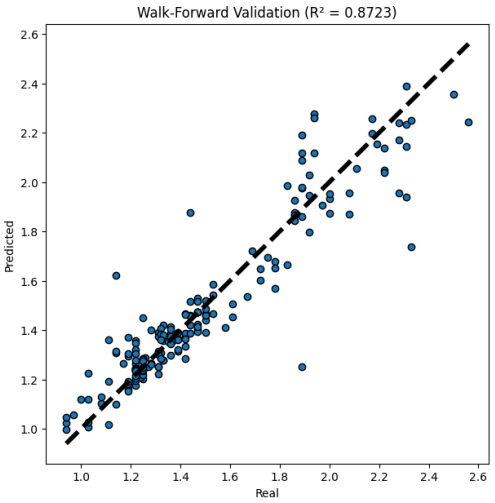

Figura 13: Scatterplot que representa los resultados obtenidos con el Random Forest con Walk-Forward
Fuente: Elaboración propia

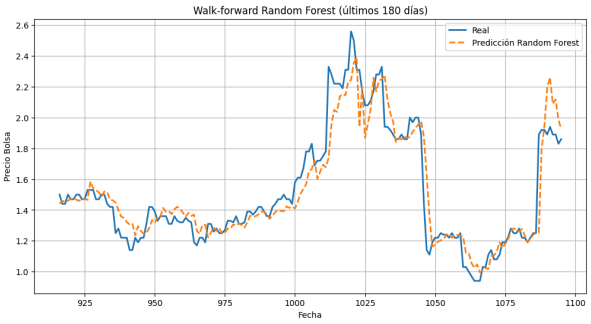

Figura 14: Diagrama que representa la variación de las predicciones y los precios reales en la validación de modelo Random Forest
Fuente: Elaboración propia

En contraste, las figuras 13 y 14 muestran el desempeño de Random Forest Regressor con walk-forward validation. El scatterplot (Figura 13) evidencia una concentración significativamente mayor de puntos sobre la diagonal, indicando predicciones más precisas. La figura 14 confirma que el modelo captura mejor tanto las tendencias de mediano plazo como los extremos de la serie temporal, demostrando su capacidad de modelar relaciones no lineales complejas presentes en los precios de la bolsa de limón.

5. Conclusión

En conclusión, se evidencia que el uso de atributos basados en condiciones climáticas, ocurrencia de desastres naturales, niveles de producción y análisis histórico del comportamiento de precios permite hacer una predicción futura más robusta que las predicciones obtenidas empleando únicamente la media móvil de precios de las últimas 2 semanas empleada como baseline.

Adicionalmente, podemos concluir que el modelo Random Forest Regressor, configurado con 400 estimadores y profundidad máxima de 20 niveles, se posiciona como mejor modelo frente a la regresión lineal múltiple. Esto sucede gracias a su capacidad de capturar relaciones complejas (no lineales) de los atributos propuestos. Si bien presenta un tiempo de procesamiento mayor, no resulta excesivo ni compromete la viabilidad del uso del modelo como herramienta, especialmente considerando la mejora significativa en la calidad de las predicciones.

6. Sugerencias de trabajos futuros

En primer lugar, se recomienda integrar fuentes de datos agrícolas directas para obtener información precisa sobre los niveles de producción del limón dado que hubo dificultades en obtener esta información de manera indirecta. Asimismo, se debería buscar mayor cantidad de fuentes de datos climáticos para cubrir datos faltantes del dataset y generar un dataset más nutrido y representativo.

En segundo lugar, se sugiere mayor experimentación con técnicas adicionales de ingeniería de atributos para considerar otro conjunto de atributos de efectos conjuntos y/o otras ventanas de tiempo a considerar para atributos lags.

En tercer lugar, se plantea la posibilidad de evaluación con modelos más avanzados, como redes neuronales, para comprobar el rendimiento y robustez en las predicciones del sistema frente a modelos más tradicionales.

Finalmente, dado que una de las principales intenciones para trabajos futuros es la aplicación del enfoque propuesto en otras regiones productoras de limón en el Perú, como Lambayeque o Tumbes, se recomienda evaluar y ajustar el modelo con datos de estas regiones para validar la generalización y desempeño regional.

7. Implicancias éticas

Se debe considerar que el modelo fue entrenado exclusivamente con datos del Gran Mercado Mayorista de Lima para el período 2022-2024, por lo que su desempeño no está validado para otras regiones o períodos con condiciones climáticas significativamente diferentes. El uso de predicciones sin comprender estas limitaciones podría llevar a decisiones comerciales inadecuadas.

Para abordar estos riesgos, es recomendable:

- Documentar de forma transparente las fuentes de datos, sus limitaciones y los supuestos del modelo.
- Monitorear sistemáticamente el desempeño de distintos subgrupos para detectar sesgos y recalibrar el modelo cuando cambien las condiciones climáticas o de mercado.

8. Link del repositorio del trabajo

https://colab.research.google.com/drive/1uDdYcOclucS1yaFB2cLDrDSI9Hah7SWJ?usp=sharing

9. Declaración de contribución de cada integrante

Integrante	Participación
Sanchez Boza, R (20221221)	Redacción de introducción de informe, planteamiento inicial de preprocesamiento en código, creación de código para el baseline, redacción de conclusiones y sugerencias de trabajos futuros y revisión de formato y ortografía del informe final.
Lau Lu, O (20220884)	Contribución en la redacción del preprocesamiento en el informe. Redacción de la parte de experimentación y los resultados del modelo Lasso con los 2 esquemas de validación. Apoyo en la revisión y corrección del informe. Recopilación de los datos de precios y variables climáticas para el dataset. Realización del esquema Walk-forward en el modelo Lasso.
Sanchez Pastor, R (20201230)	Contribución en la redacción de resultados, recolección de datos. Implementación del método de validación para random forest, corrección de data leaks, configuración de hiperparametros en el modelo de random Forest, implementación de gráficos de las series temporales, optimización de la ventana temporal para la validación de walk-forward.
Diaz Meoño, F (20222110)	Redacción del enfoque y creación del código base para la experimentación con ambos métodos
Tenipuclla Flores, G (20224918)	Contribución en la redacción de las siguientes secciones: Preprocesamiento, entrenamiento del modelo e implicancias éticas. Generación de dataset de precios de 2 años.

10. Referencias

[1]. Ministerio de Desarrollo Agrario y Riego. (2025). Situación del Limón en el Perú, un análisis económico. Recuperado 8 de octubre de 2025, de https://repositorio.midagri.gob.pe/bitstre

[2]. Ministerio de Agricultura y Riego. (2025). Situación del Limón en el Perú [Documento de análisis económico]. MIDAGRI. Recuperado 10 de octubre de 2025 de https://www.gob.pe/institucion/midagri/colecciones/350-documentos-de-analisis-economico

[3]. Ministerio de Desarrollo Agrario y Riego. (s. f.). Precios de Principales Productos de la Canasta Básica Familiar. Recuperado 8 de octubre de 2025, de https://app.powerbi.com/view?r=eyJrIjoiZTEwOGNjODItYWJjMS00ZDQ2LTllMWQtYjE5MjVmNWQ4NTZiIiwidCI6IjdmMDg0NjI3LTdmNDAtNDg3OS04OTE3LTk0Yjg2ZmQzNWYzZiJ9

[4]. Servicio Nacional de Sanidad Agraria del Perú. (s. f.). GUÍA PARA LA IMPLEMENTACIÓN DE BUENAS PRÁCTICAS AGRÍCOLAS (BPA) PARA EL CULTIVO DE LIMÓN. Recuperado 8 de octubre de 2025, de https://www.senasa.gob.pe/senasa/descargasarchivos/2020/07/Guia-BPA-LIMON.pdf

[5]. Datos Hidrometeorológicos a nivel nacional https://www.senamhi.gob.pe/?p=estaciones

[6]. Blasques, F., Gorgi, P., Koopman, S. J., & Sampi Bravo, J. (2025). Food prices and production in the aftermath of natural disasters: The case of Peru. Food prices and production in the aftermath of natural disasters: the case of Peru.https://tinbergen.nl/discussion-paper/6411/25-024-iii-food-prices-and-production-in-the-aftermath-of-natural-disasters-the-case-of-peru

[7]. Instituto Nacional de Defensa Civil - INDECI. (s.f.). Informe de emergencia. https://portal.indeci.gob.pe/informe/informe-de-emergencia/

[8]. Manogna, R. L., Dharmaji, V., & Sarang, S. (2025). Enhancing agricultural commodity price forecasting with deep learning. Scientific Reports, 15(1), 20903. https://doi.org/10.1038/s41598-025-05103-z

[9]. Porras Cuadros, D. Y. (2024). Machine Learning para predecir la demanda del limón en el Mercado Mayorista de Lima. Repositorio Institucional - UCV. https://repositorio.ucv.edu.pe/handle/20.500.12692/150041